A Survey of Context Engineering for Large Language Models

Lingrui Mei[1,6,†] **Jiayu Yao**[1,6,†] **Yuyao Ge**[1,6,†] **Yiwei Wang**[2] **Baolong Bi**[1,6,†]
Yujun Cai[3] **Jiazhi Liu**[1] **Mingyu Li**[1] **Zhong-Zhi Li**[6] **Duzhen Zhang**[6]
Chenlin Zhou[4] **Jiayi Mao**[5] **Tianze Xia**[6] **Jiafeng Guo**[1,6,†] **Shenghua Liu**[1,6,†,✉]

[1] Institute of Computing Technology, Chinese Academy of Sciences,
[2] University of California, Merced, [3] The University of Queensland
[4] Peking University, [5] Tsinghua University,
[6] University of Chinese Academy of Sciences

Abstract: The performance of Large Language Models (LLMs) is fundamentally determined by the contextual information provided during inference. This survey introduces **Context Engineering**, a formal discipline that transcends simple prompt design to encompass the systematic optimization of information payloads for LLMs. We present a comprehensive taxonomy decomposing Context Engineering into its foundational **Components** and the sophisticated **Implementations** that integrate them into intelligent systems. We first examine the foundational **Components**: (1) **Context Retrieval and Generation**, encompassing prompt-based generation and external knowledge acquisition; (2) **Context Processing**, addressing long sequence processing, self-refinement, and structured information integration; and (3) **Context Management**, covering memory hierarchies, compression, and optimization. We then explore how these components are architecturally integrated to create sophisticated **System Implementations**: (1) **Retrieval-Augmented Generation (RAG)**, including modular, agentic, and graph-enhanced architectures; (2) **Memory Systems**, enabling persistent interactions; (3) **Tool-Integrated Reasoning**, for function calling and environmental interaction; and (4) **Multi-Agent Systems**, coordinating communication and orchestration. Through this systematic analysis of over 1400 research papers, our survey not only establishes a technical roadmap for the field but also reveals a critical research gap: a fundamental asymmetry exists between model capabilities. While current models, augmented by advanced context engineering, demonstrate remarkable proficiency in *understanding* complex contexts, they exhibit pronounced limitations in *generating* equally sophisticated, long-form outputs. Addressing this gap is a defining priority for future research. Ultimately, this survey provides a unified framework for both researchers and engineers advancing context-aware AI.

† *Also affiliated with: (1)Key Laboratory of Network Data Science and Technology, ICT, CAS; (2)State Key Laboratory of AI Safety*
✉ *Corresponding Author*

Keywords: Context Engineering, Large Language Models, LLM Agent, Multi-Agent Systems

Date: July 17, 2025
Code Repository: https://github.com/Meirtz/Awesome-Context-Engineering
Contact: meilingrui23b@ict.ac.cn, liushenghua@ict.ac.cn

arXiv:2507.13334v1 [cs.CL] 17 Jul 2025

大型语言模型的上下文工程综述

凌瑞梅[1,6,†] **姚嘉宇**[1,6,†] **葛雨瑶**[1,6,†] **王怡伟**[2] **毕宝龙**[1,6,†]
蔡宇君[3] **刘佳志**[1] **李明宇**[1] **李中志**[6] **张都珍**[6]
周晨林[4] **毛嘉怡**[5] **夏天泽**[6] **郭嘉峰**[1,6,†] **刘生华**[1,6,†,✉]

[1] 中国科学院计算技术研究所,
[2] 加州大学默塞德分校, [3] 昆士兰大学
[4] 北京大学,[5] 清华大学,
[6] 中国科学院大学

摘要： 大型语言模型（LLMs）的性能从根本上由推理过程中提供的上下文信息决定。本综述介绍了 **上下文工程**，这是一个超越简单提示设计的正式学科，涵盖了为LLMs系统优化信息有效载荷。我们提出一个全面的分类，将上下文工程分解为其基础 **组件** 以及将它们集成到智能系统中的复杂 **实现**。我们首先考察基础 **组件**：(1)**上下文检索与生成**，包括基于提示的生成和外部知识获取；(2)**上下文处理**，处理长序列、自我完善和结构化信息集成；以及 (3) **上下文管理**，涵盖内存层次结构、压缩和优化。然后，我们探讨这些组件如何架构性地集成以创建复杂的 **系统实现**：(1) **检索增强生成（RAG）**，包括模块化、代理和图增强架构；(2) **内存系统**，实现持久交互；(3)**工具集成推理**，用于函数调用和环境交互；以及 (4)**多代理系统**，协调通信和编排。通过对超过1400篇研究论文的系统分析，我们的综述不仅为该领域建立了技术路线图，还揭示了一个关键研究差距：模型能力之间存在根本性不对称。虽然当前模型通过先进的上下文工程表现出色地 理解 复杂上下文，但在 生成 同样复杂的长格式输出方面存在明显限制。解决这一差距是未来研究的首要任务。最终，本综述为研究人员和工程师推进上下文感知AI提供了一个统一的框架。

† 也隶属于： *(1)*网络数据科学与技术重点实验室，中国科学院计算技术研究所； *(2)*人工智能安全国家重点实验室
✉ 通讯作者

关键词: 上下文工程，大型语言模型，LLM代理，多智能体系统

日期: 2025年7月17日
代码仓库: https://github.com/Meirtz/Awesome-Context-Engineering
联系:
meilingrui23b@ict.ac.cn,liushenghua@ict.ac.cn

Contents

目录

1. Introduction

The advent of LLMs has marked a paradigm shift in artificial intelligence, demonstrating unprecedented capabilities in natural language understanding, generation, and reasoning [103, 1059, 453]. However, the performance and efficacy of these models are fundamentally governed by the *context* they receive. This context—ranging from simple instructional prompts to sophisticated external knowledge bases—serves as the primary mechanism through which their behavior is steered, their knowledge is augmented, and their capabilities are unleashed. As LLMs have evolved from basic instruction-following systems into the core reasoning engines of complex applications, the methods for designing and managing their informational payloads have correspondingly evolved into the formal discipline of **Context Engineering** [25, 1256, 1060].

The landscape of context engineering has expanded at an explosive rate, resulting in a proliferation of specialized yet fragmented research domains. We conceptualize this landscape as being composed of foundational *components* and their subsequent *implementations*. The foundational components represent the systematic pipeline of context engineering through three critical phases: **Context Retrieval and Generation**, encompassing prompt-based generation and external knowledge acquisition [25, 591, 48]; **Context Processing**, involving long sequence processing, self-refinement mechanisms, and structured information integration [196, 735, 489]; and **Context Management**, addressing memory hierarchies, compression techniques, and optimization strategies [1362, 1074, 813].

These foundational components serve as the building blocks for more complex, application-oriented implementations that bridge LLMs to external realities. These systems include **Advanced Retrieval-Augmented Generation (RAG)**, which has evolved into modular and agentic architectures for dynamic knowledge

1. 简介

大语言模型的兴起标志着人工智能的范式转变，展示了在自然语言理解、生成和推理方面的前所未有的能力 [103, 1059, 453]。然而，这些模型的性能和功效基本上由它们接收的上下文 决定。这种上下文——从简单的指令提示到复杂的外部知识库——是引导其行为、增强其知识和释放其能力的主要机制。随着大语言模型从基本的指令跟随系统演变为复杂应用程序的核心推理引擎，设计和管理其信息有效载荷的方法也随之演变为正式的**上下文工程** [25, 1256, 1060]。

上下文工程的格局以爆炸性速度扩展，导致专业但碎片化的研究领域的激增。我们将这个格局概念化为由基础 组件 及其后续 实现组成。基础组件代表通过三个关键阶段进行的上下文工程系统管道：**上下文检索和生成**，包括基于提示的生成和外部知识获取 [25, 591, 48]；**上下文处理**，涉及长序列处理、自我完善机制和结构化信息集成[196, 735, 489]；以及 **上下文管理**，解决内存层次结构、压缩技术和优化策略 [1362, 1074, 813]。

这些基础组件是更复杂、面向应用的实现的基础，它们将LLMs与外部现实连接起来。这些系统包括**高级检索增强生成（RAG）**，它已经发展成为模块化和代理化的架构，用于动态知识

injection [591, 312, 965, 311]; explicit **Memory Systems** that mimic human cognitive faculties for persistent information retention [1182, 935, 1362]; and the entire ecosystem of **Intelligent Agent Systems**. This latter category represents the pinnacle of context engineering, where agents leverage **Function Calling** and **Tool-Integrated Reasoning** to interact with the world [931, 858, 663], and rely on sophisticated **Agent Communication** protocols and **Context Orchestration** to achieve complex goals in multi-agent configurations [356, 246, 894, 128].

While each of these domains has generated substantial innovation, they are predominantly studied in isolation. This fragmented development obscures the fundamental connections between techniques and creates significant barriers for researchers seeking to understand the broader landscape and practitioners aiming to leverage these methods effectively. The field urgently requires a unified framework that systematically organizes these diverse techniques, clarifies their underlying principles, and illuminates their interdependencies.

To address this critical gap, this survey provides the first comprehensive and systematic review of Context Engineering for LLMs. Our primary contribution is a novel, structured taxonomy that classifies the multifaceted techniques used to design, manage, and optimize context. This taxonomy organizes the field into coherent categories, distinguishing between foundational *Components* and their integration into sophisticated *System Implementations*. Through this framework, we: (1) provide a clear and structured overview of the state-of-the-art across each domain; (2) analyze the core mechanisms, strengths, and limitations of different approaches; and (3) identify overarching challenges and chart promising directions for future research. This work serves as both a technical roadmap for navigating the complex landscape of context engineering and a foundation for fostering deeper understanding and catalyzing future innovation.

The remainder of this paper is organized as follows. After discussing related work and formally defining Context Engineering, we first examine the **Foundational Components** of the field, covering Context Retrieval and Generation, Context Processing, and Context Management. We then explore their **System Implementations**, including Retrieval-Augmented Generation, Memory Systems, Tool-Integrated Reasoning, and Multi-Agent Systems. Finally, we discuss evaluation methodologies, future research directions, and conclude the survey. Figure 1 provides a comprehensive overview of our taxonomy, illustrating the hierarchical organization of techniques and their relationships within the Context Engineering landscape.

2. Related Work

The rapid maturation of LLMs has spurred a significant body of survey literature aiming to map its multifaceted landscape. This existing work, while valuable, has largely focused on specific vertical domains within the broader field of what we define as Context Engineering. Our survey seeks to complement these efforts by providing a horizontal, unifying taxonomy that distinguishes between foundational components and their integration into complex systems, thereby bridging these specialized areas.

Foundational Components Numerous surveys have addressed the foundational **Components** of context engineering that form the core technical capabilities for effective context manipulation. The challenge of **Context Retrieval and Generation** encompasses both prompt engineering methodologies and external knowledge acquisition techniques. Surveys on prompt engineering have cataloged the vast array of techniques for guiding LLM behavior, from basic few-shot methods to advanced, structured reasoning frameworks [25, 253, 1313]. External knowledge retrieval and integration techniques, particularly through knowledge graphs and structured data sources, are reviewed in works that survey representation techniques, integration

注入 [591, 312, 965, 311]; 显式**记忆系统** 模拟人类认知功能以实现持久信息保留 [1182, 935, 1362]; 以及整个 **智能体系统**。后者代表了上下文工程的顶峰，其中智能体利用**函数调用**和 **工具集成推理** 与世界交互 [931, 858, 663], 并依赖于复杂的**智能体通信** 协议和 **上下文编排** 以在多智能体配置中实现复杂目标 [356, 246, 894, 128]。

尽管这些领域都产生了大量创新，但它们主要是在孤立状态下进行研究的。这种碎片化的发展掩盖了技术和技术之间的基本联系，并为寻求了解更广泛领域的研究人员以及旨在有效利用这些方法的从业者制造了重大障碍。该领域迫切需要一个统一的框架，系统地组织这些多样化的技术，阐明其基本原理，并揭示其相互依赖关系。

为了解决这一关键差距，本调查提供了针对LLM上下文工程的首个全面且系统的综述。我们的主要贡献是一个新颖的、结构化的分类法，将用于设计、管理和优化上下文的多方面技术进行分类。该分类法将领域组织成连贯的类别，区分基础 组件 及其在复杂 系统实现 中的集成。通过这个框架，我们：(1) 提供每个领域最先进技术的清晰和结构化概述；(2) 分析不同方法的核心理念、优势和局限性；(3) 确定总体挑战并规划未来研究的有希望的方向。这项工作既是为导航上下文工程的复杂领域提供技术路线图，也是为促进更深入的理解和激发未来创新奠定基础。

本文的其余部分组织如下。在讨论相关工作并正式定义上下文工程后，我们首先考察了 **该领域的** 基础组件，**涵盖上下文检索与生成、上下文处理以及上下文管理。然后，我们探讨了它们的** 系统实现，包括检索增强生成、记忆系统、工具集成推理和多智能体系统。最后，我们讨论了评估方法、未来研究方向，并总结本次调查。图1 <style id='15'>提供了我们对分类体系的全面概述，展示了技术和它们在上下文工程领域中的关系。

2. 相关工作

LLM 的快速成熟推动了大量旨在描绘其多方面领域的调查文献。这些现有工作虽然有价值，但基本上主要集中在我们所定义的更广泛的上下文工程领域内的特定垂直领域。我们的调查通过提供一个水平、统一的分类体系来补充这些努力，该分类体系区分了基础组件及其在复杂系统中的集成，从而弥合了这些专门领域。

基础组件 许多调查已经探讨了 **上下文工程的基础** 组件，这些组件构成了有效上下文操作的核心技术能力。上下文检索和生成 **的挑战涵盖了提示工程方法** 和外部知识获取技术。关于提示工程的调查已经 [25, 253, 1313]收录了指导LLM行为的各种技术，从基本的 少样本方法到高级的、结构化的推理框架 [25, 253, 1313]。外部知识检索和整合技术，特别是通过知识图谱和结构化数据源，在调查表示技术、整合

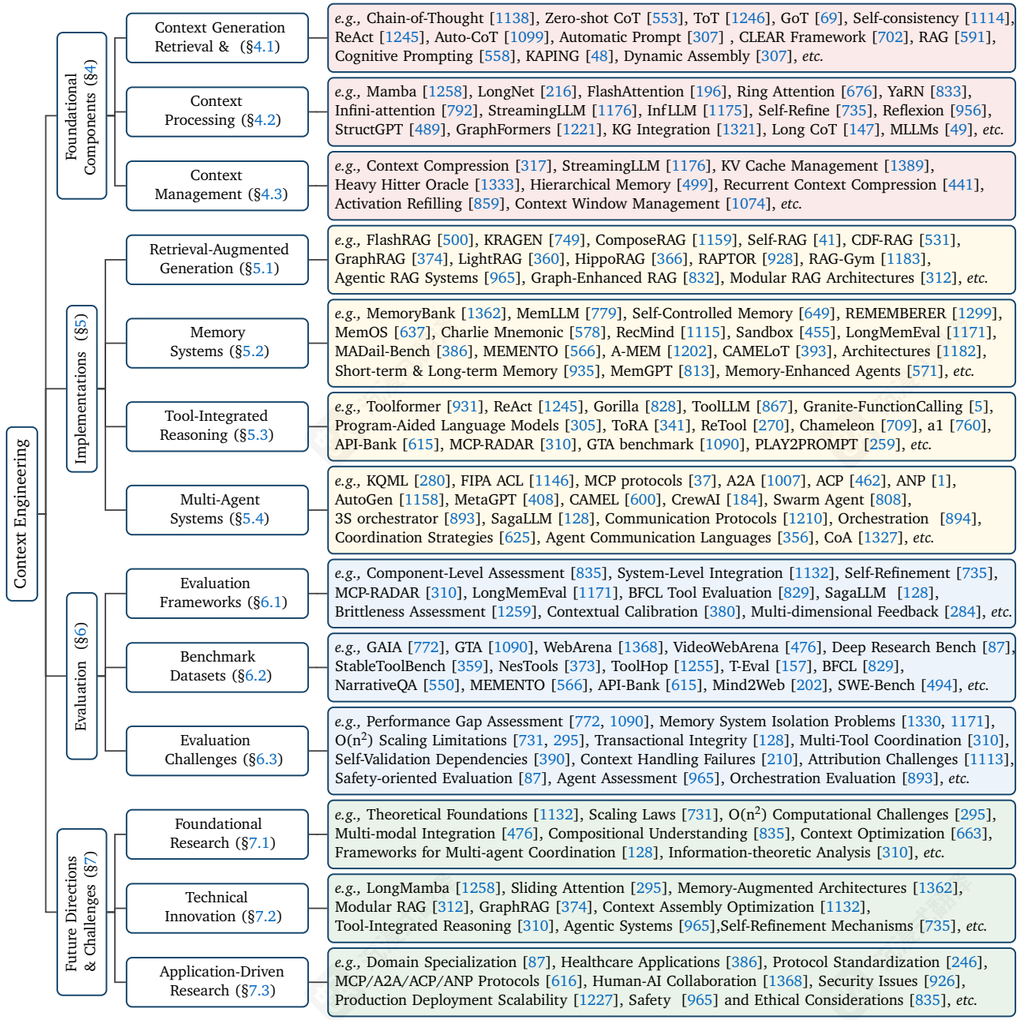

Figure 1: The taxonomy of Context Engineering in Large Language Models is categorized into foundational components, system implementations, evaluation methodologies, and future directions. Each area encompasses specific techniques and frameworks that collectively advance the systematic optimization of information payloads for LLMs.

paradigms, and applications in enhancing the factual grounding of LLMs [483, 428, 817, 889].

The domain of **Context Processing** addresses the technical challenges of handling long sequences, self-refinement mechanisms, and structured information integration. Long context processing is addressed in surveys analyzing techniques for extending context windows, optimizing attention mechanisms, and managing memory efficiently [831, 645, 1289, 268]. The internal cognitive processes of LLMs are increasingly

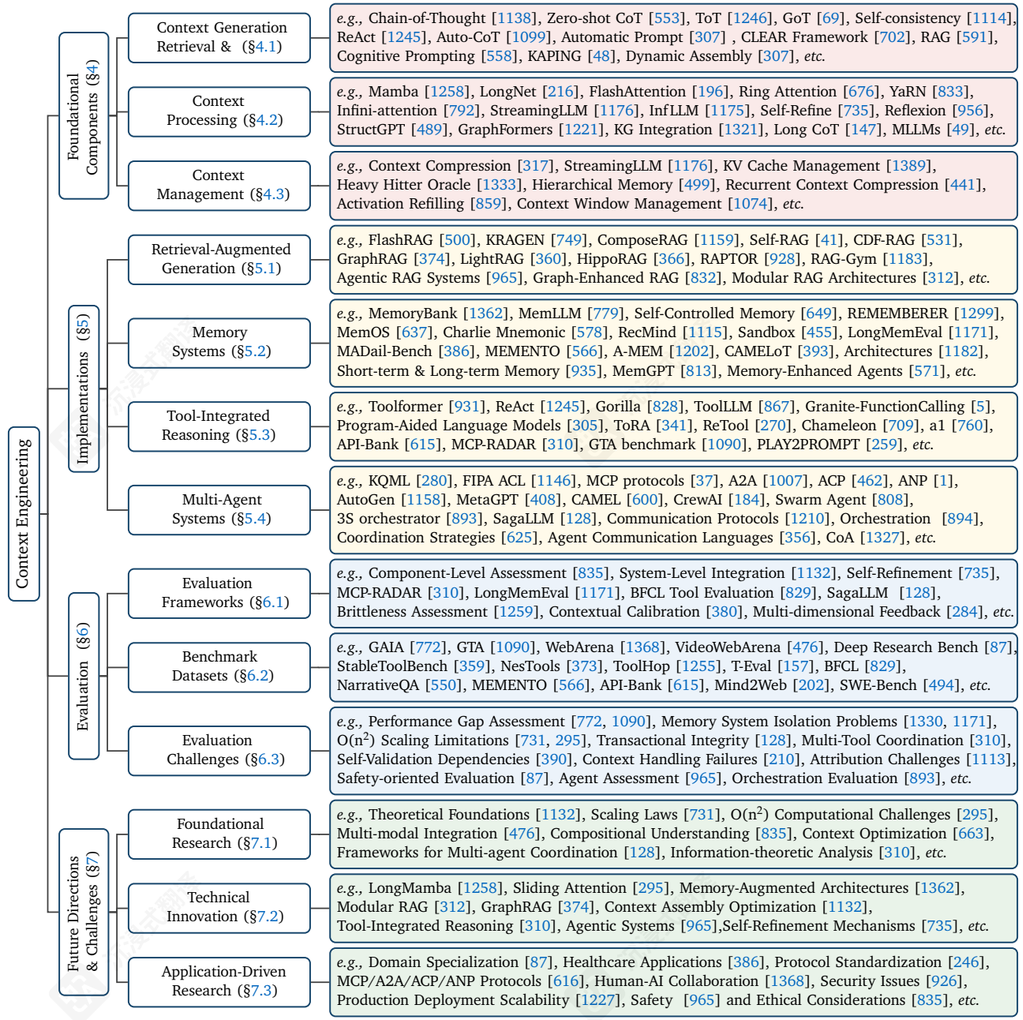

图 1：大型语言模型中的上下文工程分类为基础组件、系统实现、评估方法和未来方向。每个领域都包含特定的技术和框架，这些技术和框架共同推动了信息有效载荷对LLM的系统优化。

范式和应用，以增强LLM的事实基础 [483, 428, 817, 889]。

上下文处理 **领域** 处理处理长序列、自完善机制和结构化信息集成的技术挑战。长上下文处理在分析扩展上下文窗口、优化注意力机制和管理内存效率的技术方面的调查中得到了解决 [831, 645, 1289, 268]。LLM的内部认知过程越来越

surveyed, with works on self-contextualizing techniques and self-improvement paradigms gaining prominence [1329, 227, 1167, 935].

Finally, **Context Management** literature focuses on memory hierarchies, compression techniques, and optimization strategies that enable effective information organization and retrieval within computational constraints. While comprehensive surveys specifically dedicated to context management as a unified domain remain limited, related work on memory systems and context compression techniques provides foundational insights into these critical capabilities.

System Implementation In parallel, the literature has extensively covered the **System Implementations** that integrate foundational components into sophisticated architectures addressing real-world application requirements. The domain of **RAG** has received substantial attention, with foundational surveys tracing its development and impact on mitigating hallucinations [311, 253, 1131]. More recent work has surveyed the evolution towards modular, agentic, and graph-enhanced RAG architectures [162, 622, 120, 312, 1391].

Memory Systems that enable persistent interactions and cognitive architectures have been explored through surveys focusing on memory-enhanced agents and their applications. The broader category of **LLM-based Agents** serves as a foundational area, with comprehensive overviews of autonomous agents, their architecture, planning, and methodologies [1091, 719, 277, 843, 1340, 498, 1272].

Tool-Integrated Reasoning encompassing function calling mechanisms and agent-environment interaction are well-documented, exploring the evolution from single-tool systems to complex orchestration frameworks [663, 858, 771, 867]. The evolution towards **Multi-Agent Systems (MAS)** represents another focal point, with surveys detailing MAS workflows, infrastructure, communication protocols, and coordination mechanisms [625, 356, 246, 1235, 38, 503, 187, 458].

Evaluation The critical aspect of **evaluating** these complex systems has been thoroughly reviewed, with works analyzing benchmarks and methodologies for assessing component-level and system-level capabilities and performance [1259, 380, 835, 310]. This evaluation literature spans both foundational component assessment and integrated system evaluation paradigms.

Our Contribution While these surveys provide indispensable, in-depth analyses of their respective domains, they inherently present a fragmented view of the field. The connections between RAG as a form of external memory, tool use as a method for context acquisition, and prompt engineering as the language for orchestrating these components are often left implicit. Our work distinguishes itself by proposing *Context Engineering* as a unifying abstraction that explicitly separates foundational components from their integration in complex implementations. By organizing these disparate fields into a single, coherent taxonomy, this survey aims to elucidate the fundamental relationships between them, providing a holistic map of how context is generated, processed, managed, and utilized to steer the next generation of intelligent systems.

3. Why Context Engineering?

As Large Language Models (LLMs) evolve from simple instruction-following systems into the core reasoning engines of complex, multi-faceted applications, the methods used to interact with them must also evolve. The term “prompt engineering,” while foundational, is no longer sufficient to capture the full scope of designing, managing, and optimizing the information payloads required by modern AI systems. These systems do not

被调查，关于自上下文化技术和自改进范式的作品日益突出[1329, 227, 1167, 935]。

最后，**上下文管理** 文献关注内存层次结构、压缩技术和优化策略，这些策略能够在计算约束内实现有效信息的组织和检索。虽然专门针对上下文管理作为一个统一领域的综合调查仍然有限，但关于内存系统和上下文压缩技术的相关工作为这些关键能力提供了基础见解。

系统实现 与此同时，文献广泛涵盖了将基础组件集成到解决实际应用需求的复杂架构中的**系统实现**。**RAG** 领域获得了大量关注，基础调查追溯了其发展和对缓解幻觉的影响 [311, 253, 1131]。最近的工作调查了向模块化、代理化和图增强RAG架构的演变 [162, 622, 120, 312, 1391]。

内存系统 通过关注内存增强代理及其应用的调查，探索了支持持久交互和认知架构的内存系统。更广泛的**基于LLM的代理** 类别是一个基础领域，包括对自主代理、其架构、规划和方法的全面概述 [1091, 719, 277, 843, 1340, 498, 1272]。

工具集成推理 涵盖函数调用机制和智能体-环境交互的内容已得到充分记录，探讨了从单工具系统到复杂编排框架的演变 [663, 858, 771, 867]。向 **多智能体系统（MAS）** 的演变是另一个焦点，相关调查详细介绍了MAS工作流程、基础设施、通信协议和协调机制 [625, 356, 246, 1235, 38, 503, 187, 458]。

评估 评估这些复杂系统的关键方面已得到全面审查，相关研究分析了用于评估组件级和系统级能力与性能的基准和方法 [1259, 380, 835, 310]。这项评估文献涵盖了基础组件评估和集成系统评估范式。

我们的贡献 尽管这些调查对其各自领域提供了不可或缺的深入分析，但它们本质上呈现了该领域的碎片化视图。RAG作为外部记忆的一种形式、工具使用作为获取上下文的方法以及提示工程作为编排这些组件的语言之间的联系往往被隐含地提及。我们的工作通过提出 上下文工程作为一种统一的抽象，明确地将基础组件与其在复杂实现中的集成分离开来。通过将这些不同的领域组织成一个单一的、连贯的分类法，本调查旨在阐明它们之间基本的关系，提供一个整体地图，说明上下文是如何被生成、处理、管理和利用来引导下一代智能系统。

3. 为什么需要上下文工程?

随着大型语言模型（LLMs）从简单的指令跟随系统演变为复杂、多方面的应用的核心推理引擎，与之交互的方法也必须演变。虽然“提示工程”是一个基础概念，但它已经不足以涵盖设计、管理和优化现代AI系统所需的信息有效载荷的全部范围。这些系统并不

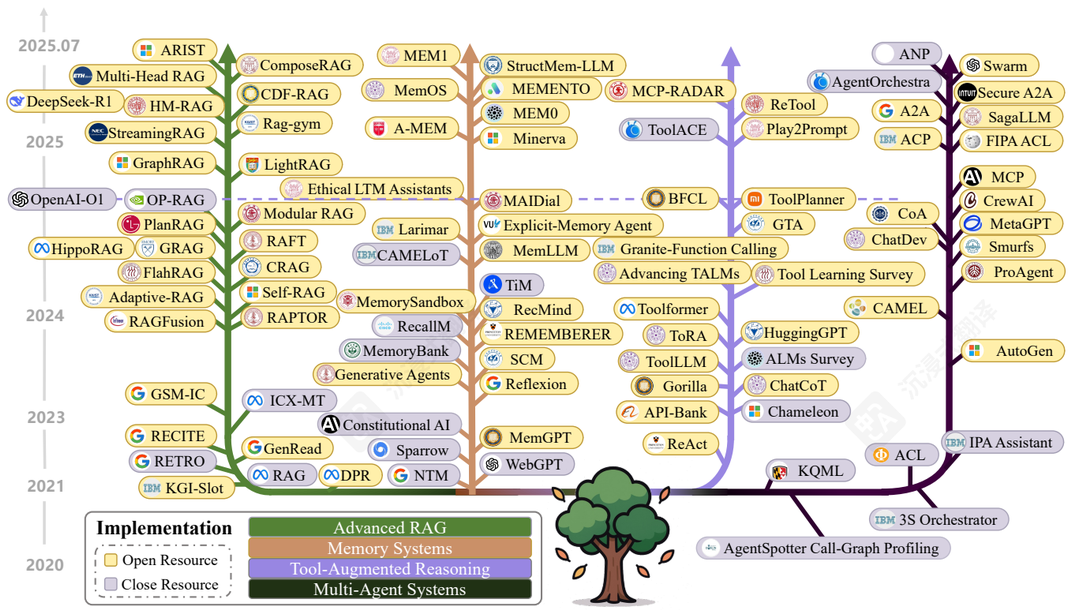

Figure 2: Context Engineering Evolution Timeline: A comprehensive visualization of the development trajectory of Context Engineering implementations from 2020 to 2025, showing the evolution from foundational RAG systems to sophisticated multi-agent architectures and tool-integrated reasoning systems.

operate on a single, static string of text; they leverage a dynamic, structured, and multifaceted information stream. To address this, we introduce and formalize the discipline of **Context Engineering**.

3.1. Definition of Context Engineering

To formally define Context Engineering, we begin with the standard probabilistic model of an autoregressive LLM. The model, parameterized by θ, generates an output sequence $Y = (y_1, \ldots, y_T)$ given an input context C by maximizing the conditional probability:

$$P_\theta(Y|C) = \prod_{t=1}^{T} P_\theta(y_t|y_{<t}, C) \tag{1}$$

Historically, in the paradigm of prompt engineering, the context C was treated as a monolithic, static string of text, i.e., $C =$ prompt. This view is insufficient for modern systems.

Context Engineering re-conceptualizes the context C as a dynamically structured set of informational components, $c_1, c_2, \ldots, c_n$. These components are sourced, filtered, and formatted by a set of functions, and finally orchestrated by a high-level assembly function, $\mathcal{A}$:

$$C = \mathcal{A}(c_1, c_2, \ldots, c_n) \tag{2}$$

The components c_i are not arbitrary; they map directly to the core technical domains of this survey:

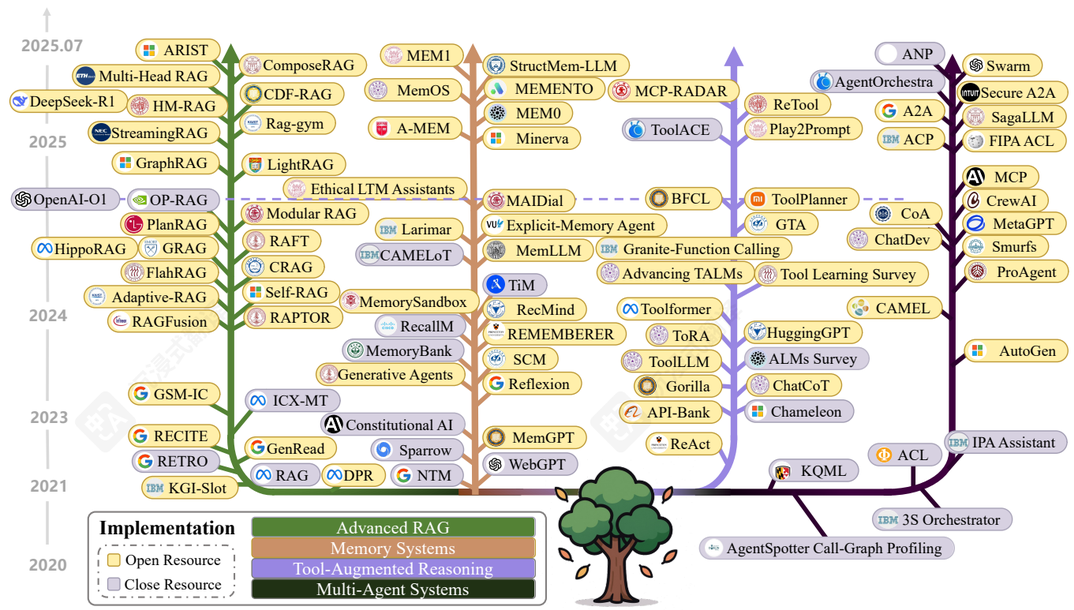

图 2：上下文工程演化时间线：对 2020 年至 2025 年上下文工程实现发展轨迹的综合可视化，展示了从基础 RAG 系统到复杂的多智能体架构和工具集成推理系统的演化。

它们操作的是单个、静态的文本字符串；它们利用的是动态的、结构化的、多方面的信息流。为了解决这个问题，我们引入并形式化了**上下文工程**这一学科。

3.1. 上下文工程的定义

为了形式化定义上下文工程，我们从自回归大语言模型的标准概率模型开始。该模型由 θ 参数化，给定输入上下文 C 时，通过最大化条件概率生成输出序列 $Y = (y_1, \ldots, y_T)$：

$$P_\theta(Y|C) = \prod_{t=1}^{T} P_\theta(y_t|y_{<t}, C) \tag{1}$$

历史上，在提示工程的范式下，上下文 C 被视为一个整体的、静态的文本字符串，即 $C =$ 提示。这种观点对于现代系统来说是不够的。

上下文工程将上下文 C 重新概念化为一个动态结构化的信息组件集合，$c_1, c_2, \ldots, c_n$。这些组件由一组函数进行提取、过滤和格式化，最后由一个高级组装函数 $\mathcal{A}$ 协调：

$$C = \mathcal{A}(c_1, c_2, \ldots, c_n) \tag{2}$$

这些组件 c_i 不是随意的；它们直接映射到这项调查的核心技术领域 y:

- c_{instr}: System instructions and rules (**Context Retrieval and Generation**, Sec. 4.1).
- c_{know}: External knowledge, retrieved via functions like RAG or from integrated knowledge graphs (**RAG**, Sec. 5.1; **Context Processing**, Sec. 4.2).
- c_{tools}: Definitions and signatures of available external tools (**Function Calling** & **Tool-Integrated Reasoning**, Sec. 5.3).
- c_{mem}: Persistent information from prior interactions (**Memory Systems**, Sec. 5.2; **Context Management**, Sec. 4.3).
- c_{state}: The dynamic state of the user, world, or multi-agent system (**Multi-Agent Systems** & **Orchestration**, Sec. 5.4).
- c_{query}: The user's immediate request.

The Optimization Problem of Context Engineering. From this perspective, Context Engineering is the formal optimization problem of finding the ideal set of context-generating functions (which we denote collectively as $\mathcal{F} = \{\mathcal{A}, \text{Retrieve}, \text{Select}, \dots\}$) that maximizes the expected quality of the LLM's output. Given a distribution of tasks $\mathcal{T}$, the objective is:

$$\mathcal{F}^* = \arg\max_{\mathcal{F}} \mathbb{E}_{\tau\sim\mathcal{T}}[\text{Reward}(P_\theta(Y|C_{\mathcal{F}}(\tau)), Y_\tau^*)] \tag{3}$$

where τ is a specific task instance, $C_{\mathcal{F}}(\tau)$ is the context generated by the functions in $\mathcal{F}$ for that task, and Y_τ^* is the ground-truth or ideal output. This optimization is subject to hard constraints, most notably the model's context length limit, $|C| \leq L_{\text{max}}$.

Mathematical Principles and Theoretical Frameworks. This formalization reveals deeper mathematical principles. The assembly function $\mathcal{A}$ is a form of **Dynamic Context Orchestration**, a pipeline of formatting and concatenation operations, $\mathcal{A} = \text{Concat} \circ (\text{Format}_1, \dots, \text{Format}_n)$, where each function must be optimized for the LLM's architectural biases (e.g., attention patterns).

The retrieval of knowledge, $c_{\text{know}} = \text{Retrieve}(\dots)$, can be framed as an **Information-Theoretic Optimality** problem. The goal is to select knowledge that maximizes the mutual information with the target answer Y^*, given the query c_{query}:

$$\text{Retrieve}^* = \arg\max_{\text{Retrieve}} I(Y^*; c_{\text{know}}|c_{\text{query}}) \tag{4}$$

This ensures that the retrieved context is not just semantically similar, but maximally informative for solving the task.

Furthermore, the entire process can be viewed through the lens of **Bayesian Context Inference**. Instead of deterministically constructing the context, we infer the optimal context posterior $P(C|c_{\text{query}}, \text{History}, \text{World})$. Using Bayes' theorem, this posterior is proportional to the likelihood of the query given the context and the prior probability of the context's relevance:

$$P(C|c_{\text{query}}, \dots) \propto P(c_{\text{query}}|C) \cdot P(C|\text{History}, \text{World}) \tag{5}$$

The decision-theoretic objective is then to find the context C^* that maximizes the expected reward over the distribution of possible answers:

$$C^* = \arg\max_{C} \int P(Y|C, c_{\text{query}}) \cdot \text{Reward}(Y, Y^*)\, dY \cdot P(C|c_{\text{query}}, \dots) \tag{6}$$

This Bayesian formulation provides a principled way to handle uncertainty, perform adaptive retrieval by updating priors, and maintain belief states over context in multi-step reasoning tasks.

- c_{instr}: 系统指令和规则（**上下文检索和生成**，第 4.1节）。
- c_{know}: 外部知识，通过 RAG 等函数检索或来自集成知识库 aphs (**RAG**，第 5.1节； **上下文处理**，第 4.2节）。
- c_{tools}: 可用外部工具的定义和签名（**函数调用** & **工具集成 ated 推理**，第 5.3节）。
- c_{mem}: 先前交互的持久信息（**记忆系统**，第 5.2节； **上下文管理 age-ment**, 第 4.3)
- $c_{\text{状态}}$ **state**: 用户、世界或多智能体系统的动态状态（**多智能体系统**& **Orche**
- c查询：用户的即时请求。

上下文工程的优化问题。从这个角度来看，上下文工程是寻找理想的一组上下文生成函数（我们统称为 $\mathcal{F} = \{\mathcal{A}, \text{Retrieve}, \text{Select}, \dots\}$）的正式优化问题，这些函数最大化了大语言模型输出的预期质量。给定一个任务分布 $\mathcal{T}$，目标是：

$$\mathcal{F}^* = \arg\max_{\mathcal{F}} \mathbb{E}_{\tau\sim\mathcal{T}}[\text{Reward}(P_\theta(Y|C_{\mathcal{F}}(\tau)), Y_\tau^*)] \tag{3}$$

where τ 是一个特定的任务实例，$C_{\mathcal{F}}(\tau)$ 是由 $\mathcal{F}$ 中的函数为该任务生成的上下文，而 Y_τ^* 是真实值或理想输出。此优化受硬约束约束，最值得注意的是模型的上下文长度限制，$|C| \leq L_{\text{max}}$。

数学原理和理论框架。这种形式化揭示了更深层次的数学原理。组装函数 $\mathcal{A}$ 是一种 **动态上下文编排**，一系列格式化和连接操作，$\mathcal{A} = \text{Concat} \circ (\text{Format}_1, \dots, \text{Format}_n)$，其中每个函数都必须针对LLM的架构偏差（例如，注意力模式）进行优化。

知识的检索，$c_{\text{know}} = \text{Retrieve}(\dots)$，可以被视为一个 **信息论最优性**问题。目标是在给定查询 c_{query} 的情况下，选择最大化与目标答案Y^*的互信息量的知识：

$$\text{Retrieve}^* = \arg\max_{\text{Retrieve}} I(Y^*; c_{\text{know}}|c_{\text{query}}) \tag{4}$$

这确保了检索到的上下文不仅语义相似，而且对于解决任务具有最大信息量。

此外，整个过程可以通过**贝叶斯上下文推理**的视角来理解。我们不是确定性地构建上下文，而是推断出最优的上下文后验概率$P(C|c_{\text{query}}, \text{History}, \text{World})$。使用贝叶斯定理，这个后验概率与给定上下文的查询似然以及上下文的先验相关性概率成正比：

$$P(C|c_{\text{query}}, \dots) \propto P(c_{\text{query}}|C) \cdot P(C|\text{History}, \text{World}) \tag{5}$$

那么，决策理论的目标就是找到最大化可能答案分布上预期奖励的上下文C^*：

$$C^* = \arg\max_{C} \int P(Y|C, c_{\text{query}}) \cdot \text{Reward}(Y, Y^*)\, dY \cdot P(C|c_{\text{query}}, \dots) \tag{6}$$

这种贝叶斯形式化提供了一种原则性的方法来处理不确定性，通过更新先验值进行自适应检索，并在多步推理任务中保持上下文信念状态。

Dimension	Prompt Engineering	Context Engineering
Model	$C =$ prompt (static string)	$C = \mathcal{A}(c_1, c_2, \ldots, c_n)$ (dynamic, structured assembly)
Target	$\arg\max_{\text{prompt}} P_\theta(Y\|\text{prompt})$	$\mathcal{F}^* = \arg\max_{\mathcal{F}} \mathbb{E}_{\tau\sim\mathcal{T}}[\text{Reward}(P_\theta(Y\|C_{\mathcal{F}}(\tau)), Y^*_\tau)]$
Complexity	Manual or automated search over a string space.	System-level optimization of $\mathcal{F} = \{\mathcal{A}, \text{Retrieve}, \text{Select}, \ldots\}$.
Information	Information content is fixed within the prompt.	Aims to maximize task-relevant information under constraint $\|C\| \leq L_{\max}$.
State	Primarily stateless.	Inherently stateful, with explicit components for c_{mem} and c_{state}.
Scalability	Brittleness increases with length and complexity.	Manages complexity through modular composition.
Error Analysis	Manual inspection and iterative refinement.	Systematic evaluation and debugging of individual context functions.

Table 1: Comparison of Prompt Engineering and Context Engineering Paradigms.

Comparison of Paradigms The formalization of Context Engineering highlights its fundamental distinctions from traditional prompt engineering. The following table summarizes the key differences.

In summary, Context Engineering provides the formal, systematic framework required to build, understand, and optimize the sophisticated, context-aware AI systems that are coming to define the future of the field. It shifts the focus from the "art" of prompt design to the "science" of information logistics and system optimization.

Context Scaling Context scaling encompasses two fundamental dimensions that collectively define the scope and sophistication of contextual information processing. The first dimension, **length scaling**, addresses the computational and architectural challenges of processing ultra-long sequences, extending context windows from thousands to millions of tokens while maintaining coherent understanding across extended narratives, documents, and interactions. This involves sophisticated attention mechanisms, memory management techniques, and architectural innovations that enable models to maintain contextual coherence over vastly extended input sequences.

The second, equally critical dimension is **multi-modal and structural scaling**, which expands context beyond simple text to encompass multi-dimensional, dynamic, cross-modal information structures. This includes temporal context (understanding time-dependent relationships and sequences), spatial context (interpreting location-based and geometric relationships), participant states (tracking multiple entities and their evolving conditions), intentional context (understanding goals, motivations, and implicit objectives), and cultural context (interpreting communication within specific social and cultural frameworks).

Modern context engineering must address both dimensions simultaneously, as real-world applications require models to process not only lengthy textual information but also diverse data types including structured knowledge graphs, multimodal inputs (text, images, audio, video), temporal sequences, and implicit contextual cues that humans naturally understand. This multi-dimensional approach to context scaling represents a fundamental shift from parameter scaling toward developing systems capable of understanding complex, ambiguous contexts that mirror the nuanced nature of human intelligence in facing a complex world [1036].

维度	提示工程	Context Engineering
模型	C = 提示（静态字符串）	$C = \mathcal{A}(c_1, c_2, \ldots, c_n)$ (动态，结构化组合)
目标	$\arg\max_{\text{prompt}} P_\theta(Y\|\text{prompt})$	$\mathcal{F}^* = \arg\max_{\mathcal{F}} \mathbb{E}_{\tau\sim\mathcal{T}}[\text{Reward}(P_\theta(Y\|C_{\mathcal{F}}(\tau)), Y^*_\tau)]$
Complexity	Manual or automated search over a string space.	$\mathcal{F} = \{\mathcal{A}$,Retrieve, }Select, ...
Information	Information content is fixed within the prompt.	旨在约束下最大化任务相关信息 $\|C\| \leq L_{\max}$.
状态	主要无状态。	Inherently stateful, with explicit components for c_{mem} and c_{state}.
可扩展性	脆性随长度和复杂度增加而增加。通过模块化组合管理复杂性。	
错误分析	人工检查和迭代优化。	对单个上下文功能的系统评估和调试 ns.

Table 1: Comparison of Prompt Engineering and Context Engineering Paradigms.

范式比较 上下文工程的正式化突出了它与传统提示工程的根本区别。下表总结了主要差异。

总之，上下文工程提供了构建、理解和优化复杂、上下文感知的AI系统的正式、系统化框架，这些系统正定义着该领域的未来。它将重点从提示设计的“艺术”转移到信息物流和系统优化的“科学”上。

上下文扩展 上下文扩展包含两个基本维度，这些维度共同定义了上下文信息处理的范围和复杂性。第一个维度，**长度扩展**，解决了处理超长序列的计算和架构挑战，将上下文窗口从数千扩展到数百万个标记，同时保持跨扩展叙事、文档和交互的连贯理解。这涉及复杂的注意力机制、内存管理技术和架构创新，使模型能够在大大扩展的输入序列中保持上下文连贯性。

第二个同样关键的维度是**多模态和结构扩展**，它将上下文扩展到简单的文本之外，以涵盖多维、动态、跨模态的信息结构。这包括时间上下文（理解时间依赖关系和序列）、空间上下文（解释基于位置和几何关系的关系）、参与者状态（跟踪多个实体及其演变条件）、意图上下文（理解目标、动机和隐含目标）和文化上下文（解释特定社会和文化框架内的交流）。

现代上下文工程必须同时处理这两个维度，因为实际应用要求模型不仅要处理大量的文本信息，还要处理包括结构化知识图谱、多模态输入（文本、图像、音频、视频）、时间序列以及人类自然理解的隐式上下文线索等多样化数据类型。这种多维度的上下文扩展方法代表了从参数扩展向开发能够理解复杂、模糊上下文的系统的根本性转变，这些系统能够反映人类在面对复杂世界时智能的微妙性 [1036]。

3.2. Why Context Engineering

3.2.1. Current Limitations

Large Language Models face critical technical barriers necessitating sophisticated context engineering approaches. The self-attention mechanism imposes quadratic computational and memory overhead as sequence length increases, creating substantial obstacles to processing extended contexts and significantly impacting real-world applications such as chatbots and code comprehension models [1017, 977]. Commercial deployment compounds these challenges through repeated context processing that introduces additional latency and token-based pricing costs [1017].

Beyond computational constraints, LLMs demonstrate concerning reliability issues including frequent hallucinations, unfaithfulness to input context, problematic sensitivity to input variations, and responses that appear syntactically correct while lacking semantic depth or coherence [951, 1279, 523].

The prompt engineering process presents methodological challenges through approximation-driven and subjective approaches that focus narrowly on task-specific optimization while neglecting individual LLM behavior [800]. Despite these challenges, prompt engineering remains critical for effective LLM utilization through precise and contextually rich prompts that reduce ambiguity and enhance response consistency [964].

3.2.2. Performance Enhancement

Context engineering delivers substantial performance improvements through techniques like retrieval-augmented generation and superposition prompting, achieving documented improvements including 18-fold enhancement in text navigation accuracy, 94% success rates, and significant gains from careful prompt construction and automatic optimization across specialized domains [267, 768, 681].

Structured prompting techniques, particularly chain-of-thought approaches, enable complex reasoning through intermediate steps while enhancing element-aware summarization capabilities that integrate fine-grained details from source documents [1138, 750, 1120]. Few-shot learning implementations through carefully selected demonstration examples yield substantial performance gains, including 9.90% improvements in BLEU-4 scores for code summarization and 175.96% in exact match metrics for bug fixing [306].

Domain-specific context engineering proves especially valuable in specialized applications, with execution-aware debugging frameworks achieving up to 9.8% performance improvements on code generation benchmarks and hardware design applications benefiting from specialized testbench generation and security property verification [1360, 873, 44]. These targeted approaches bridge the gap between general-purpose model training and specialized domain requirements.

3.2.3. Resource Optimization

Context engineering provides efficient alternatives to resource-intensive traditional approaches by enabling intelligent content filtering and direct knowledge transmission through carefully crafted prompts [630, 670]. LLMs can generate expected responses even when relevant information is deleted from input context, leveraging contextual clues and prior knowledge to optimize context length usage while maintaining response quality, particularly valuable in domains with significant data acquisition challenges [630, 670].

Specialized optimization techniques further enhance efficiency gains through context awareness and responsibility tuning that significantly reduce token consumption, dynamic context optimization employing

3.2. 为什么需要上下文工程

3.2.1. 当前局限性

大型语言模型面临关键的技术障碍，需要复杂的上下文工程方法。自注意力机制随着序列长度的增加导致计算和内存开销呈平方级增长，为处理长上下文创造了重大障碍，并显著影响聊天机器人和代码理解模型等实际应用 [1017, 977]。商业部署通过重复的上下文处理增加了额外的延迟和基于token的定价成本，进一步加剧了这些挑战 [1017]。

除了计算限制，LLMs还表现出令人担忧的可靠性问题，包括频繁的幻觉、对输入上下文的不忠实、对输入变化的敏感性问题，以及语法正确但缺乏语义深度或连贯性的响应 [951, 1279, 523]。

提示工程过程通过近似驱动和主观的方法提出了方法论挑战，这些方法狭隘地关注特定任务的优化，而忽略了单个LLM的行为 [800]。尽管存在这些挑战，提示工程通过精确且上下文丰富的提示仍然对有效的LLM利用至关重要，这些提示减少了歧义并增强了响应一致性[964]。

3.2.2. Performance Enhancement

上下文工程通过检索增强生成和叠加提示等技术，实现了显著的性能提升，取得了包括文本导航精度提升18倍、成功率达到94%以及通过精心构建的提示和跨专业领域的自动优化获得显著收益等已记录的改进成果 [267, 768, 681]。

结构化提示技术，特别是思维链方法，通过中间步骤实现复杂推理，同时增强了对源文档中细粒度细节的元素感知总结能力 [1138, 750, 1120]。通过精心选择的演示示例实现的少样本学习方法，带来了显著的性能提升，包括代码摘要的BLEU-4分数提升9.90%，以及漏洞修复的精确匹配指标提升175.96% [306]。

特定领域的上下文工程在专业应用中尤其有价值，执行感知调试框架在代码生成基准测试中实现了高达9.8%的性能提升，硬件设计应用则受益于专门的测试平台生成和安全属性验证 [1360, 873, 44]。这些有针对性的方法弥合了通用模型训练和特定领域需求之间的差距。

3.2.3. 资源优化

上下文工程通过智能内容过滤和精心设计的提示直接传递知识，为资源密集型传统方法提供了高效的替代方案 [630, 670]。LLM可以在输入上下文中删除相关信息时生成预期响应，利用上下文线索和先验知识优化上下文长度使用，同时保持响应质量，这在数据获取面临重大挑战的领域尤为宝贵 [630, 670]。

专门的优化技术通过上下文感知和责任调整进一步提高了效率，显著减少了token消耗，动态上下文优化采用

precise token-level content selection, and attention steering mechanisms for long-context inference [426, 944, 350]. These approaches maximize information density while reducing processing overhead and maintaining performance quality [944, 350].

3.2.4. *Future Potential*

Context engineering enables flexible adaptation mechanisms through in-context learning that allows models to adapt to new tasks without explicit retraining, with context window size directly influencing available examples for task adaptation [617]. Advanced techniques integrate compression and selection mechanisms for efficient model editing while maintaining contextual coherence [619]. This adaptability proves especially valuable in low-resource scenarios, enabling effective utilization across various prompt engineering techniques including zero-shot approaches, few-shot examples, and role context without requiring domain-specific fine-tuning [924, 129, 1075].

Sophisticated context engineering techniques including in-context learning, chain-of-thought, tree-of-thought, and planning approaches establish foundations for nuanced language understanding and generation capabilities while optimizing retrieval and generation processes for robust, context-aware AI applications [797, 974].

Future research directions indicate substantial potential for advancing context-sensitive applications through chain-of-thought augmentation with logit contrast mechanisms [953], better leveraging different context types across domains, particularly in code intelligence tasks combining syntax, semantics, execution flow, and documentation [1094], and understanding optimal context utilization strategies as advanced language models continue demonstrating prompt engineering's persistent value [1079]. Evolution toward sophisticated filtering and selection mechanisms represents a critical pathway for addressing transformer architectures' scaling limitations while maintaining performance quality.

4. Foundational Components

Context Engineering is built upon three fundamental components that collectively address the core challenges of information management in large language models: **Context Retrieval and Generation** sources appropriate contextual information through prompt engineering, external knowledge retrieval, and dynamic context assembly; **Context Processing** transforms and optimizes acquired information through long sequence processing, self-refinement mechanisms, and structured data integration; and **Context Management** tackles efficient organization and utilization of contextual information through addressing fundamental constraints, implementing sophisticated memory hierarchies, and developing compression techniques. These foundational components establish the theoretical and practical basis for all context engineering implementations, forming a comprehensive framework where each component addresses distinct aspects of the context engineering pipeline while maintaining synergistic relationships that enable comprehensive contextual optimization and effective context engineering strategies.

4.1. Context Retrieval and Generation

Context Retrieval and Generation forms the foundational layer of context engineering, encompassing the systematic retrieval and construction of relevant information for LLMs. This component addresses the critical challenge of sourcing appropriate contextual information through three primary mechanisms: prompt-based generation that crafts effective instructions and reasoning frameworks, external knowledge retrieval that

精确的token级内容选择，以及用于长上下文推理的注意力引导机制 [426, 944,350]。这些方法在最大化信息密度的同时减少处理开销并保持性能质量 [944, 350]。

3.2.4. 未来潜力

上下文工程通过情境学习使模型能够通过灵活的适应机制适应新任务，而无需显式重新训练，上下文窗口大小直接影响可用于任务适应的示例 [617]。高级技术集成了压缩和选择机制，以高效地编辑模型，同时保持上下文连贯性 [619]。这种适应性在低资源场景中尤其有价值，能够有效利用各种提示工程技术，包括零样本方法、少样本示例和角色上下文，而无需特定领域的微调 [924, 129, 1075]。

复杂的上下文工程技术，包括情境学习、思维链、思维树和规划方法，为精细的语言理解和生成能力奠定了基础，同时优化了检索和生成过程，以实现稳健的、上下文感知的AI应用[797, 974]。

未来的研究方向表明，通过思维链增强和logit对比机制 [953],，在跨领域更好地利用不同类型上下文，特别是在结合语法、语义、执行流程和文档的代码智能任务中，对于推进上下文敏感应用具有巨大潜力 [1094], ，并且随着先进语言模型继续展示提示工程的持久价值 [1079]，理解最佳上下文利用策略至关重要。向复杂的过滤和选择机制发展是解决Transformer架构扩展限制并保持性能质量的关键途径。

4. 基础组件

上下文工程建立在三个基本组件之上，这些组件共同应对大型语言模型中信息管理的核心挑战：**上下文检索和生成**通过提示工程、外部知识检索和动态上下文组装获取适当的上下文信息；**上下文处理**通过长序列处理、自完善机制和结构化数据集成转换和优化获取的信息；以及**上下文管理**通过解决基本约束、实施复杂的内存层次结构和开发压缩技术来处理上下文信息的有效组织和利用。这些基础组件为所有上下文工程实现奠定了理论和实践基础，形成一个综合框架，其中每个组件解决上下文工程管道的不同方面，同时保持协同关系，以实现全面的上下文优化和有效的上下文工程策略。

4.1. 上下文检索与生成

上下文检索与生成构成了上下文工程的基础层，涵盖了为大型语言模型系统性地检索和构建相关信息的任务。该组件通过三种主要机制解决获取适当上下文信息的关键挑战：基于提示的生成机制，用于制作有效的指令和推理框架；外部知识检索机制，用于

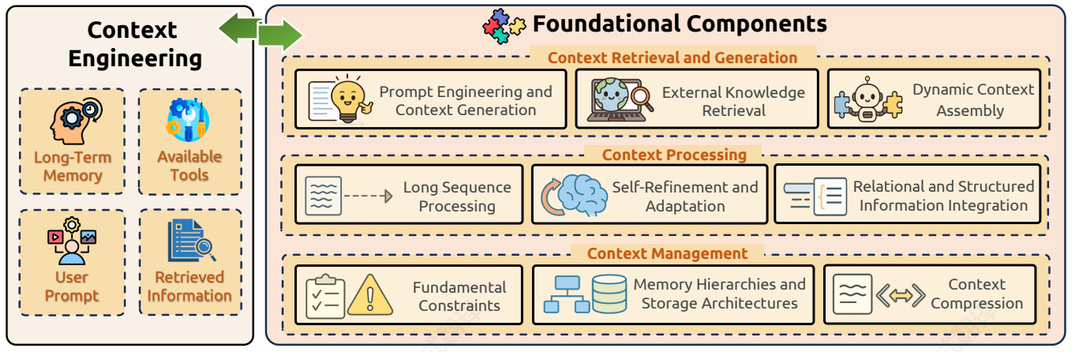

Figure 3: Context Engineering Framework: A comprehensive taxonomy of Context Engineering components including Context Retrieval and Generation, Context Processing, and Context Management, integrated into System Implementations such as RAG systems, memory architectures, tool-integrated reasoning, and multi-agent coordination mechanisms.

accesses dynamic information sources, and dynamic context assembly that orchestrates acquired components into coherent, task-optimized contexts.

4.1.1. Prompt Engineering and Context Generation

Prompt engineering and context generation forms the foundational layer of context retrieval, encompassing strategic input design that combines art and science to craft effective instructions for LLMs. The CLEAR Framework—conciseness, logic, explicitness, adaptability, and reflectiveness—governs effective prompt construction, while core architecture integrates task instructions, contextual information, input data, and output indicators [702, 1133, 569, 209, 25].

Zero-Shot and Few-Shot Learning Paradigms Zero-shot prompting enables task performance without prior examples, relying exclusively on instruction clarity and pre-trained knowledge [1361, 336, 553, 67, 1046]. Few-shot prompting extends this capability by incorporating limited exemplars to guide model responses, demonstrating task execution through strategic example selection [1361, 401, 103, 546, 788, 1371]. In-context learning facilitates adaptation to novel tasks without parameter updates by leveraging demonstration examples within prompts, with performance significantly influenced by example selection and ordering strategies [365, 103, 1287, 1016, 920, 846, 1139, 348, 576].

Chain-of-Thought Foundations Chain-of-Thought (CoT) prompting decomposes complex problems into intermediate reasoning steps, mirroring human cognition [1138, 401, 336, 939, 603]. Zero-shot CoT uses trigger phrases like "Let's think step by step," improving MultiArith accuracy from 17.7% to 78.7% [553, 1099, 472, 662], with Automatic Prompt Engineer refinements yielding additional gains [1215, 526].

Tree-of-Thoughts (ToT) organizes reasoning as hierarchical structures with exploration, lookahead, and backtracking capabilities, increasing Game of 24 success rates from 4% to 74% [1246, 217, 557, 598]. Graph-of-Thoughts (GoT) models reasoning as arbitrary graphs with thoughts as vertices and dependencies as edges, improving quality by 62% and reducing costs by 31% compared to ToT [69, 826, 1366].

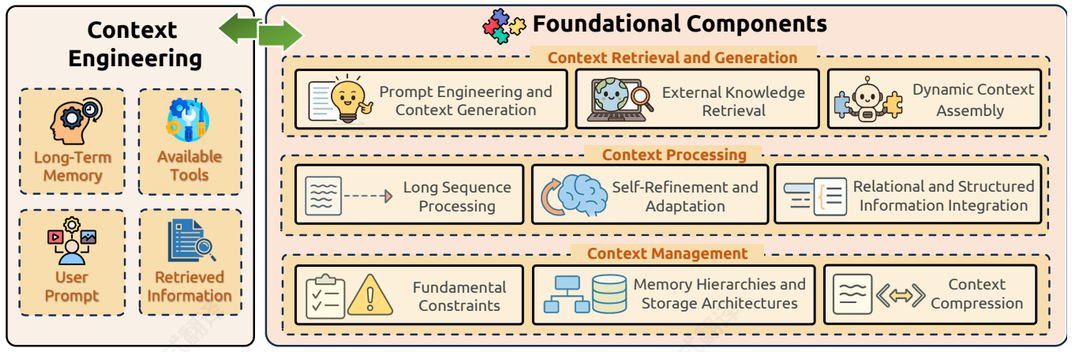

图 3：上下文工程框架：一个全面的上下文工程组件分类，包括上下文检索和生成、上下文处理以及上下文管理，集成到系统实现中，如RAG系统、内存架构、工具集成推理和多智能体协调机制。

acce动态信息源，以及动态上下文组装，协调获取的组件，形成任务优化的上下文。 s into

4.1.1. 提示工程和上下文生成

提示工程和上下文生成构成了上下文检索的基础层，包括结合艺术和科学的战略输入设计，以制作有效的LLM指令。CLEAR框架——简洁性、逻辑性、明确性、适应性和反思性——指导有效的提示构建，而核心架构集成了任务指令、上下文信息、输入数据和输出指标 [702, 1133, 569, 209, 25]。

零样本和少样本学习范式 零样本提示能够在没有先例的情况下执行任务，完全依赖于指令清晰度和预训练知识 [1361, 336, 553, 67, 1046]。少样本提示通过结合有限的示例来扩展这种能力，以指导模型响应，通过战略示例选择展示任务执行 [1361, 401, 103, 546, 788, 1371]。上下文学习通过在提示中利用演示示例，在不更新参数的情况下适应新任务，其性能显著受示例选择和排序策略的影响 [365, 103, 1287, 1016, 920, 846, 1139, 348, 576]。

思维链基础 思维链（CoT）提示将复杂问题分解为中间推理步骤，模拟人类认知 [1138, 401, 336, 939, 603]。零样本思维链使用"让我们一步步思考"等触发短语，在自动提示工程师改进下将MultiArith准确率从17.7%提高到78.7%[553, 1099, 472, 662], ，并带来额外增益 [1215, 526]。

思维树（ToT）将推理组织为具有探索、前瞻和回溯能力的层次结构，将24点游戏成功率从4%提高到74% [1246, 217, 557, 598]。思维图（GoT）将推理建模为任意图，其中思想作为顶点，依赖关系作为边，与ToT相比，质量提高62%，成本降低31% [69, 826, 1366]。

Cognitive Architecture Integration Cognitive prompting implements structured human-like operations including goal clarification, decomposition, filtering, abstraction, and pattern recognition, enabling systematic multi-step task resolution through deterministic, self-adaptive, and hybrid variants [558, 557, 1205, 1164]. Guilford's Structure of Intellect model provides psychological foundations for categorizing cognitive operations such as pattern recognition, memory retrieval, and evaluation, enhancing reasoning clarity, coherence, and adaptability [556, 191]. Advanced implementations incorporate cognitive tools as modular reasoning operations, with GPT-4.1 performance on AIME2024 increasing from 26.7% to 43.3% through structured cognitive operation sequences [243, 1030].

Method	Description
Self-Refine [735, 916]	Enables LLMs to improve outputs through iterative feedback and refinement cycles using the same model as the generator, feedback provider, and refiner, without supervised training.
Multi-Aspect Feedback [799]	Integrates multiple feedback modules (frozen LMs and external tools), each focusing on specific error categories to enable more comprehensive, independent evaluation.
N-CRITICS [789]	Implements an ensemble of critics that evaluate an initial output. Compiled feedback from the generating LLM and other models guides refinement until a stopping criterion is met.
ISR-LLM [1373]	Improves LLM-based planning by translating natural language to formal specifications, creating an initial plan, and then systematically refining it with a validator.
SELF [704]	Teaches LLMs meta-skills (self-feedback, self-refinement) with limited examples, then has the model continuously self-evolve by generating and filtering its own training data.
ProMiSe [884]	Addresses self-refinement in smaller LMs using principle-guided iterative refinement, combining proxy metric thresholds with few-shot refinement and rejection sampling.
A2R [577]	Augments LLMs through Metric-based Iterative Feedback Learning, using explicit evaluation across multiple dimensions (e.g., correctness) to generate feedback and refine outputs.
Experience Refinement [857]	Enables LLM agents to refine experiences during task execution by learning from recent (successive) or all previous (cumulative) experiences, prioritizing high-quality ones.
I-SHEEP [654]	Allows LLMs to continuously self-align from scratch by generating, assessing, filtering, and training on high-quality synthetic datasets without external guidance.
CaP [1271]	Uses external tools to refine chain-of-thought (CoT) responses, addressing the limitation of models that get stuck in non-correcting reasoning loops.
Agent-R [1277]	Enables language agents to reflect "on the fly" through iterative self-training, using Monte Carlo Tree Search (MCTS) to construct training data that corrects erroneous paths.
GenDiE [610]	Enhances context faithfulness with sentence-level optimization, combining generative and discriminative training to give LLMs self-generation and self-scoring capabilities.
Self-Developing [466]	Enables LLMs to autonomously discover, implement, and refine their own improvement algorithms by generating them as code, evaluating them, and using DPO to recursively improve.
SR-NLE [1121]	Improves the faithfulness of post-hoc natural language explanations via an iterative critique and refinement process using self-feedback and feature attribution.

Table 2: Self-refinement methods in large language models and their key characteristics.

4.1.2. External Knowledge Retrieval

External knowledge retrieval represents a critical component of context retrieval, addressing fundamental limitations of parametric knowledge through dynamic access to external information sources including databases, knowledge graphs, and document collections.

Retrieval-Augmented Generation Fundamentals RAG combines parametric knowledge stored in model parameters with non-parametric information retrieved from external sources, enabling access to current, domain-specific knowledge while maintaining parameter efficiency [591, 311, 253]. FlashRAG provides comprehensive evaluation and modular implementation of RAG systems, while frameworks like KRAGEN

认知架构集成 认知提示实现结构化类人操作，包括目标澄清、分解、过滤、抽象和模式识别，通过确定性、自适应和混合变体实现系统化多步任务解决 [558, 557, 1205, 1164]。吉尔福德的智力结构模型为分类认知操作（如模式识别、记忆检索和评估）提供了心理基础，增强了推理的清晰度、连贯性和适应性 [556, 191]。高级实现将认知工具作为模块化推理操作，通过结构化认知操作序列使 GPT-4.1 在 AIME2024 上的表现从 26.7% 提升至 43.3% [243, 1030]。

方法	描述
自我优化 [735, 916]	使 LLMs to improve outputs through iterative feedback and refinement cycles using the same model as the generator, 提供反馈并提供细化，无需监督训练。
多方面反馈 [799]	整合多个反馈模块（冻结的LLM和外 nal tools), each focusing on specific error categories to enable 更全面、独立的评估。
N-CRITICS [789]	实现了一组评价初始输出的评论家。收集来自生成LLM的反馈 d other 模型指导改进，直到满足停止标准。
ISR-LLM [1373]	通过将自然语言翻译来改进基于LLM的规划 age to formal specifications, creating an initial plan, and then systematically refining it with a validator.
SELF [704]	使用有限的示例向LLM教授元技能（自我反馈、自我完善），然后让模型继续 sly self- 通过生成并过滤其自身的训练数据来进化。
ProMiSe [884]	使用原则解决较小语言模型中的自我完善问题 -guided iterative refinement, combining proxy metric thresholds 结合少量样本细化和拒绝采样。
A2R [577]	通过基于指标的迭代反馈增强语言模型 edback Learning, using explicit evaluation across multiple dimensions (e.g., correctness) to generate feedback and refine outputs.
Experience Refinement [857]	Enables LLM agents to refine experiences during task execution by learning from recent (successive) or all previous (累积)经验，优先考虑高质量的经验。
I-SHEEP [654]	允许 LLM 从头开始通过 generating, assessing, filtering, and training on high-quality 合成数据集而不需要外部指导。
CaP [1271]	使用外部工具来改进思维链 t (CoT) responses, addressing the limitation of models that get stuck in 非纠正推理循环。
Agent-R [1277]	使语言代理能够反映 ct "on the fly" through iterative self-training, using Monte Carlo Tree Search (MCTS) to 构建校正错误路径的训练数据。
GenDiE [610]	Enhances context faithfulness with sentence-level optimization, combining generative and discri最小化训练 to give LLMs 自我生成和自我评分能力。
自我发展 [466]	Enables LLMs to autonomously discover, implement, and refine their own improvement algorithms by generating them as 代码，评估它们，并使用 DPO 递归改进。
SR-NLE [1121]	提高了后验自然语言解释的忠实度 ations via an iterative critique and refinement process using 自我反馈和特征归因。

Table 2: Self-refinement methods in large language models and their key characteristics.

4.1.2. 外部知识检索

外部知识检索代表了上下文检索的一个关键组成部分，通过动态访问外部信息源（包括数据库、知识图谱和文档集合）来解决参数化知识的根本局限性。

检索增强生成基础 RAG结合模型参数中存储的参数化知识以及从外部来源检索的非参数化信息，能够在保持参数效率的同时访问当前、特定领域的知识 [591, 311, 253]。FlashRAG提供了对RAG系统的全面评估和模块化实现，而KRAGEN等框架

and ComposeRAG demonstrate advanced retrieval strategies with substantial performance improvements across diverse benchmarks [500, 749, 1159].

Self-RAG introduces adaptive retrieval mechanisms where models dynamically decide when to retrieve information and generate special tokens to control retrieval timing and quality assessment [41]. Advanced implementations include RAPTOR for hierarchical document processing, HippoRAG for memory-inspired retrieval architectures, and Graph-Enhanced RAG systems that leverage structured knowledge representations for improved information access [928, 366, 360].

Knowledge Graph Integration and Structured Retrieval Knowledge graph integration addresses structured information retrieval through frameworks like KAPING, which retrieves relevant facts based on semantic similarities and prepends them to prompts without requiring model training [48, 673]. KARPA provides training-free knowledge graph adaptation through pre-planning, semantic matching, and relation path reasoning, achieving state-of-the-art performance on knowledge graph question answering tasks [258].

Think-on-Graph enables sequential reasoning over knowledge graphs to locate relevant triples, conducting exploration to retrieve related information from external databases while generating multiple reasoning pathways [1000, 720]. StructGPT implements iterative reading-then-reasoning approaches that construct specialized functions to collect relevant evidence from structured data sources [489].

Agentic and Modular Retrieval Systems Agentic RAG systems treat retrieval as dynamic operations where agents function as intelligent investigators analyzing content and cross-referencing information [648, 162, 965]. These systems incorporate sophisticated planning and reflection mechanisms requiring integration of task decomposition, multi-plan selection, and iterative refinement capabilities [438, 1183].

Modular RAG architectures enable flexible composition of retrieval components through standardized interfaces and plug-and-play designs. Graph-Enhanced RAG systems leverage structured knowledge representations for improved information access, while Real-time RAG implementations address dynamic information requirements in streaming applications [312, 1391].

4.1.3. Dynamic Context Assembly

Dynamic context assembly represents the sophisticated orchestration of acquired information components into coherent, task-optimized contexts that maximize language model performance while respecting computational constraints.

Assembly Functions and Orchestration Mechanisms The assembly function $\mathcal{A}$ encompasses template-based formatting, priority-based selection, and adaptive composition strategies that must adapt to varying task requirements, model capabilities, and resource constraints [702, 1133, 569]. Contemporary orchestration mechanisms manage agent selection, context distribution, and interaction flow control in multi-agent systems, enabling effective cooperation through user input processing, contextual distribution, and optimal agent selection based on capability assessment [894, 53, 171].

Advanced orchestration frameworks incorporate intent recognition, contextual memory maintenance, and task dispatching components for intelligent coordination across domain-specific agents. The Swarm Agent framework utilizes real-time outputs to direct tool invocations while addressing limitations in static tool registries and bespoke communication frameworks [808, 263, 246].

和 ComposeRAG 展示了先进的检索策略，并在各种基准测试中显著提升了性能 [500, 749, 1159]。

Self-RAG 引入了自适应检索机制，其中模型动态决定何时检索信息，并生成特殊标记来控制检索时机和质量评估 [41]。高级实现包括用于分层文档处理的 RAPTOR、受记忆启发的检索架构 HippoRAG，以及利用结构化知识表示以改进信息访问的 Graph-Enhanced RAG 系统 [928, 366, 360]。

知识图谱集成和结构化检索 知识图谱集成通过 KAPING 等框架解决结构化信息检索问题，该框架根据语义相似性检索相关事实，并将它们添加到提示中，而无需模型训练 [48, 673]。KARPA 通过预规划、语义匹配和关系路径推理提供无训练的知识图谱适应，在知识图谱问答任务上达到最先进性能 [258]。

Think-on-Graph 实现了知识图谱上的顺序推理以定位相关三元组，进行探索以从外部数据库检索相关信息，同时生成多个推理路径 [1000, 720]。StructGPT 实施了迭代阅读-推理方法，构建专用函数从结构化数据源收集相关证据 [489]。

代理式和模块化检索系统 代理式 RAG 系统将检索视为动态操作，其中代理充当智能调查员分析内容并交叉引用信息[648, 162, 965]。这些系统结合了复杂的规划和反思机制，需要集成任务分解、多计划选择和迭代优化能力 [438, 1183]。

模块化 RAG 架构通过标准接口和即插即用设计，支持检索组件的灵活组合。图增强 RAG 系统利用结构化知识表示来改进信息访问，而实时 RAG 实现则解决流式应用中的动态信息需求 [312, 1391]。

4.1.3. 动态上下文组装

动态上下文组装代表了对获取的信息组件进行复杂编排，以形成连贯、任务优化的上下文，从而在尊重计算约束的同时最大化语言模型性能。

组装函数和编排机制 组装函数 $\mathcal{A}$ 包括基于模板的格式化、基于优先级的选中和自适应组合策略，这些策略必须适应不同的任务需求、模型能力和资源约束 [702, 1133, 569]。当前的编排机制管理多代理系统中的代理选择、上下文分配和交互流控制，通过用户输入处理、上下文分配和基于能力评估的最优代理选择，实现有效协作 [894, 53, 171]。

高级编排框架集成了意图识别、上下文记忆维护和任务调度组件，以实现跨特定领域代理的智能协调。Swarm Agent 框架利用实时输出来指导工具调用，同时解决静态工具注册表和定制化通信框架的局限性 [808, 263, 246]。

Multi-Component Integration Strategies Context assembly must address cross-modal integration challenges, incorporating diverse data types including text, structured knowledge, temporal sequences, and external tool interfaces while maintaining coherent semantic relationships [529, 1221, 496]. Verbalization techniques convert structured data including knowledge graph triples, table rows, and database records into natural language sentences, enabling seamless integration with existing language systems without architectural modifications [12, 782, 1064, 13].

Programming language representations of structured data, particularly Python implementations for knowledge graphs and SQL for databases, outperform traditional natural language representations in complex reasoning tasks by leveraging inherent structural properties [1166]. Multi-level structurization approaches reorganize input text into layered structures based on linguistic relationships, while structured data representations leverage existing LLMs to extract structured information and represent key elements as graphs, tables, or relational schemas [681, 1125, 1324].

Automated Assembly Optimization Automated prompt engineering addresses manual optimization limitations through systematic prompt generation and refinement algorithms. Automatic Prompt Engineer (APE) employs search algorithms for optimal prompt discovery, while LM-BFF introduces automated pipelines combining prompt-based fine-tuning with dynamic demonstration incorporation, achieving up to 30% absolute improvement across NLP tasks [307, 417, 590]. Promptbreeder implements self-referential evolutionary systems where LLMs improve both task-prompts and mutation-prompts governing these improvements through natural selection analogies [275, 508].

Self-refine enables iterative output improvement through self-critique and revision across multiple iterations, with GPT-4 achieving approximately 20% absolute performance improvement through this methodology [735, 670]. Multi-agent collaborative frameworks simulate specialized team dynamics with agents assuming distinct roles (analysts, coders, testers), resulting in 29.9-47.1% relative improvement in Pass@1 metrics compared to single-agent approaches [434, 1257].

Tool integration frameworks combine Chain-of-Thought reasoning with external tool execution, automating intermediate reasoning step generation as executable programs strategically incorporating external data. LangChain provides comprehensive framework support for sequential processing chains, agent development, and web browsing capabilities, while specialized frameworks like Auto-GPT and Microsoft's AutoGen facilitate complex AI agent development through user-friendly interfaces [963, 1087, 25, 867].

4.2. Context Processing

Context Processing focuses on transforming and optimizing acquired contextual information to maximize its utility for LLMs. This component addresses challenges in handling ultra-long sequence contexts, enables iterative self-refinement and adaptation mechanisms, and facilitates integration of multimodal, relational and structured information into coherent contextual representations.

4.2.1. Long Context Processing

Ultra-long sequence context processing addresses fundamental computational challenges arising from transformer self-attention's $O(n^2)$ complexity, which creates significant bottlenecks as sequence lengths increase and substantially impacts real-world applications [1059, 731, 295, 268, 416]. Increasing Mistral-7B input from 4K to 128K tokens requires 122-fold computational increase, while memory constraints during

多组件集成策略 上下文组装必须解决跨模态集成挑战，整合多种数据类型，包括文本、结构化知识、时间序列和外部工具接口，同时保持连贯的语义关系 [529, 1221, 496]。语言化技术将结构化数据（包括知识图谱三元组、表格行和数据库记录）转换为自然语言句子，使无缝集成现有语言系统，无需架构修改 [12, 782, 1064, 13]。

结构化数据的编程语言表示，特别是用于知识图谱的 Python 实现和用于数据库的 SQL，通过利用固有的结构属性 [1166]，在复杂推理任务中优于传统的自然语言表示。多级结构化方法根据语言关系重组输入文本为分层结构，而结构化数据表示利用现有的 LLM 提取结构化信息，并将关键元素表示为图、表格或关系模式 [681, 1125, 1324]。

自动化组装优化 自动化提示工程通过系统化提示生成和优化算法解决手动优化限制。自动提示工程师（APE）采用搜索算法进行最优提示发现，而 LM-BFF 引入结合基于提示的微调和动态演示整合的自动化管道，在 NLP 任务中实现高达 30% 的绝对改进 [307, 417, 590]。Promptbreeder 实现自指的进化系统，其中 LLM 通过自然选择类比改进任务提示和治理这些改进的变异提示 [275, 508]。

Self-refine 通过自我批评和多轮迭代中的修订实现输出改进，GPT-4 通过此方法实现了约 20% 的绝对性能提升 [735, 670]。多智能体协作框架模拟了专业团队动态，智能体承担不同角色（分析师、编码员、测试员），与单智能体方法相比，Pass@1 指标相对提升了 29.9-47.1% [434, 1257]。

工具集成框架结合思维链推理与外部工具执行，将中间推理步骤自动生成可执行程序，策略性地整合外部数据。LangChain 提供了全面的框架支持，用于顺序处理链、智能体开发和网络浏览功能，而 Auto-GPT 和微软的 AutoGen 等专业框架通过用户友好的界面促进了复杂 AI 智能体开发 [963, 1087, 25, 867]。

4.2. 上下文处理

上下文处理专注于转换和优化获取的上下文信息，以最大化其对大型语言模型（LLM）的效用。该组件解决了处理超长序列上下文中的挑战，支持迭代自我完善和自适应机制，并促进多模态、关系和结构化信息整合为连贯的上下文表示。

4.2.1. 长上下文处理

超长序列上下文处理解决了由 Transformer 自注意力机制 $O(n_2)$ 复杂度引起的根本性计算挑战，随着序列长度的增加，这会创建显著的瓶颈，并严重影响实际应用 [1059, 731, 295, 268, 416]。将 Mistral-7B 的输入从 4K 增加到 128K 个 token 需要 122 倍的计算增加，而内存限制在

prefilling and decoding stages create substantial resource demands, with Llama 3.1 8B requiring up to 16GB per 128K-token request [1032, 1227, 425].

Architectural Innovations for Long Context State Space Models (SSMs) maintain linear computational complexity and constant memory requirements through fixed-size hidden states, with models like Mamba offering efficient recurrent computation mechanisms that scale more effectively than traditional transformers [1258, 347, 346]. Dilated attention approaches like LongNet employ exponentially expanding attentive fields as token distance grows, achieving linear computational complexity while maintaining logarithmic dependency between tokens, enabling processing of sequences exceeding one billion tokens [216].

Toeplitz Neural Networks (TNNs) model sequences with relative position encoded Toeplitz matrices, reducing space-time complexity to log-linear and enabling extrapolation from 512 training tokens to 14,000 inference tokens [868, 869]. Linear attention mechanisms reduce complexity from $O(N^2)$ to $O(N)$ by expressing self-attention as linear dot-products of kernel feature maps, achieving up to 4000× speedup when processing very long sequences [522]. Alternative approaches like non-attention LLMs break quadratic barriers by employing recursive memory transformers and other architectural innovations [547].

Position Interpolation and Context Extension Position interpolation techniques enable models to process sequences beyond original context window limitations by intelligently rescaling position indices rather than extrapolating to unseen positions [150]. Neural Tangent Kernel (NTK) approaches provide mathematically grounded frameworks for context extension, with YaRN combining NTK interpolation with linear interpolation and attention distribution correction [833, 471, 1021].

LongRoPE achieves 2048K token context windows through two-stage approaches: first fine-tuning models to 256K length, then conducting positional interpolation to reach maximum context length [218]. Position Sequence Tuning (PoSE) demonstrates impressive sequence length extensions up to 128K tokens by combining multiple positional interpolation strategies [1377]. Self-Extend techniques enable LLMs to process long contexts without fine-tuning by employing bi-level attention strategies—grouped attention and neighbor attention—to capture dependencies among distant and adjacent tokens [499].

Optimization Techniques for Efficient Processing Grouped-Query Attention (GQA) partitions query heads into groups that share key and value heads, striking a balance between multi-query attention and multi-head attention while reducing memory requirements during decoding [16, 1341]. FlashAttention exploits asymmetric GPU memory hierarchy to achieve linear memory scaling instead of quadratic requirements, with FlashAttention-2 providing approximately twice the speed through reduced non-matrix multiplication operations and optimized work distribution [196, 195].

Ring Attention with Blockwise Transformers enables handling extremely long sequences by distributing computation across multiple devices, leveraging blockwise computation while overlapping communication with attention computation [676]. Sparse attention techniques include Shifted sparse attention (S^2-Attn) in LongLoRA and SinkLoRA with SF-Attn, which achieve 92% of full attention perplexity improvement with significant computation savings [1304, 1217].

Efficient Selective Attention (ESA) proposes token-level selection of critical information through query and key vector compression into lower-dimensional representations, enabling processing of sequences up to 256K tokens [1084]. BigBird combines local attention with global tokens that attend to entire sequences,

预填充和解码阶段会消耗大量资源，Llama 3.1 8B 每个包含 128K 个 token 的请求需要高达 16GB 的资源 [1032, 1227, 425]。

用于长上下文的状态空间模型（SSMs）通过固定大小的隐藏状态保持线性计算复杂度和恒定的内存需求，像 Mamba 这样的模型提供了高效的循环计算机制，其扩展效果比传统 Transformer 更好 [1258, 347, 346]。扩张注意力方法如 LongNet 随着token距离的增长，采用指数扩展的注意力字段，实现线性计算复杂度，同时保持 token 之间的对数依赖关系，能够处理超过十亿个 token 的序列 [216]。

Toeplitz 神经网络（TNNs）使用相对位置编码的 Toeplitz 矩阵对序列进行建模，将时空复杂度降低到对数线性，并能够从 512 个训练 token 外推到 14,000 个推理 token [868, 869]。线性注意力机制通过将自注意力表达为核特征图的线性点积，将复杂度从 $O(N_2)$ 降低到 $O(N)$，在处理非常长的序列时能够实现高达 4000× 的加速 [522]。其他方法如非注意力 LLMs 通过采用递归内存 Transformer 和其他架构创新打破二次方壁垒 [547]。

位置插值和上下文扩展 位置插值技术使模型能够通过智能地重新缩放位置索引来处理超出原始上下文窗口限制的序列，而不是推断到未见过的位置 [150]。神经切线核 (NTK) 方法为上下文扩展提供了数学基础框架，YaRN 结合 NTK 插值与线性插值和注意力分布校正 [833, 471, 1021]。

LongRoPE 通过两阶段方法实现 2048K 令牌上下文窗口：首先将模型微调到 256K 长度，然后进行位置插值以达到最大上下文长度 [218]。位置序列调整 (PoSE) 通过结合多种位置插值策略，展示了令人印象深刻的序列长度扩展，最高可达 128K 令牌 [1377]。自扩展技术使大型语言模型能够在不微调的情况下处理长上下文，通过采用双层注意力策略——分组注意力和邻近注意力——来捕获远距离和相邻令牌之间的依赖关系 [499]。

高效处理的优化技术 分组查询注意力 (GQA) 将查询头分区为共享键和值头的组，在多查询注意力和多头注意力之间取得平衡，同时减少解码期间的内存需求 [16, 1341]。FlashAttention 利用非对称 GPU 内存层次结构以线性内存扩展代替二次要求，FlashAttention-2 通过减少非矩阵乘法操作和优化工作分配提供大约两倍的速度 [196, 195]。

使用分块Transformer的Ring Attention能够通过在多个设备上分配计算来处理极长序列，利用分块计算的同时重叠通信与注意力计算 [676]。稀疏注意力技术包括LongLoRA和SinkLoRA中的Shifted sparse attention (S_2-Attn)以及具有SF-Attn的SinkLoRA，它们在显著节省计算量的同时实现了92%的完整注意力困惑度提升 [1304, 1217]。

高效的选择性注意力（ESA）通过将查询和键向量压缩到低维表示中，提出了一种基于token级别的关键信息选择，能够处理长达256K token的序列 [1084]。BigBird结合了局部注意力与全局token，这些全局token能够关注整个序列。

plus random connections, enabling efficient processing of sequences up to 8× longer than previously possible [1285].

Memory Management and Context Compression Memory management strategies include Rolling Buffer Cache techniques that maintain fixed attention spans, reducing cache memory usage by approximately 8× on 32K token sequences [1341]. StreamingLLM enables processing infinitely long sequences without fine-tuning by retaining critical "attention sink" tokens together with recent KV cache entries, demonstrating up to 22.2× speedup over sliding window recomputation with sequences up to 4 million tokens [1176].

Infini-attention incorporates compressive memory into vanilla attention, combining masked local attention with long-term linear attention in single Transformer blocks, enabling processing of infinitely long inputs with bounded memory and computation [792]. Heavy Hitter Oracle (H_2O) presents efficient KV cache eviction policies based on observations that small token portions contribute most attention value, improving throughput by up to 29× while reducing latency by up to 1.9× [1333].

Context compression techniques like QwenLong-CPRS implement dynamic context optimization mechanisms enabling multi-granularity compression guided by natural language instructions [944]. InfLLM stores distant contexts in additional memory units and employs efficient mechanisms to retrieve token-relevant units for attention computation, allowing models pre-trained on sequences of a few thousand tokens to effectively process sequences up to 1,024K tokens [1175].

4.2.2. Contextual Self-Refinement and Adaptation

Self-refinement enables LLMs to improve outputs through cyclical feedback mechanisms mirroring human revision processes, leveraging self-evaluation through conversational self-interaction via prompt engineering distinct from reinforcement learning approaches [735, 916, 25, 1211].

Foundational Self-Refinement Frameworks The Self-Refine framework uses the same model as generator, feedback provider, and refiner, demonstrating that identifying and fixing errors is often easier than producing perfect initial solutions [735, 1313, 227]. Reflexion maintains reflective text in episodic memory buffers for future decision-making through linguistic feedback [956], while structured guidance proves essential as simplistic prompting often fails to enable reliable self-correction [672, 587].

Multi-Aspect Feedback integrates frozen language models and external tools focusing on specific error categories to enable more comprehensive, independent evaluation [799]. The N-CRITICS framework implements ensemble-based evaluation where initial outputs are assessed by both generating LLMs and other models, with compiled feedback guiding refinement until task-specific stopping criteria are fulfilled [789].

The A2R framework adopts explicit evaluation across multiple dimensions including correctness and citation quality, formulating natural language feedback for each aspect and iteratively refining outputs [577]. ISR-LLM improves LLM-based planning by translating natural language to formal specifications, creating an initial plan, and then systematically refining it with a validator [1373].

Meta-Learning and Autonomous Evolution SELF teaches LLMs meta-skills (self-feedback, self-refinement) with limited examples, then has the model continuously self-evolve by generating and filtering its own training data [704]. Self-rewarding mechanisms enable models to improve autonomously through iterative

增加随机连接，实现了对长达 8× 此前无法处理的更长时间序列[1285]的序列的高效处理。

内存管理与上下文压缩 内存管理策略包括滚动缓冲区缓存技术，这些技术维护固定的注意力跨度，通过约8× 减少了32K token序列的缓存内存使用。StreamingLLM通过将关键的“注意力吸收”token与最近的KV缓存条目保留在一起，实现了对无限长序列的处理，无需微调，展示了在高达4百万token序列[1176]的情况下，比滑动窗口重新计算快22.2× 的速度提升。

Infini-attention将压缩内存整合到普通的注意力机制中，结合掩码局部注意力和长期线性注意力在单个Transformer块中，实现了在有限的内存和计算量下对无限长输入的处理 [792]。Heavy Hitter Oracle (H_2O)基于观察到小部分token贡献了大部分注意力值的观察结果，提出了高效的KV缓存驱逐策略，通过最多 29× 提升了吞吐量，同时将延迟降低了最多1.9× [1333]。

上下文压缩技术如QwenLong-CPRS实现了动态上下文优化机制，通过自然语言指令指导多粒度压缩[944]。InfLLM将远距离上下文存储在额外的内存单元中，并采用高效机制检索与token相关的单元进行注意力计算，允许在数千个token序列上预训练的模型有效处理高达1,024K token的序列 [1175]。

4.2.2. 上下文自精炼与自适应

自精炼使LLM能够通过循环反馈机制改进输出，这些机制类似于人类修订过程，通过提示工程实现对话式自我交互进行自我评估，这与强化学习方法不同 [735, 916, 25, 1211]。

基础自精炼框架 Self-Refine框架使用与生成器、反馈提供者和精炼器相同的模型，证明识别和修复错误通常比产生完美的初始解决方案更容易 [735, 1313, 227]。Reflexion通过语言反馈在情景记忆缓冲区中维护反思性文本，用于未来的决策 [956], ，而结构化指导被证明至关重要，因为简单的提示往往无法实现可靠的自我纠正 [672, 587]。

多方面反馈整合了冻结的语言模型和专注于特定错误类别的外部工具，以实现更全面、独立的评估[799]。N-CRITICS框架实现了基于集成的方法的评估，其中初始输出由生成式LLM和其他模型共同评估，通过编译的反馈进行改进，直到满足特定任务的停止标准 [789]。

A2R框架在多个维度上采用显式评估，包括正确性和引用质量，为每个方面制定自然语言反馈，并迭代改进输出 [577]。ISR-LLM通过将自然语言转换为形式规范来改进基于LLM的规划，创建初始计划，然后使用验证器系统地改进它 [1373]。

元学习和自主进化 SELF通过有限的示例教授LLM元技能（自我反馈、自我改进），然后让模型通过生成和过滤其自己的训练数据来持续自我进化 [704]。自我奖励机制使模型能够通过迭代自主改进

self-judgment, where a single model adopts dual roles as performer and judge, maximizing rewards it assigns itself [1163, 1278].

The Creator framework extends this paradigm by enabling LLMs to create and use their own tools through a four-module process encompassing creation, decision-making, execution, and recognition [946, 856]. The Self-Developing framework represents the most autonomous approach, enabling LLMs to discover, implement, and refine their own improvement algorithms through iterative cycles generating algorithmic candidates as executable code [466].

In-context learning fundamentally represents a form of meta-learning where models learn optimization strategies during pre-training that generalize across diverse tasks, enabling rapid adaptation to novel challenges during inference [179, 1165]. Meta-in-context learning demonstrates that in-context learning abilities can be recursively improved through in-context learning itself, adaptively reshaping model priors over expected tasks and modifying in-context learning strategies [177].

Memory-Augmented Adaptation Frameworks Memory augmentation represents a powerful approach for implementing meta-learning through frameworks like Memory of Amortized Contexts, which uses feature extraction and memory-augmentation to compress information from new documents into compact modulations stored in memory banks [1011]. Context-aware Meta-learned Loss Scaling addresses outdated knowledge challenges by meta-training small autoregressive models to dynamically reweight language modeling loss for each token during online fine-tuning [430].

Decision-Pretrained Transformers demonstrate how transformers can be trained to perform in-context reinforcement learning, solving previously unseen RL problems by generalizing beyond pretraining distribution [1013, 582]. Context-based meta-reinforcement learning methods enhance performance through direct supervision of context encoders, improving sample efficiency compared to end-to-end training approaches [1072].

Long Chain-of-Thought and Advanced Reasoning Long Chain-of-Thought has emerged as a significant evolution characterized by substantially longer reasoning traces enabling thorough problem exploration, as implemented in advanced models including OpenAI-o1, DeepSeek-R1, QwQ, and Gemini 2.0 Flash Thinking [147, 718, 1214]. LongCoT effectiveness appears linked to context window capacity, with empirical evidence suggesting larger context windows often lead to stronger reasoning performance [1229].

Extended reasoning enables self-reflection and error correction mechanisms allowing models to identify and rectify mistakes during problem-solving processes [1334]. The effectiveness of increasing reasoning step length, even without adding new information, considerably enhances reasoning abilities across multiple datasets through test-time scaling [1345].

Optimization strategies address computational inefficiencies due to verbose reasoning traces through self-generated shorter reasoning paths via best-of-N sampling, adaptive reasoning modes including Zero-Thinking and Less-Thinking approaches, and explicit compact CoT methods reducing token usage while maintaining reasoning quality [791, 1348, 697]. Auto Long-Short Reasoning enables dynamic adjustment of reasoning path length according to question complexity, helping models decide when longer chains are necessary [715].

自我判断，其中单个模型采用表演者和裁判的双重角色，最大化其分配给自己的奖励 [1163, 1278]。

Creator框架通过一个包含创建、决策、执行和识别四个模块的过程，扩展了这一范式，使大型语言模型能够创建和使用自己的工具 [946, 856]。Self-Developing框架代表了最自主的方法，使大型语言模型能够通过迭代循环发现、实现和改进自己的改进算法，并生成算法候选作为可执行代码 [466]。

上下文学习从根本上代表了一种元学习形式，模型在预训练过程中学习能够泛化到各种任务中的优化策略，从而在推理时能够快速适应新的挑战 [179, 1165]。元上下文学习表明，上下文学习能力可以通过上下文学习本身进行递归式改进，自适应地重塑模型对预期任务的先验知识，并修改上下文学习策略 [177]。

记忆增强自适应框架记忆增强代表了一种强大的实现元学习的方法，例如记忆的摊销上下文（Memory of Amortized Contexts）框架，它利用特征提取和记忆增强将新文档的信息压缩成存储在内存库中的紧凑调制 [1011]。上下文感知元学习损失缩放通过元训练小型自回归模型来解决过时知识挑战，在在线微调期间动态地为每个标记重新权衡语言建模损失 [430]。

决策预训练Transformer展示了如何训练Transformer以执行情境强化学习，通过超越预训练分布 [1013, 582] 来解决以前未见过的RL问题。基于情境的元强化学习方法通过直接监督情境编码器来提高性能，与端到端训练方法相比，提高了样本效率 [1072]。

长思维链和高级推理 长思维链已成为一个重要的发展，其特点是推理轨迹显著更长，能够进行彻底的问题探索，如在OpenAI-o1、DeepSeek-R1、QwQ和Gemini 2.0 Flash Thinking [147, 718, 1214] 等高级模型中实现。长CoT的有效性似乎与上下文窗口容量有关，实证证据表明更大的上下文窗口通常会导致更强的推理性能 [1229]。

扩展推理使自我反思和错误纠正机制得以实现，允许模型在问题解决过程中识别和纠正错误 [1334]。增加推理步长（即使不添加新信息）的有效性通过测试时扩展，在多个数据集上显著提高了推理能力 [1345]。

优化策略通过最佳-N采样生成较短的推理路径、自适应推理模式（包括零思考和无思考方法）以及显式紧凑的CoT方法来解决冗长推理轨迹导致的计算低效问题，同时减少token使用量并保持推理质量 [791, 1348, 697]。自动长短推理能够根据问题复杂度动态调整推理路径长度，帮助模型决定何时需要更长的链条 [715]。

4.2.3. Multimodal Context

Multimodal Large Language Models (MLLMs) extend context engineering beyond text by integrating diverse data modalities including vision, audio, and 3D environments into unified contextual representations. This expansion introduces new challenges in modality fusion, cross-modal reasoning, and long-context processing while enabling sophisticated applications that leverage rich multimodal contextual understanding.

Multimodal Context Integration

Foundational Techniques Multimodal MLLMs expand upon traditional LLMs by integrating data from diverse modalities like vision, audio, and 3D environments [105, 49, 957]. A primary integration method converts visual inputs into discrete tokens concatenated with text tokens, conditioning the LLM's generative process on a combined representation [1286]. This is often facilitated by Visual Prompt Generators (VPGs) trained on image-caption pairs to map visual features into the LLM's embedding space [607]. The dominant architectural paradigm connects specialized, external multimodal encoders—such as CLIP for vision or CLAP for audio—to the LLM backbone via alignment modules like Q-Former or simple MLPs [19, 86, 609, 1130], a modular design that allows for independent encoder updates without retraining the entire model [618].

Advanced Integration Strategies More sophisticated approaches enable deeper modality fusion. Cross-modal attention mechanisms learn fine-grained dependencies between textual and visual tokens directly within the LLM's embedding space, enhancing semantic understanding for tasks like image editing [564, 901, 102]. To manage lengthy inputs, hierarchical designs process modalities in stages to ensure scalability [155], while the "browse-and-concentrate" paradigm fuses the contexts of multiple images before LLM ingestion to overcome the limitations of isolated processing [1134]. Some research bypasses the adaptation of text-only LLMs, opting for unified training paradigms that jointly pre-train models on multimodal data and text corpora from the start to mitigate alignment challenges [1381, 1224]. Other methods leverage text as a universal semantic space, using LLM in-context learning to improve generalization across diverse modality combinations [1050]. For video, context integration techniques range from prompt tuning to adapter-based methods that transform video content into a sequence for reasoning [1080]. The development of these models is often constrained by the need for vast, high-quality multimodal data and significant computational resources [1295, 609, 211].

Core Challenges in Multimodal Context Processing

Modality Bias and Reasoning Deficiencies A primary obstacle in MLLM development is modality bias, where models favor textual inputs, generating plausible but multimodally ungrounded responses by relying on learned linguistic patterns rather than integrated visual or auditory information [1358, 24, 315, 1325]. This issue is exacerbated by training methodologies; for instance, VPGs trained on simple image-captioning tasks learn to extract only salient features for captions, neglecting other visual details crucial for more complex, instruction-based tasks, which fundamentally limits deep multimodal understanding [607, 504]. Consequently, MLLMs frequently struggle with fine-grained spatial or temporal reasoning, such as precise object localization or understanding detailed event sequences in videos [1031, 957], particularly in complex domains like social media where interpreting the interplay of text and images to understand misinformation or sarcasm is difficult [505]. Effective multimodal reasoning requires not just comprehending each modality

4.2.3. 多模态上下文

多模态大型语言模型（MLLMs）通过将视觉、音频和3D环境等多种数据模态整合到统一的上下文表示中，将上下文工程扩展到文本之外。这种扩展在模态融合、跨模态推理和长上下文处理方面引入了新的挑战，同时支持利用丰富的多模态上下文理解的复杂应用。

多模态上下文集成

基础技术 多模态MLLMs通过整合来自视觉、音频和3D环境等多种模态的数据扩展了传统LLMs [105, 49, 957]。主要集成方法将视觉输入转换为离散的token，并与文本token连接，在组合表示上对LLM的生成过程进行条件化 [1286]。这通常通过在图像-文本对上训练的视觉提示生成器（VPGs）来实现，以将视觉特征映射到LLM的嵌入空间 [607]。主导的架构范式通过Q-Former或简单的MLPs等对齐模块将专门的外部多模态编码器（如视觉的CLIP或音频的CLAP）连接到LLM骨干 [19, 86, 609, 1130]，a模块化设计，允许独立更新编码器而无需重新训练整个模型 [618]。

高级集成策略 更复杂的方法支持更深度的模态融合。跨模态注意力机制直接在LLM的嵌入空间内学习文本和视觉标记之间的细粒度依赖关系，从而增强图像编辑等任务的语义理解 [564,901, 102]。为了管理长输入，分层设计分阶段处理模态以确保可扩展性[155]，，而“浏览和集中”范式在LLM处理前融合多个图像的上下文以克服孤立处理的局限性 [1134]。一些研究绕过纯文本LLM的适配，选择统一的训练范式，从一开始就在多模态数据和文本语料库上联合预训练模型，以缓解对齐挑战 [1381, 1224]。其他方法利用文本作为通用语义空间，使用LLM上下文学习来提高跨不同模态组合的泛化能力 [1050]。对于视频，上下文集成技术范围从提示调整到基于适配器的方法，这些方法将视频内容转换为推理序列 [1080]。这些模型的开发通常受限于对大量高质量多模态数据和显著计算资源的需要 [1295, 609, 211]。

多模态上下文处理中的核心挑战

模态偏差和推理缺陷 在 MLLM 开发中的主要障碍是模态偏差，模型倾向于文本输入，通过依赖学习到的语言模式而不是整合的视觉或听觉信息 [1358, 24, 315, 1325] 来生成看似合理但多模态未扎根的响应。这个问题由于训练方法而加剧；例如，在简单的图像-标题任务上训练的 VPG 只学习用于标题的显著特征，忽视对更复杂、基于指令的任务至关重要的其他视觉细节，这从根本上限制了深度多模态理解 [607, 504]。因此，MLLM 经常难以进行细粒度的空间或时间推理，例如精确的对象定位或理解视频中的详细事件序列 [1031, 957]，，特别是在像社交媒体这样复杂的领域，解释文本和图像之间的相互作用以理解虚假信息或讽刺是困难的 [505]。有效的多模态推理不仅需要理解每种模态

but also inferring their combined holistic meaning [385]. Compounding these issues is our limited mechanistic understanding of MLLMs themselves; their internal workings are largely a black box, hindering the development of better architectures [1274].

Advanced Contextual Capabilities and Future Directions

In-Context and Long-Context Learning A key capability of MLLMs is in-context learning, where models adapt to new tasks from multimodal examples in the prompt without weight updates [1397, 1398, 551]. Link-context learning (LCL) enhances this by providing demonstrations with explicit causal links, improving generalization [1012]. However, in-context learning is constrained by fixed context windows, as image tokens consume significant space, limiting many-shot learning [437]. Performance is also sensitive to input order and the relative importance of each modality varies by task [1020, 1197]. Processing long multimodal contexts, crucial for applications like video analysis, remains a major research frontier [1086]. Innovations include adaptive hierarchical token compression for video [1119], variable visual position encoding (V2PE) [1381], specialized modules like ContextQFormer for conversational memory [589], and dynamic, query-aware frame selection for video [581]. MLLMs also show emergent communication efficiency over extended interactions, a phenomenon still under investigation [436].

Emerging Applications The ability to process rich multimodal context is unlocking new applications. MLLMs are used for predictive reasoning, such as forecasting human activity from visual scenes [1382], and have demonstrated impressive perception and cognitive capabilities across various multimodal benchmarks [290]. In VQA, context is leveraged for more precise answers, for instance, by prompting the MLLM to generate its own descriptive text context of an image [1346] or by integrating external knowledge via RAG [993, 105]. Other applications include planning digital actions based on sensory inputs [605], enhancing surgical decision support through memory-augmented context comprehension [418], and enabling nuanced video understanding by integrating visual information with speech and audio cues [642, 1193, 7]. Researchers have also extended MLLMs to emerging modalities like tactile information, event data, and graph structures [1358, 1023, 1213]. The growing importance of these real-world use cases has spurred the development of comprehensive evaluation frameworks to assess contextual comprehension [1109]. These advancements enable applications previously impossible with text-only models, such as image captioning and sophisticated multimodal reasoning [1173, 677, 139].

4.2.4. Relational and Structured Context

Large language models face fundamental constraints processing relational and structured data including tables, databases, and knowledge graphs due to text-based input requirements and sequential architecture limitations [489, 47, 1136]. Linearization often fails to preserve complex relationships and structural properties, with performance degrading when information is dispersed throughout contexts [586, 585, 938].

Knowledge Graph Embeddings and Neural Integration Advanced encoding strategies address structural limitations through knowledge graph embeddings that transform entities and relationships into numerical vectors, enabling efficient processing within language model architectures [12, 1250, 930, 1194]. Graph neural networks capture complex relationships between entities, facilitating multi-hop reasoning across

但也推断它们的整体综合意义 [385]。这些问题叠加起来是我们对MLLMs本身的有限机制理解；它们内部的工作原理基本上是一个黑箱，阻碍了更好架构的开发 [1274]。

高级上下文能力和未来方向

情境学习和长情境学习 MLLMs的一个关键能力是情境学习，其中模型通过提示中的多模态示例适应新任务，而无需更新权重 [1397, 1398, 551]。链接情境学习（LCL）通过提供具有明确因果链接的演示来增强这一点，从而提高泛化能力 [1012]。然而，情境学习受限于固定的上下文窗口，因为图像标记消耗大量空间，限制了多示例学习 [437]。性能也受输入顺序的影响，并且每个模态的相对重要性因任务而异 [1020, 1197]。处理长多模态上下文对于视频分析等应用至关重要，仍然是主要的科研前沿 [1086]。创新包括用于视频的自适应分层标记压缩 [1119]，变量视觉位置编码（V2PE）[1381]，专门模块，如用于对话记忆的ContextQFormer [589]，以及用于视频的动态、查询感知帧选择 [581]。MLLMs在扩展交互中也表现出涌现的通信效率，这一现象仍在研究中 [436]。

新兴应用 处理丰富的多模态上下文的能力正在解锁新的应用。MLLMs用于预测推理，例如从视觉场景中预测人类活动 [1382]，，并在各种多模态基准测试[290]中展示了令人印象深刻的感知和认知能力。在VQA中，利用上下文以获得更精确的答案，例如，通过提示MLLM生成图像的自描述性文本上下文 [1346]，或通过RAG[993, 105]集成外部知识。其他应用包括根据感官输入 [605]，规划数字操作，通过记忆增强的上下文理解增强手术决策支持 [418]，，以及通过将视觉信息与语音和音频提示相结合来启用细致的视频理解 [642, 1193, 7]。研究人员还扩展MLLMs以适应新兴的模态，如触觉信息、事件数据和图结构[1358, 1023, 1213]。这些现实世界的用例日益重要，推动了全面评估框架的开发以评估上下文理解 [1109]。这些进步使得仅使用文本模型的以前不可能的应用成为可能，例如图像描述和复杂的multimodal 推理 [1173, 677, 139]。

4.2.4. 关系和结构化上下文

大型语言模型在处理关系和结构化数据（包括表格、数据库和知识图谱）时面临基本限制，这是由于基于文本的输入要求和顺序架构限制所致 [489, 47, 1136]。线性化往往无法保留复杂关系和结构化属性，当信息分散在整个上下文中时，性能会下降 [586, 585, 938]。

知识图谱嵌入和神经集成 高级编码策略通过知识图谱嵌入解决结构化限制，将实体和关系转换为数值向量，从而在语言模型架构中实现高效处理 [12, 1250, 930, 1194]。图神经网络捕获实体之间的复杂关系，促进跨多跳推理

knowledge graph structures through specialized architectures like GraphFormers that nest GNN components alongside transformer blocks [974, 404, 1221, 483].

GraphToken demonstrates substantial improvements by explicitly representing structural information, achieving up to 73 percentage points enhancement on graph reasoning tasks through parameter-efficient encoding functions [836]. Heterformer and other hybrid GNN-LM architectures perform contextualized text encoding and heterogeneous structure encoding in unified models, addressing the computational challenges of scaling these integrated systems [496, 465, 751].

Method	Approach	Performance	Key Innovation
ODA [1001]	Observation-driven agent framework	12.87% and 8.9% improvements	Recursive observation with action-reflection
RAG-KG [1206]	Historical issue KG construction	77.6% MRR, 0.32 BLEU improvement	Query parsing and sub-graph retrieval
KARPA [258]	Training-free KG adaptation	State-of-the-art KGQA performance	Pre-planning relation paths
Faithful Reasoning [720]	Planning-retrieval-reasoning framework	N/A	LLM-KG synergy with relation paths

Table 3: Knowledge graph integration methods for enhanced reasoning in large language models.

Verbalization and Structured Data Representations Verbalization techniques convert structured data including knowledge graph triples, table rows, and database records into natural language sentences, enabling seamless integration with existing language systems without architectural modifications [12, 782, 1064, 13]. Multi-level structurization approaches reorganize input text into layered structures based on linguistic relationships, while structured data representations leverage existing LLMs to extract structured information and represent key elements as graphs, tables, or relational schemas [681, 1125, 1324, 1035, 602].

Programming language representations of structured data, particularly Python implementations for knowledge graphs and SQL for databases, outperform traditional natural language representations in complex reasoning tasks by leveraging inherent structural properties [1166]. Resource-efficient approaches using structured matrix representations offer promising directions for reducing parameter counts while maintaining performance on structured data tasks [343].

Integration Frameworks and Synergized Approaches The integration of knowledge graphs with language models follows distinct paradigms characterized by different implementation strategies and performance trade-offs [817, 1140]. Pre-training integration methods like K-BERT inject knowledge graph triples during training to internalize factual knowledge, while inference-time approaches enable real-time knowledge access without requiring complete model retraining [690, 1237, 712].

KG-enhanced LLMs incorporate structured knowledge to improve factual grounding through retrieval-based augmentation methods like KAPING, which retrieves relevant facts based on semantic similarities and prepends them to prompts without requiring model training [48, 673, 591]. More sophisticated implementations embed KG-derived representations directly into model latent spaces through adapter modules and cross-attention mechanisms, with Text2Graph mappers providing linking between input text and KG embedding spaces [132, 1066, 428].

Synergized approaches create unified systems where both technologies play equally important roles, addressing fundamental limitations through bidirectional reasoning driven by data and knowledge [817, 853, 1111]. GreaseLM facilitates deep interaction across all model layers, allowing language context representations to be grounded by structured world knowledge while linguistic nuances inform graph

通过专用架构（如GraphFormers）嵌套GNN组件和transformer模块来构建知识图谱结构 [974, 404, 1221, 483]。

GraphToken通过显式表示结构信息，在图推理任务上实现了显著提升，通过参数高效的编码函数达到了73个百分点的改进 [836]。Heterformer和其他hybridGNN-LM架构在统一模型中执行上下文化文本编码和异构结构编码，解决了扩展这些集成系统的计算挑战 [496, 465, 751]。

方法	方法	性能	关键创新
ODA [1001]	基于观察的代理框架 12.87% 和 8.9% 的改进		递归观察与行动- flection
RAG-KG [1206]	历史问题 KG 构建	77.6% MRR,0.32 BLEU 改进 查询解析和子图检索	al
KARPA [258]	无需训练的KG适配	最先进的KGQA性能预规划关系路径	
忠实推理 **ning** [720]	规划-检索-推理 framework N/A		LLM与关系路径的协同

Table 3: Knowledge graph integration methods for enhanced reasoning in large language models.

语言化与结构化数据表示 语言化技术将结构化数据（包括知识图谱三元组、表格行和数据库记录）转换为自然语言句子，从而无需架构修改即可与现有语言系统无缝集成 [12, 782, 1064, 13]。多级结构化方法根据语言关系将输入文本重新组织为分层结构，而结构化数据表示则利用现有的LLM提取结构化信息，并将关键元素表示为图、表格或关系模式 [681, 1125, 1324, 1035, 602]。

结构化数据的编程语言表示，特别是用于知识图谱的Python实现和用于数据库的SQL，通过利用固有的结构属性，在复杂推理任务中优于传统的自然语言表示 [1166]。使用结构化矩阵表示的资源高效方法为减少参数数量同时保持结构化数据任务的性能提供了有前景的方向 [343]。

集成框架与协同方法 知识图谱与语言模型的集成遵循不同的范式，其特点在于不同的实现策略和性能权衡 [817, 1140]。预训练集成方法如K-BERT在训练过程中注入知识图谱三元组以内化事实知识，而推理时方法则支持实时知识访问，无需完全重新训练模型 [690, 1237, 712]。

KG增强型LLM结合结构化知识来通过基于检索的增强方法（如KAPING）提高事实基础，该方法根据语义相似性检索相关事实并将它们添加到提示中，而无需模型训练 [48, 673, 591]。更复杂的实现通过适配器模块和交叉注意力机制将KG派生的表示直接嵌入到模型潜在空间中，而Text2Graph映射器在输入文本和KG嵌入空间之间提供链接 [132, 1066, 428]。

协同方法创建统一系统，其中两种技术都发挥着同等重要的作用，通过数据和知识驱动的双向推理来解决基本限制 [817,853, 1111]。GreaseLM促进所有模型层之间的深度交互，允许语言上下文表示由结构化世界知识进行基础，而语言细微差别则影响图

representations [1321]. QA-GNN implements bidirectional attention mechanisms connecting question-answering contexts and knowledge graphs through joint graph formation and mutual representation updates via graph-based message passing [1250, 974].

Applications and Performance Enhancement Structured data integration significantly enhances LLM capabilities across multiple dimensions, with knowledge graphs providing structured information that reduces hallucinations by grounding responses in verifiable facts and improving factual accuracy through clearly defined information sources [1002, 1342, 200, 565]. Knowledge graphs enhance reasoning capabilities by providing structured entity relationships that enable complex multi-hop reasoning and logical inferences, with their rich repository of hierarchical knowledge significantly improving precision and reliability of inferences [1166, 208, 1018].

Real-world applications demonstrate substantial improvements across specialized domains. Healthcare systems combine structured medical knowledge with contextual understanding through Retrieval-Augmented Generation frameworks to improve disease progression modeling and clinical decision-making [842, 583]. Scientific research platforms organize findings into structured knowledge supporting hypothesis generation and research gap identification, while business analytics systems balance rule-based precision with AI pattern recognition for more actionable insights [1326, 1062].

Question answering systems benefit from natural language interfaces over structured data sources, with integration creating more robust systems capable of handling multimodal queries and providing personalized responses that overcome static knowledge base limitations [1317, 1116, 914, 1206]. Research demonstrates that structured knowledge representations can improve summarization performance by 40% and 14% across public datasets compared to unstructured memory approaches, with Chain-of-Key strategies providing additional performance gains through dynamic structured memory updates [459].

Method	Data Type	Integration Method	Key Innovation	Task Scope
K-LAMP [48]	Knowledge graphs	Retrieval-based augmentation	KAPING framework	Zero-shot QA
Pan et al. [817]	Knowledge graphs	Pre-training & inference integration	Synergized LLMs + KGs	Multi-domain reasoning
StructLM [1392]	Tables, graphs, databases	Instruction tuning	1.1M example dataset	18 datasets, 8 SKG tasks
Shao et al. [938]	Tables, databases, KGs	Linearization methods	Schema linking & syntax prediction	Text-to-SQL tasks

Table 4: Representative approaches for structured data integration in large language models.

4.3. Context Management

Context Management addresses the efficient organization, storage, and utilization of contextual information within LLMs. This component tackles fundamental constraints imposed by finite context windows, develops sophisticated memory hierarchies and storage architectures, and implements compression techniques to maximize information density while maintaining accessibility and coherence.

4.3.1. Fundamental Constraints

LLMs face fundamental constraints in context management stemming from finite context window sizes inherent in most architectures, which significantly reduce model efficacy on tasks requiring deep understanding of lengthy documents while imposing substantial computational demands that hinder applications requiring quick responses and high throughput [1074]. Although extending context windows enables models to handle

表示 [1321]。 QA-GNN 实现了双向注意力机制，通过联合图形成和基于图的消息传递进行相互表示更新，将问答上下文和知识图谱连接起来 [1250, 974]。

应用与性能提升 结构化数据集成在多个维度上显著增强了 LLM 的能力，知识图谱提供了结构化信息，通过将响应与可验证的事实相结合来减少幻觉，并通过明确定义的信息来源提高事实准确性 [1002, 1342, 200, 565]。知识图谱通过提供结构化的实体关系来增强推理能力，从而实现复杂的multi-hop推理和逻辑推断，其丰富的层次知识库显著提高了推断的精确性和可靠性[1166, 208, 1018]。

现实世界的应用在专业领域展示了显著的改进。医疗保健系统通过检索增强生成框架将结构化医学知识与上下文理解相结合，以改进疾病进展建模和临床决策 [842, 583]。科学研究平台将发现组织成结构化知识，支持假设生成和研究差距识别，而商业分析系统在基于规则的精确性与AI模式识别之间取得平衡，以获得更具操作性的见解 [1326, 1062]。

问答系统受益于自然语言界面而非结构化数据源，集成可创建更强大的系统能够处理多模态查询并提供个性化响应以克服静态知识库的局限性 [1317, 1116, 914, 1206]。研究表明，与无结构化记忆方法相比，结构化知识表示可提高摘要性能，在公共数据集上分别提升40%和14%，而链式关键策略通过动态结构化记忆更新提供额外的性能提升 [459]。

方法	数据类型	集成方法	关键创新	任务范围
K-LAMP [48]	知识图谱	基于检索的增强	KAPING框架	零样本问答
Pan等人 [817]	知识图谱	预训练&推理集成	协同LLMs + KGs	多领域阅读 soning
StructLM [1392]	表格，图表，数据库	指令微调	1.1M 示例数据集	18 个数据集，8 个 SK G tasks
Shao 等人 [938]	表格，数据库，知识图谱	线性化方法	Schema linking & syntax	x prediction Text-to-SQL tasks

Table 4: Representative approaches for structured data integration in large language models.

4.3. 上下文管理

上下文管理负责在大型语言模型中高效地组织、存储和利用上下文信息。该组件解决了由有限上下文窗口带来的基本约束，开发了复杂的内存层次结构和存储架构，并实施了压缩技术，以在保持可访问性和连贯性的同时最大化信息密度。

4.3.1. 基本约束

大型语言模型在上下文管理方面面临基本约束，这些约束源于大多数架构中固有的有限上下文窗口大小，这显著降低了模型在需要深入理解长文档的任务上的效能，同时带来了巨大的计算需求，阻碍了需要快速响应和高吞吐量的应用 [1074]。尽管扩展上下文窗口使模型能够处理

entire documents and capture longer-range dependencies, traditional transformer architectures experience quadratic computational complexity growth as sequence length increases, making processing extremely long texts prohibitively expensive [999]. While innovative approaches like LongNet have reduced this complexity to linear, balancing window size and generalization capabilities remains challenging [999, 216].

Empirical evidence reveals the "lost-in-the-middle" phenomenon, where LLMs struggle to access information positioned in middle sections of long contexts, performing significantly better when relevant information appears at the beginning or end of inputs [128, 685, 648]. This positional bias severely impacts performance in extended chain-of-thought reasoning tasks where critical earlier results become susceptible to forgetting, with performance degrading drastically by as much as 73% compared to performance with no prior context [128, 1138, 377].

LLMs inherently process each interaction independently, lacking native mechanisms to maintain state across sequential exchanges and robust self-validation mechanisms, constraints stemming from fundamental limits identified in Gödel's incompleteness theorems [128, 368]. This fundamental statelessness necessitates explicit management systems to maintain coherent operation sequences and ensure robust failure recovery mechanisms [128]. Context management faces opposing challenges of context window overflow, where models "forget" prior context due to exceeding window limits, and context collapse, where enlarged context windows or conversational memory cause models to fail in distinguishing between different conversational contexts [985]. Research demonstrates that claimed benefits of chain-of-thought prompting don't stem from genuine algorithmic learning but rather depend on problem-specific prompts, with benefits deteriorating as problem complexity increases [984]. The computational overhead of long-context processing creates additional challenges in managing key-value caches which grow substantially with input length, creating bottlenecks in both latency and accuracy, while multi-turn and longitudinal interaction challenges further complicate context management as limited effective context hinders longitudinal knowledge accumulation and token demands of many-shot prompts constrain space available for system and user inputs while slowing inference [911, 719, 389].

4.3.2. Memory Hierarchies and Storage Architectures

Modern LLM memory architectures employ sophisticated hierarchical designs organized into methodological approaches to overcome fixed context window limitations. OS-inspired hierarchical memory systems implement virtual memory management concepts, with MemGPT exemplifying this approach through systems that page information between limited context windows (main memory) and external storage, similar to traditional operating systems [813]. These architectures consist of main context containing system instructions, FIFO message queues, and writable scratchpads, alongside external context holding information accessible through explicit function calls, with memory management through function-calling capabilities enabling autonomous paging decisions [831]. PagedAttention, inspired by virtual memory and paging techniques in operating systems, manages key-value cache memory in LLMs [57].

Dynamic memory organizations implement innovative systems based on cognitive principles, with MemoryBank using Ebbinghaus Forgetting Curve theory to dynamically adjust memory strength according to time and significance [1202, 1362]. ReadAgent employs episode pagination to segment content, memory gisting to create concise representations, and interactive look-up for information retrieval [1202]. Compressor-retriever architectures support life-long context management by using base model forward functions to compress and retrieve context, ensuring end-to-end differentiability [1236].

Architectural adaptations enhance model memory capabilities through internal modifications including augmented attention mechanisms, refined key-value cache mechanisms, and modified positional encodings

完整文档并捕获更长距离的依赖关系，传统的 transformer 架构在序列长度增加时会出现二次计算复杂度增长，使得处理极长文本的成本变得极其昂贵 [999]。虽然像 LongNet 这样的创新方法将这种复杂度降低到线性，但平衡窗口大小和泛化能力仍然具有挑战性 [999, 216]。

实证研究表明存在“中间丢失”现象，其中大型语言模型难以访问长上下文中位于中间位置的信息，当相关信息出现在输入的开头或结尾时表现明显更好 [128, 685, 648]。这种位置偏差严重影响了在扩展的思维链推理任务中的性能，其中关键的早期结果容易受到遗忘的影响，与没有先验上下文时的性能相比，性能下降高达 73%[128, 1138, 377]。

大型语言模型本质上独立处理每个交互，缺乏在顺序交换中保持状态的原生机制和强大的自我验证机制，这些限制源于哥德尔不完备性定理中确定的基本限制 [128, 368]。这种基本的无状态性需要显式的管理系统来维护连贯的操作序列并确保强大的故障恢复机制 [128]。上下文管理面临相反的挑战：上下文窗口溢出，其中模型由于超出窗口限制而“忘记”先前的上下文，以及上下文崩溃，其中扩大的上下文窗口或对话记忆导致模型无法区分不同的对话上下文 [985]。研究表明，思维链提示的所谓好处并非来自真正的算法学习，而是依赖于特定问题的提示，随着问题复杂性的增加，这些好处会逐渐恶化 [984]。长上下文处理的计算开销在管理键值缓存时创造了额外的挑战，这些缓存随着输入长度的增加而显著增长，在延迟和准确性方面都造成了瓶颈，而多轮和纵向交互挑战进一步使上下文管理复杂化，因为有限的有效上下文阻碍了纵向知识积累，而多轮提示的标记需求限制了系统和用户输入可用空间，同时减缓了推理 [911, 719, 389]。

4.3.2. 内存层次和存储架构

现代LLM内存架构采用复杂的层次化设计，通过方法论途径克服固定的上下文窗口限制。受操作系统启发的层次化内存系统实现了虚拟内存管理概念，MemGPT通过在有限的上下文窗口（主内存）和外部存储之间分页信息来体现这一方法，类似于传统操作系统 [813]。这些架构包括包含系统指令、FIFO消息队列和可写临时区域的主上下文，以及通过显式函数调用可访问的外部上下文，通过函数调用能力实现内存管理，支持自主分页决策 [831]。受操作系统虚拟内存和分页技术启发的PagedAttention管理LLM中的键值缓存内存 [57]。

动态内存组织基于认知原理实现创新系统，MemoryBank使用艾宾浩斯遗忘曲线理论根据时间和重要性动态调整内存强度 [1202, 1362]。ReadAgent采用场景分页来分割内容，内存梗概来创建简洁表示，以及交互式查找来检索信息 [1202]。压缩器-检索器架构通过使用基础模型前向函数来压缩和检索上下文，支持终身上下文管理，并确保端到端可微分 [1236]。

架构适配通过内部修改增强模型内存能力，包括增强的注意力机制、精炼的键值缓存机制和修改的位置编码

[160, 1352]. Knowledge-organization methods structure memory into interconnected semantic networks enabling adaptive management and flexible retrieval, while retrieval mechanism-oriented approaches integrate semantic retrieval with memory forgetting mechanisms [515, 1362, 444].

System configurations balance efficiency and scalability through organizational approaches where centralized systems coordinate tasks efficiently but struggle with scalability as topics increase, leading to context overflow, while decentralized systems reduce context overflow but increase response time due to inter-agent querying [396]. Hybrid approaches balance shared knowledge with specialized processing for semi-autonomous operation, addressing challenges in balancing computational efficiency with contextual fidelity while mitigating memory saturation where excessive storage of past interactions leads to retrieval inefficiencies [160, 396]. Context Manager Components provide fundamental capabilities for snapshot creation, restoration of intermediate generation states, and overall context window management for LLMs [757].

4.3.3. Context Compression

Context compression techniques enable LLMs to handle longer contexts efficiently by reducing computational and memory burden while preserving critical information. Autoencoder-based compression achieves significant context reduction through In-context Autoencoder (ICAE), which achieves 4× context compression by condensing long contexts into compact memory slots that LLMs can directly condition on, significantly enhancing models' ability to handle extended contexts with improved latency and memory usage during inference [317]. Recurrent Context Compression (RCC) efficiently expands context window length within constrained storage space, addressing challenges of poor model responses when both instructions and context are compressed by implementing instruction reconstruction techniques [441].

Memory-augmented approaches enhance context management through kNN-based memory caches that store key-value pairs of past inputs for later lookup, improving language modeling capabilities through retrieval-based mechanisms [393]. Contrastive learning approaches enhance memory retrieval accuracy, while side networks address memory staleness without requiring LLM fine-tuning, and consolidated representation methods dynamically update past token representations, enabling arbitrarily large context windows without being limited by fixed memory slots [393].

Hierarchical caching systems implement sophisticated multi-layer approaches, with Activation Refilling (ACRE) employing Bi-layer KV Cache where layer-1 cache captures global information compactly and layer-2 cache provides detailed local information, dynamically refilling L1 cache with query-relevant entries from L2 cache to integrate broad understanding with specific details [859]. Infinite-LLM addresses dynamic context length management through DistAttention for distributing attention computation across GPU clusters, liability mechanisms for borrowing memory across instances, and global planning coordination [935]. KCache optimizes inference by storing K Cache in high-bandwidth memory while keeping V Cache in CPU memory, selectively copying key information based on attention calculations [935].

Multi-agent distributive processing represents an emerging approach using LLM-based multi-agent methods to handle massive inputs in distributed manner, addressing core bottlenecks in knowledge synchronization and reasoning processes when dealing with extensive external knowledge [699]. Analysis of real-world key-value cache access patterns reveals high cache reusability in workloads like RAG and agents, highlighting the need for efficient distributed caching systems with optimized metadata management to reduce redundancy and improve speed [1389]. These compression techniques can be combined with other long-context modeling approaches to further enhance LLMs' capacity to process and utilize extended contexts efficiently while reducing computational overhead and preserving information integrity [317].

[160, 1352]. 知识组织方法将记忆结构化为相互连接的语义网络，实现自适应管理和灵活检索，而检索机制导向的方法将语义检索与记忆遗忘机制相结合 [515, 1362, 444]。

系统配置通过组织方法平衡效率和可扩展性，其中集中式系统高效协调任务，但随着主题增加难以扩展，导致上下文溢出，而分布式系统减少上下文溢出，但由于代理间查询增加响应时间 [396]。混合方法平衡共享知识与专业处理，实现半自主运行，解决在平衡计算效率与上下文保真度时面临的挑战，同时缓解记忆饱和问题，即过度存储过去交互导致检索效率低下 [160, 396]。上下文管理组件为快照创建、中间生成状态的恢复以及LLM的整体上下文窗口管理提供基本功能 [757]。

4.3.3. 上下文压缩

上下文压缩技术使LLM能够通过减少计算和内存负担来高效处理更长的上下文，同时保留关键信息。基于自动编码器的压缩通过情境自动编码器（ICAE）实现显著的上下文压缩，通过将长上下文压缩成LLM可以直接条件化的紧凑内存槽，显著提高模型处理扩展上下文的能力，并在推理过程中改进延迟和内存使用 4× [317]。循环上下文压缩（RCC）在受限存储空间内高效扩展上下文窗口长度，解决当指令和上下文都被压缩时模型响应质量差的问题，通过实现指令重建技术 [441]。

基于内存的增强方法通过基于kNN的内存缓存来增强上下文管理，这些缓存存储过去输入的键值对以供后续查找，通过检索机制提高语言建模能力 [393]。对比学习方法增强了内存检索的准确性，而侧网络在不需要LLM微调的情况下解决了内存陈旧问题，而整合表示方法动态更新过去标记的表示，使得上下文窗口可以任意大而不受固定内存槽的限制 [393]。

分层缓存系统实现了复杂的多层方法，激活重填（ACRE）采用双层KV缓存，其中层1缓存紧凑地捕获全局信息，层2缓存提供详细的局部信息，动态地从L2缓存中用与查询相关的条目填充L1缓存，以将广泛的理解与具体细节相结合 [859]。无限-LLM通过分布式注意力（DistAttention）在GPU集群中分配注意力计算、跨实例借用的责任机制以及全局规划协调来解决动态上下文长度管理问题 [935]。KCache通过在高带宽内存中存储K缓存，同时将V缓存保留在CPU内存中，根据注意力计算选择性地复制键信息来优化推理 [935]。

多智能体分布式处理代表了一种新兴方法，使用基于LLM的多智能体方法以分布式方式处理大量输入，解决了在处理大量外部知识时知识同步和推理过程中的核心瓶颈 [699]。对现实世界键值缓存访问模式的分析揭示了RAG和智能体等工作负载中缓存的高复用性，突出了对具有优化元数据管理的有效分布式缓存系统的需求，以减少冗余并提高速度 [1389]。这些压缩技术可以与其他长上下文建模方法相结合，以进一步增强LLM处理和有效利用扩展上下文的能力，同时减少计算开销并保持信息完整性 [317]。

Method	Strategy	Efficiency	Accuracy	Length Mgmt	Scalability
O1-Pruner [718]	RL fine-tuning	N/A	+Acc, -Overhead	Auto pruning	+Efficiency
InftyThink [1214]	Iterative + summarization	Complexity reduction	+3-13%	Iterative control	Scalable
Long-CoT Survey [147]	Long CoT + reasoning	+Efficiency frameworks	+Complex domains	Deep exploration	Test-time scaling
PREMISE [1273]	Prompt opt + diagnostics	Gradient-inspired opt	Maintained/+Acc	-87.5% tokens	Performance maintained
Prune-on-Logic [721]	Structure-aware pruning	Selective pruning	+Accuracy	Selective framework	Logic-based opt

Table 5: Long-chain reasoning methods and their characteristics in large language models. O1-Pruner uses reinforcement learning-style fine-tuning to shorten reasoning chains while maintaining accuracy. InftyThink employs iterative reasoning with intermediate summarization to reduce computational complexity. Long-CoT Survey explores long chain-of-thought characteristics that enhance reasoning abilities through efficiency improvements and enhanced knowledge frameworks. PREMISE optimizes prompts with trace-level diagnostics using gradient-inspired optimization, achieving 87.5% token reduction. Prune-on-Logic performs structure-aware pruning of logic graphs through selective removal of low-utility reasoning steps.

4.3.4. Applications

Effective context management extends LLMs' capabilities beyond simple question-answering to enable sophisticated applications leveraging comprehensive contextual understanding across multiple domains. Document processing and analysis capabilities enable LLMs to handle entire documents or comprehend full articles rather than fragments, allowing for contextually relevant responses through comprehensive understanding of input material, particularly valuable for inherently long sequential data such as gene sequences, legal documents, and technical literature where maintaining coherence across extensive content is critical [999].

Extended reasoning capabilities facilitated by context management techniques support complex reasoning requiring maintenance and building upon intermediate results across extended sequences. By capturing longer-range dependencies, these systems support multi-step problem solving where later reasoning depends on earlier calculations or deductions, enabling sophisticated applications in fields requiring extensive contextual awareness like complex decision support systems and scientific research assistance [999, 160].

Collaborative and multi-agent systems benefit from effective context management in multi-turn dialogues or sequential tasks where maintaining consistent state and synchronizing internal information between collaborating models is essential [154]. These capabilities support applications including distributed task processing, collaborative content creation, and multi-agent problem-solving where contextual coherence across multiple interactions must be maintained [154].

Enhanced conversational interfaces leverage robust context management to seamlessly handle extensive conversations without losing thread coherence, enabling more natural, persistent dialogues that closely resemble human conversations [883]. Task-oriented LLM systems benefit from structured context management approaches, with sliding window storage implementing minimal context management systems that permanently append prompts and responses to context stores, and Retrieval-Augmented Generation systems supplementing LLMs with access to external sources of dynamic information [212, 926]. These capabilities support applications like personalized virtual assistants, long-term tutoring systems, and therapeutic conversational agents that maintain continuity across extended interactions [883].

Memory-augmented applications implement strategies enabling LLMs to persistently store, manage,

方法	策略	效率	准确率	长度管理	可扩展性
O1-Pruner [718]	RL微调	N/A	+Acc, -Overhead	自动剪枝	+效率
InftyThink [1214]	迭代 + 摘要 复杂性降低		+3-13%	迭代控制	可扩展
长-CoT 调查 [147] 长 CoT +推理		+效率框架	+复杂领域 深入探索		测试时扩展
前提 [1273]	提示 opt + 诊断	梯度启发式 opt	保持/+Acc	-87.5% tokens	性能主 tained
基于逻辑剪枝 [721]	结构感知剪枝	选择性剪枝	+ 准确率	选择性框架 逻辑优化	

表 5：大型语言模型中的长链推理方法及其特点。O1-Pruner 使用强化学习风格的微调来缩短推理链，同时保持准确性。InftyThink 采用带中间摘要的迭代推理来降低计算复杂度。Long-CoT 调查通过效率提升和增强知识框架来增强推理能力的长链思维特征。PREMISE 通过梯度启发式优化进行提示优化，并使用跟踪级诊断，实现了 87.5% 的 token 减少。Prune-on-Logic 通过选择性地移除低效用推理步骤，对逻辑图进行结构感知剪枝。

4.3.4. 应用

有效的上下文管理扩展了 LLM 的能力，使其超越简单的问答，能够通过跨多个领域的全面上下文理解来支持复杂应用。文档处理和分析能力使 LLM 能够处理整个文档或理解完整文章，而不是片段，通过全面理解输入材料，允许基于上下文的响应，这对于基因序列、法律文件和技术文献等固有的长序列数据尤其有价值，在这些数据中保持跨大量内容的连贯性至关重要 [999]。

通过上下文管理技术实现的扩展推理能力支持需要维护和扩展中间结果的复杂推理，这些结果跨越了扩展序列。通过捕获更长范围的依赖关系，这些系统支持多步骤问题解决，其中后续推理依赖于早期的计算或推论，从而在需要广泛上下文感知的领域（如复杂的决策支持系统和科学研究辅助）实现高级应用 [999, 160]。

协作和多智能体系统受益于有效的上下文管理，在多轮对话或顺序任务中，维护一致状态和在不同协作模型之间同步内部信息至关重要 [154]。这些能力支持包括分布式任务处理、协作内容创建和多智能体问题解决等应用，其中需要在多个交互中保持上下文连贯性 [154]。

增强的对话界面利用强大的上下文管理，无缝处理大量对话而不丢失线程连贯性，实现更自然、持续的对话，这些对话与人类对话非常相似 [883]。面向任务的LLM系统受益于结构化的上下文管理方法，滑动窗口存储实现最小上下文管理系统，永久将提示和响应追加到上下文存储中，检索增强生成系统通过访问外部动态信息源补充LLM [212, 926]。这些能力支持个性化虚拟助手、长期辅导系统和治疗性对话代理等应用，这些应用在扩展交互中保持连续性 [883]。

基于记忆的应用实现策略，使大型语言模型能够持久存储、管理 e,

and dynamically retrieve relevant contextual information, supporting applications requiring knowledge accumulation over time through building personalized user models via continuous interaction, implementing effective knowledge management across extended interactions, and supporting long-term planning scenarios depending on historical context [160]. Advanced memory frameworks like Contextually-Aware Intelligent Memory (CAIM) enhance long-term interactions by incorporating cognitive AI principles through modules that enable storage and retrieval of user-specific information while supporting contextual and time-based relevance filtering [1143]. Memory management for LLM agents incorporates processes analogous to human memory reconsolidation, including deduplication, merging, and conflict resolution, with approaches like Reflective Memory Management combining prospective and retrospective reflection for dynamic summarization and retrieval optimization [1167, 382]. Case-based reasoning systems provide theoretical foundations for LLM agent memory through architectural components that enable cognitive integration and persistent context storage techniques that implement caching strategies for faster provisioning of necessary context [383, 381]. The benefits extend beyond processing longer texts to fundamentally enhancing LLM interaction quality through improved comprehension, more relevant responses, and greater continuity across extended engagements, significantly expanding LLMs' utility and resolving limitations imposed by restricted context windows [883].

5. System Implementations

Building upon the foundational components of Context Engineering, this section examines sophisticated system implementations that integrate these components into practical, intelligent architectures. These implementations represent the evolution from theoretical frameworks to deployable systems that leverage context engineering principles. We present four major categories of system implementations. **RAG** systems demonstrate external knowledge integration through modular architectures and graph-enhanced approaches. **Memory Systems** showcase persistent context management through sophisticated memory architectures enabling long-term learning. **Tool-Integrated Reasoning** transforms language models into world interactors through function calling and environment interaction. **Multi-Agent Systems** present coordinated approaches through communication protocols and orchestration mechanisms. Each implementation builds upon foundational components while addressing specific challenges in context utilization, demonstrating how theoretical principles translate into practical systems.

5.1. Retrieval-Augmented Generation

Retrieval-Augmented Generation bridges the gap between parametric knowledge and dynamic information access by integrating external knowledge sources with language model generation. This implementation enables models to access current, domain-specific information through modular architectures, agentic frameworks, and graph-enhanced approaches that extend beyond static training data.

5.1.1. Modular RAG Architectures

Modular RAG shifts from linear retrieval-generation architectures toward reconfigurable frameworks with flexible component interaction [311, 1131, 591]. Unlike Naive RAG and Advanced RAG's query rewriting, Modular RAG introduces hierarchical architectures: top-level RAG stages, middle-level sub-modules, and bottom-level operational units [312, 730]. This transcends linear structures through routing, scheduling, and fusion mechanisms enabling dynamic reconfiguration [312].

The formal representation RAG = R, G operates through sophisticated module arrangements enabling

并动态检索相关的上下文信息，支持通过持续交互构建个性化用户模型、实现跨长时间的知识积累、有效管理长期交互中的知识，以及支持依赖历史上下文的长期规划场景 [160]。具有认知AI原理的先进记忆框架，如情境感知智能记忆（CAIM），通过支持存储和检索用户特定信息的模块，结合认知AI原理，增强长期交互，同时支持基于上下文和时间的关联性过滤 [1143]。LLM代理的记忆管理包含类似于人类记忆重组的过程，包括去重、合并和冲突解决，反射式记忆管理等方法结合前瞻性和回顾性反思，实现动态摘要和检索优化 [1167, 382]。基于案例的推理系统通过支持认知集成和持久上下文存储技术的架构组件，为LLM代理的记忆提供理论基础，这些技术实现缓存策略，以更快地提供必要的上下文[383, 381]。这些优势不仅超越了处理更长时间文本的能力，还通过提高理解能力、提供更相关的响应以及增强长期交互的连续性，从根本上提升了LLM交互质量，显著扩展了LLM的实用性，并解决了由受限上下文窗口带来的限制 [883]。

5. 系统实现

在上下文工程的基石组件之上，本节探讨了将这些组件集成到实际智能架构中的复杂系统实现。这些实现代表了从理论框架到利用上下文工程原理的可部署系统的演变。我们介绍了四种主要的系统实现类别。**RAG** 系统通过模块化架构和图增强方法展示了外部知识集成。**记忆系统** 通过复杂的记忆架构展示了持久上下文管理，支持长期学习。**工具集成推理** 通过函数调用和环境交互将语言模型转变为世界交互者。**多智能体系统** 通过通信协议和编排机制展示了协调方法。每种实现都基于基础组件，同时解决上下文利用中的特定挑战，展示了理论原理如何转化为实际系统。

5.1. 检索增强生成

检索增强生成通过将外部知识源与语言模型生成集成，弥合了参数化知识与动态信息访问之间的差距。这种实现使模型能够通过模块化架构、代理框架和图增强方法（这些方法超越了静态训练数据）访问当前、特定领域的信息。

5.1.1. 模块化 ***RAG*** 架构

模块化 RAG 从线性检索-生成架构转向可重构的框架，具有灵活的组件交互 [311, 1131, 591]。与 Naive RAG 和 Advanced RAG 的查询重写不同，模块化 RAG 引入分层架构：顶层 RAG 阶段、中层子模块和底层操作单元 [312, 730]。通过路由、调度和融合机制，这种架构超越了线性结构，实现动态重构 [312]。

形式化表示 RAG = R, G 通过复杂的模块排列操作，实现

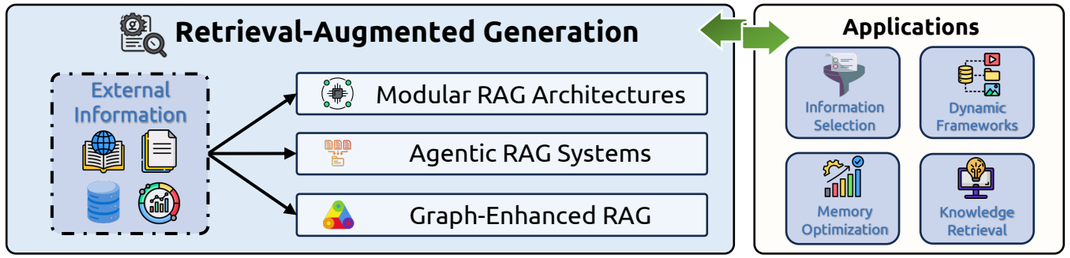

Figure 4: Retrieval-Augmented Generation Framework: Overview of RAG system architectures including Modular RAG, Agentic RAG Systems, and Graph-Enhanced RAG approaches for external context integration.

Rewrite-Retrieve-Read models and Generate-Read approaches, incorporating adaptive search modules, RAGFusion for multi-query processing, routing modules for optimal data source selection, and hybrid retrieval strategies addressing retrieval accuracy and context relevance [311, 491, 908, 1045, 880, 95].

Contemporary frameworks demonstrate significant improvements in retrieval accuracy and trustworthiness [1372]. FlashRAG provides a modular toolkit with 5 core modules and 16 subcomponents enabling independent adjustment and pipeline combination [500]. KRAGEN enhances biomedical problem-solving by integrating knowledge graphs with vector databases, utilizing biomedical knowledge graph-optimized prompt generation to address hallucination in complex reasoning [397, 749, 973]. ComposeRAG implements atomic modules for Question Decomposition and Query Rewriting, incorporating self-reflection mechanisms for iterative refinement [1159]. This modularity facilitates integration with fine-tuning and reinforcement learning, enabling customization for specific applications and comprehensive toolkits supporting diverse NLP tasks [312, 912, 4].

5.1.2. Agentic RAG Systems

Agentic RAG embeds autonomous AI agents into the RAG pipeline, enabling dynamic, context-sensitive operations guided by continuous reasoning [965, 277]. These systems leverage reflection, planning, tool use, and multi-agent collaboration to manage retrieval strategies dynamically and adapt workflows to complex task requirements [965]. RAG and agent workflows align through query rewriting corresponding to semantic comprehension, while retrieval phases correspond to planning and execution [622].

LLM-based autonomous agents extend basic language model capabilities through multimodal perception, tool utilization, and external memory integration [1160, 1091, 931, 843]. External long-term memory serves as a knowledge datastore enabling agents to incorporate and access information over extended periods [1160, 382]. Unlike static approaches, Agentic RAG treats retrieval as dynamic operation where agents function as intelligent investigators analyzing content and cross-referencing information [648, 162].

Implementation paradigms encompass prompt-based methods requiring no additional training and training-based approaches optimizing models through reinforcement learning for strategic tool invocation [648, 1318, 965]. Advanced systems enable LLM agents to query vector databases, access SQL databases, or utilize APIs within single workflows, with methodological advances focusing on reasoning capabilities, tool integration, memory mechanisms, and instruction fine-tuning for autonomous decision-making [703, 6].

Core capabilities include reasoning and planning components through task decomposition, multi-plan selection, and memory-augmented planning strategies enabling agents to break down complex tasks and

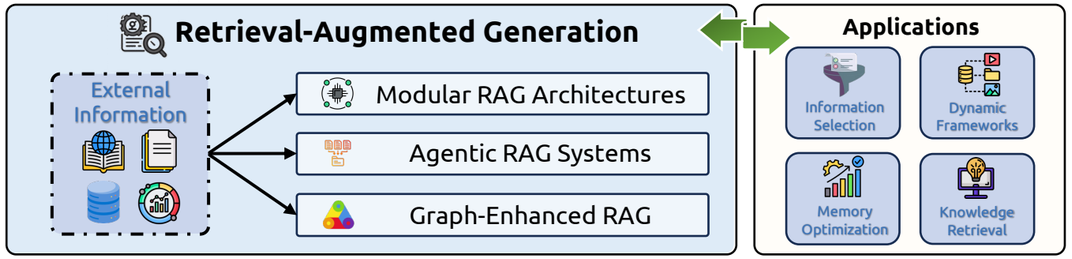

图 4：检索增强生成框架：RAG系统架构概述，包括模块化RAG、代理式RAG系统和图增强RAG方法用于外部上下文集成。

重写-检索-读取模型和生成-读取方法，结合自适应搜索模块、用于多查询处理的RAGFusion、用于最佳数据源选择的路由模块以及解决检索准确性和上下文相关性的混合检索策略 [311, 491, 908, 1045, 880, 95]。

当代框架在检索准确性和可信度方面表现出显著改进 [1372]。FlashRAG提供了一个模块化工具包，包含5个核心模块和16个子组件，支持独立调整和流程组合 [500]。KRAGEN通过将知识图谱与向量数据库集成，增强生物医学问题解决能力，利用针对生物医学知识图谱优化的提示生成来处理复杂推理中的幻觉 [397, 749, 973]。ComposeRAG实现了用于问题分解和查询重写的原子模块，结合自我反思机制进行迭代优化 [1159]。这种模块化便于与微调和强化学习集成，支持针对特定应用的定制和全面支持多样化NLP任务的工具包 [312, 912, 4]。

5.1.2. 基于代理的 ***RAG*** 系统

基于代理的 RAG 将自主 AI 代理嵌入到 RAG 管道中，通过持续的推理 [965, 277] 引导动态、上下文敏感的操作。这些系统利用反思、规划、工具使用和多代理协作来动态管理检索策略，并根据复杂任务需求调整工作流程 [965]。RAG 和代理工作流程通过与语义理解的对应查询重写而一致，而检索阶段则对应于规划和执行 [622]。

基于 LLM 的自主代理通过多模态感知、工具利用和外部内存集成扩展了基本语言模型的功能 [1160, 1091, 931, 843]。外部长期记忆作为知识数据存储，使代理能够在长时间内整合和访问信息 [1160, 382]。与静态方法不同，基于代理的 RAG 将检索视为动态操作，其中代理充当智能调查员分析内容并交叉引用信息 [648, 162]。

实现范例包括无需额外训练的基于提示的方法和通过强化学习优化模型以进行策略性工具调用的基于训练的方法 [648, 1318, 965]。高级系统使 LLM 代理能够查询向量数据库、访问 SQL 数据库或在单个工作流程中利用 API，方法学进步集中在推理能力、工具集成、记忆机制和指令微调以实现自主决策 [703, 6]。

核心能力包括通过任务分解、多计划选择和记忆增强规划策略实现推理和规划组件，使代理能够分解复杂任务和

select appropriate strategies [438, 439]. PlanRAG improves decision-making through plan-then-retrieve approaches, enabling agents to evaluate multiple information sources and optimize retrieval strategies, while SLA management frameworks address reconfigurable multi-agent architectures [162, 461]. Tool utilization enables systems to employ diverse resources including search engines, calculators, and APIs, with frameworks like ReAct and Reflexion exemplifying how interleaving reasoning with actions enhances adaptability [162, 1160, 956]. Memory mechanisms provide external long-term storage, while adaptive retrieval strategies enable autonomous analysis of complexity and context [162, 1128].

Self-reflection and adaptation mechanisms enable Agentic RAG systems to operate in dynamic environments through iterative feedback loops refining operations based on previous interaction outcomes [1183, 686]. Advanced memory systems like MemoryBank implement update mechanisms inspired by the Ebbinghaus Forgetting Curve, enhancing agents' ability to retrieve and apply learnings from past interactions [1362, 165]. CDF-RAG employs closed-loop processes combining causal graph retrieval with reinforcement learning-driven query refinement and hallucination correction [531]. Self-RAG trains models that retrieve passages on demand while reflecting on retrievals and generations, using reflection tokens to control behavior during inference [239, 41].

5.1.3. Graph-Enhanced RAG

Graph-based Retrieval-Augmented Generation shifts from document-oriented approaches toward structured knowledge representations capturing entity relationships, domain hierarchies, and semantic connections [120, 1353, 360, 1391]. This enables extraction of specific reasoning paths providing relevant information to language models while supporting multi-hop reasoning through structured pathway navigation [120]. Graph structures minimize context drift and hallucinations by leveraging interconnectivity for enhanced context-aware retrieval and logical coherence [512, 806].

Knowledge graphs serve as foundational representations encapsulating entities and interrelationships in structured formats enabling efficient querying and semantic relationship capture [162, 1058]. Graph-based knowledge representations categorize into knowledge-based GraphRAG using graphs as knowledge carriers, index-based GraphRAG employing graphs as indexing tools, and hybrid GraphRAG combining both approaches [1199]. Sophisticated implementations include GraphRAG's hierarchical indexing with community detection, PIKE's multi-level heterogeneous knowledge graphs organizing documents into three-layer hierarchies, and EMG-RAG's Editable Memory Graph architecture [313].

Graph Neural Networks enhance RAG systems by addressing limitations in handling structured knowledge, with GNNs excelling at capturing entity associations and improving knowledge consistency [228, 116]. GNN-RAG implementations adopt lightweight architectures for effective knowledge graph element retrieval, improving graph structure capture before interfacing with language models [1370, 162]. The integration process encompasses graph building through node and edge extraction, retrieval based on queries, and generation incorporating retrieved information [1370].

Multi-hop reasoning capabilities enable graph-based systems to synthesize information across multiple connected knowledge graph nodes, facilitating complex query resolution requiring interconnected fact integration [1058, 166]. These systems employ structured representations capturing semantic relationships between entities and domain hierarchies in ways that unstructured text cannot [1058, 166]. Advanced frameworks like Hierarchical Lexical Graph preserve statement provenance while clustering topics for flexible retrieval and linking entities for graph-based traversal [329]. Systems like GraphRAG, LightRAG, and derivatives implement dual-level retrieval, hierarchical indexing, and graph-enhanced strategies enabling robust multilevel reasoning [1174, 313].

选择合适的策略 [438, 439]. PlanRAG 通过计划-检索方法改进决策，使代理能够评估多个信息源并优化检索策略，而 SLA 管理框架则处理可重构的多代理架构 [162, 461]. 工具利用使系统能够使用各种资源，包括搜索引擎、计算器和 API，ReAct 和 Reflexion 等框架展示了如何将推理与行动交错以提高适应性 [162, 1160, 956]. 内存机制提供外部长期存储，而自适应检索策略能够自主分析复杂性和上下文 [162, 1128].

自我反思和适应机制使 Agentic RAG 系统能够通过迭代反馈循环在动态环境中运行，这些反馈循环根据先前的交互结果细化操作[1183, 686]。高级内存系统如 MemoryBank 实现了受艾宾浩斯遗忘曲线启发的更新机制，增强了代理从过去交互中检索和应用学习的能力[1362, 165]。CDF-RAG 采用闭环流程，结合因果图检索与强化学习驱动的查询优化和幻觉校正 [531]。Self-RAG 训练能够在需要时检索段落并对检索和生成进行反思的模型，使用反思标记在推理过程中控制行为 [239, 41]。

5.1.3. 图增强式*RAG*

基于图的检索增强生成从文档导向的方法转向结构化知识表示，捕获实体关系、领域层次和语义连接 [120, 1353, 360, 1391]。这能够提取特定的推理路径为语言模型提供相关信息，同时通过结构化路径导航支持多跳推理 [120]。图结构通过利用互连性来最小化上下文漂移和幻觉，从而增强上下文感知检索和逻辑连贯性 [512, 806]。

知识图谱作为基础表示形式，以结构化格式封装实体和相互关系，实现高效查询和语义关系捕获 [162, 1058]。基于图的 knowledge representations 分为使用图作为知识载体的知识型 GraphRAG、使用图作为索引工具的索引型 GraphRAG，以及结合这两种方法的混合型 GraphRAG [1199]。复杂的实现包括 GraphRAG 的层次索引与社区检测、PIKE 的多级异构知识图谱将文档组织成三层层次结构，以及 EMG-RAG 的可编辑记忆图架构 [313]。

图神经网络通过解决处理结构化知识时的局限性来增强 RAG 系统，GNN 在捕获实体关联和提高知识一致性方面表现出色 [228, 116]。GNN-RAG 实现采用轻量级架构，以有效检索知识图谱元素，并在与语言模型接口之前改进图结构捕获 [1370, 162]。集成过程包括通过节点和边提取构建图，基于查询进行检索，以及结合检索到的信息进行生成 [1370]。

多跳推理能力使基于图的系统能够跨多个连接的知识图谱节点综合信息，促进需要整合相互关联事实的复杂查询解析 [1058, 166]。这些系统采用结构化表示来捕获实体之间的语义关系和领域层次结构，而这种方式非结构化文本无法做到 [1058, 166]。Hierarchical Lexical Graph 等高级框架在聚类主题以实现灵活检索的同时保留语句来源，并通过图结构遍历链接实体 [329]。GraphRAG、LightRAG 及其衍生系统实现双层检索、分层索引和图增强策略，支持强大的多级推理 [1174, 313]。

Prominent architectures demonstrate diverse approaches to graph-enhanced retrieval, with optimization strategies showing significant improvements in retrieval effectiveness [106]. LightRAG integrates graph structures with vector representations through dual-level retrieval paradigms improving efficiency and content quality [412, 717]. HippoRAG leverages Personalized PageRank over knowledge graphs achieving notable improvements in multi-hop question answering [1088, 746, 366]. HyperGraphRAG proposes hypergraph structured representations advancing beyond binary relations [717]. RAPTOR provides hierarchical summary tree construction for recursive context generation, while PathRAG introduces pruning techniques for graph-based retrieval [1349, 928, 134]. These structured approaches enable transparent reasoning pathways with explicit entity connections, reducing noise and improving semantic understanding while overcoming traditional RAG challenges [1174, 512].

5.1.4. Applications

Real-time RAG systems address critical challenges in production environments where dynamic knowledge bases require continuous updates and low-latency responses [1339, 528]. Core challenges include efficient deployment and processing pipeline optimization, with existing frameworks lacking plug-and-play solutions necessitating system-level optimizations [1339]. Integration of streaming data introduces complications as traditional architectures demonstrate poor accuracy with frequently changing information and decreased efficiency as document volumes grow [514].

Dynamic retrieval mechanisms advance over static approaches by continuously updating strategies during generation, adjusting goals and semantic vector spaces in real-time based on generation states and identified knowledge gaps [384]. Current limitations in determining optimal retrieval timing and query formulation are addressed through Chain-of-Thought reasoning, iterative retrieval processes, decomposed prompting, and LLM-generated content for dynamic retrieval enabling adaptive information selection, with approaches extending to adaptive control mechanisms enhancing generation quality through reflective tags [992, 530, 85, 533, 1239].

Low-latency retrieval approaches leverage graph-based methods demonstrating significant promise in speed-accuracy optimization, with dense passage retrieval techniques providing foundational improvements [519]. LightRAG's dual-level retrieval system enhances information discovery while integrating graph structures with vector representations for efficient entity relationship retrieval, reducing response times while maintaining relevance [360]. Multi-stage retrieval pipelines optimize computational efficiency through techniques like graph-based reranking, enabling dynamic access to current information while reducing storage requirements [974].

Scalability solutions incorporate distributed processing architectures with efficient data partitioning, query optimization, and fault tolerance mechanisms adapting to changing stream conditions [1040, 35]. Memory optimization through transformed heavy hitters streaming algorithms intelligently filters irrelevant documents while maintaining quality, particularly valuable for frequently changing content [514]. Production frameworks demonstrate efficiency gains through modular RAG architectures supporting pre-retrieval processes like query expansion and post-retrieval refinements such as compression and selection, enabling fine-tuning of individual components [1069].

Incremental indexing and dynamic knowledge updates ensure systems adapt to new information without full retraining, particularly crucial in rapidly evolving domains like cybersecurity and climate finance applications [830, 1056]. Modern frameworks incorporate dynamic knowledge retrieval methods enabling continuous strategy adjustment based on evolving input and contextual information, enhancing interactivity and semantic understanding while increasing applicability across cross-domain integration [384]. Advanced

突出的架构展示了多样化的图增强检索方法，优化策略在检索有效性方面表现出显著改进 [106]。LightRAG通过双层检索范式将图结构与向量表示相结合，提高了效率和内容质量 [412, 717]。HippoRAG利用知识图上的个性化PageRank，在多跳问答方面取得了显著改进 [1088, 746, 366]。HyperGraphRAG提出了超图结构表示，超越了二元关系 [717]。RAPTOR提供分层摘要树构建用于递归上下文生成，而PathRAG引入了剪枝技术用于基于图的检索 [1349, 928, 134]。这些结构化方法支持透明的推理路径，具有明确的实体连接，减少噪声并提高语义理解，同时克服了传统RAG的挑战 [1174, 512]。

5.1.4. 应用

实时RAG系统解决了生产环境中关键挑战，动态知识库需要持续更新和低延迟响应 [1339, 528]。核心挑战包括高效的部署和处理管道优化，现有框架缺乏即插即用解决方案，需要系统级优化 [1339]。流数据的集成引入了复杂性，因为传统架构在信息频繁变化和文档量增长时准确性差且效率降低 [514]。

动态检索机制通过在生成过程中持续更新策略，根据生成状态和识别的知识空白实时调整目标和语义向量空间，从而超越静态方法 [384]。当前在确定最佳检索时机和查询公式方面的局限性通过思维链推理、迭代检索过程、分解提示和LLM生成的动态检索内容得到解决，这些方法支持自适应信息选择，并扩展到通过反思标签增强生成质量的自适应控制机制[992, 530, 85, 533, 1239]。

低延迟检索方法利用基于图的方法在速度-精度优化方面展现出显著前景，密集段落检索技术提供了基础性改进[519]。LightRAG的双级检索系统增强了信息发现，同时将图结构与向量表示集成，以实现高效的实体关系检索，减少响应时间并保持相关性 [360]。多阶段检索管道通过基于图的重新排序等技术优化计算效率，支持对当前信息的动态访问，同时减少存储需求 [974]。

可扩展性解决方案结合分布式处理架构，包括高效的数据分区、查询优化和容错机制，以适应变化的流条件 [1040, 35]。通过转换重型打击流算法进行内存优化，能够智能地过滤不相关的文档，同时保持质量，这对于经常变化的内容尤其有价值 [514]。生产框架通过支持查询扩展等预检索过程以及压缩和选择等后检索改进的模块化RAG架构，展示了效率提升，能够对单个组件进行微调 [1069]。

增量索引和动态知识更新确保系统适应新信息而无需完整重新训练，这在快速发展的领域（如网络安全和气候金融应用）中尤为重要 [830, 1056]。现代框架结合了动态知识检索方法，能够根据不断变化的输入和上下文信息持续调整策略，增强交互性和语义理解，同时提高跨领域集成的适用性 [384]。高级

agent-based approaches demonstrate sophisticated task allocation capabilities in complex environments, such as coordinated UAV operations requiring real-time decision-making, with applications extending to grounded planning for embodied agents [1315, 975]. Dynamic Retrieval Augmented Generation frameworks like DRAGON-AI showcase specialized implementations for ontology generation, combining textual and logical components while incorporating self-memory mechanisms enabling iterative improvement [1043]. These advances represent significant evolution toward seamlessly integrating real-time knowledge with flexible retrieval capabilities in dynamic environments.

5.2. Memory Systems

Memory Systems enable LLMs to transcend stateless interactions by implementing persistent information storage, retrieval, and utilization mechanisms. This implementation transforms models from pattern-matching processors into sophisticated agents capable of learning, adaptation, and long-term contextual understanding across extended interactions.

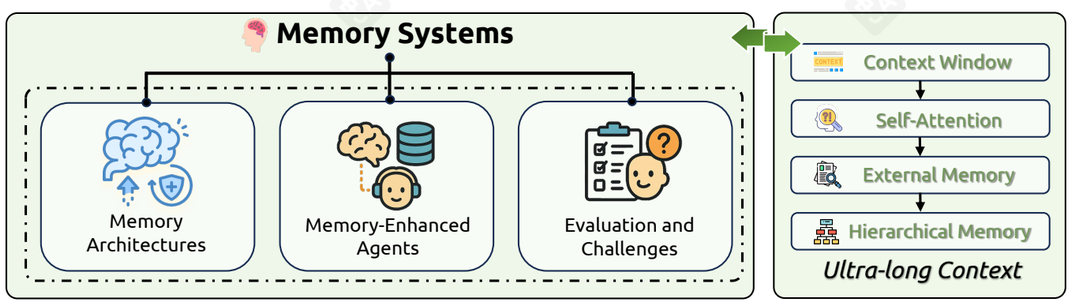

Figure 5: Memory Systems Framework: Overview of memory architectures, memory-enhanced agents, and evaluation challenges for ultra-long context processing in LLMs.

5.2.1. Memory Architectures

Memory distinguishes sophisticated language systems from pattern-matching models, enabling information processing, storage, and utilization across natural language tasks [1182, 1167, 296]. LLMs face considerable memory system constraints despite breakthroughs in text generation and multi-turn conversations [1182]. Neural memory mechanisms struggle with inadequate structured information storage and reliance on approximate vector similarity calculations rather than precise symbolic operations, challenging accurate storage and retrieval for multi-hop reasoning [423]. These limitations represent critical challenges for developing AI systems operating effectively in complex real-world applications [544].

Memory Classification Frameworks LLM memory systems can be organized into multiple classification frameworks. The primary temporal classification divides memory into three categories: sensory memory (input prompts), short-term memory (immediate context processing), and long-term memory (external databases or dedicated structures) [935]. From a persistence perspective, short-term memory includes key-value caches and hidden states existing only within single sessions, while long-term memory encompasses text-based storage and knowledge embedded in model parameters, persisting across multiple interaction cycles [935, 818].

基于代理的方法在复杂环境中展示了复杂的任务分配能力，例如需要实时决策的协调无人机操作，其应用扩展到具身代理的接地规划 [1315, 975]。动态检索增强生成框架如DRAGON-AI展示了本体生成的专门实现，结合文本和逻辑组件，同时结合自我记忆机制以实现迭代改进 [1043]。这些进步代表了在动态环境中无缝集成实时知识与灵活检索能力的显著演变。

5.2. 记忆系统

记忆系统使LLM能够通过实现持久信息存储、检索和利用机制来超越无状态交互。这种实现将模型从模式匹配处理器转变为能够学习、适应和跨长时间交互进行长期上下文理解的复杂代理。

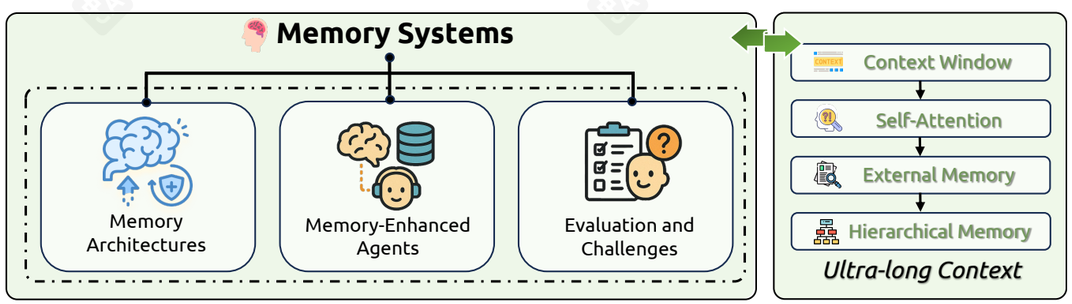

图 5：记忆系统框架：记忆架构、记忆增强代理以及LLM中超长上下文处理的评估挑战的概述。

5.2.1. 记忆架构

记忆将复杂语言系统与模式匹配模型区分开来，使信息处理、存储和利用能够在自然语言任务中实现 [1182, 1167, 296]。尽管在文本生成和多轮对话方面取得了突破，LLM仍然面临相当大的记忆系统限制 [1182]。神经记忆机制在结构化信息存储不足和依赖近似向量相似度计算而非精确符号运算方面存在困难，这挑战了多跳推理的准确存储和检索 [423]。这些限制代表了为开发在复杂现实世界应用中有效运行的AI系统所面临的重大挑战 [544]。

记忆分类框架 LLM记忆系统可以组织成多个分类框架。主要的时间分类将记忆分为三类：感觉记忆（输入提示）、短期记忆（即时上下文处理）和长期记忆（外部数据库或专用结构） [935]。从持久性角度来看，短期记忆包括仅存在于单个会话中的键值缓存和隐藏状态，而长期记忆则涵盖基于文本的存储和嵌入在模型参数中的知识，持久存在于多个交互周期中 [935, 818]。

Implementation-based classifications identify parametric memory (knowledge encoded in model weights), ephemeral activation memory (context-limited runtime states), and plaintext memory accessed through Retrieval-Augmented Generation methods [637]. Current implementations lack sophisticated lifecycle management and multi-modal integration, limiting long-term knowledge evolution. Feed-forward network layers serve as key-value tables storing memory, functioning as "inner lexicon" for word retrieval and creating mechanisms analogous to human associative memory [518, 325, 326, 764, 464]. These classification schemes reflect attempts to develop LLM memory architectures paralleling human cognitive systems [1167].

Short-Term Memory Mechanisms Short-term memory in LLMs operates through the context window, serving as working memory maintaining immediate access to previously processed tokens [1282]. This functionality is implemented through key-value caches storing token representations but disappearing when sessions terminate [891]. Architectural variations demonstrate significant differences: transformer-based models implement working memory systems flexibly retrieving individual token representations across arbitrary delays, while LSTM architectures maintain coarser, rapidly-decaying semantic representations weighted toward earliest items [40].

Modern LLM short-term memory frequently manifests as in-context learning, reflecting models' ability to acquire and process information temporarily within context windows [1180, 103]. This enables few-shot learning and task adaptation without parameter updates. Research identifies three primary memory configurations: full memory (utilizing entire context history), limited memory (using context subsets), and memory-less operation (without historical context) [1044]. Despite advances expanding context windows to millions of tokens, LLMs struggle with effective reasoning over extended contexts, particularly when relevant information appears in middle positions [891, 685].

Long-Term Memory Implementations LLMs face significant challenges maintaining long-term memory due to context window limitations and catastrophic forgetting [114]. External memory-based methods address these limitations by utilizing physical storage to cache historical information, allowing relevant history retrieval without maintaining all information within constrained context windows [682, 1362]. These approaches contrast with internal memory-based methods focusing on reducing self-attention computational costs to expand sequence length [682, 287].

Long-term memory implementations categorize into knowledge-organization methods (structuring memory into interconnected semantic networks), retrieval mechanism-oriented approaches (integrating semantic retrieval with forgetting curve mechanisms), and architecture-driven methods (implementing hierarchical structures with explicit read-write operations) [515, 1362, 444]. Memory storage representations can be further divided into token-level memory (information stored as structured text for direct retrieval) and latent-space memory (utilizing high-dimensional vectors for abstract and compact information representation) [1216, 1124]. Advanced approaches incorporate psychological principles, with MemoryBank implementing Ebbinghaus Forgetting Curve theory for selective memory preservation based on temporal factors [1362], emotion-aware frameworks employing Mood-Dependent Memory theory [444], and memorization mechanisms balancing performance advantages with privacy concerns through extraction vulnerability analysis [1041, 122, 123].

Memory Access Patterns and Structures LLMs exhibit characteristic memory access patterns with notable similarities to human cognitive processes, demonstrating clear primacy and recency effects when recalling

基于实现的分类识别了参数化内存（知识编码在模型权重中）、短暂激活内存（上下文限制的运行时状态）和通过检索增强生成方法访问的明文内存 [637]。当前的实现缺乏复杂的生命周期管理和多模态集成，限制了长期知识进化。前馈网络层作为存储内存的关键值表，充当“内部词汇”用于单词检索，并创建类似于人类联想记忆的机制 [518, 325, 326, 764, 464]。这些分类方案反映了开发与人类认知系统平行的LLM内存架构的尝试 [1167]。

短期记忆机制 LLM中的短期记忆通过上下文窗口运行，作为工作内存维护对先前处理过的标记的即时访问 [1282]。此功能通过存储标记表示的关键值缓存实现，但在会话终止时消失 [891]。架构差异显著：基于transformer的模型灵活地检索跨任意延迟的单独标记表示，而LSTM架构维护较粗、快速衰减的语义表示，更侧重于最早的项目 [40]。

现代大型语言模型的短期记忆通常表现为情境学习，反映了模型在上下文窗口 [1180, 103] 内临时获取和处理信息的能力。这使得无需更新参数即可进行少样本学习和任务适应。研究表明存在三种主要的记忆配置：全记忆（利用整个上下文历史）、有限记忆（使用上下文子集）和无记忆操作（不使用历史上下文） [1044]。尽管通过扩展上下文窗口到数百万个标记取得了进展，但大型语言模型在处理扩展上下文时仍面临挑战，尤其是在相关信息出现在中间位置时 [891, 685]。

长期记忆实现 由于上下文窗口限制和灾难性遗忘 [114]，大型语言模型在维持长期记忆方面面临重大挑战。基于外部存储的方法通过利用物理存储来缓存历史信息，从而解决这些限制，允许在不将所有信息保留在受限上下文窗口内的情况下检索相关历史信息 [682, 1362]。这些方法与基于内部记忆的方法形成对比，后者专注于通过减少自注意力计算成本来扩展序列长度 [682, 287]。

长期记忆实现方式可分为知识组织方法（将记忆结构化为相互连接的语义网络）、检索机制导向的方法（将语义检索与遗忘曲线机制相结合）和架构驱动的方法（实现具有显式读写操作的层次结构） [515, 1362, 444]。记忆存储表示可以进一步分为标记级记忆（信息作为结构化文本存储以供直接检索）和潜在空间记忆（利用高维向量进行抽象和紧凑的信息表示）[1216, 1124]。高级方法结合了心理学原理，MemoryBank 实现了艾宾浩斯遗忘曲线理论，基于时间因素进行选择性记忆保留 [1362]，情绪感知框架采用了依赖情绪的记忆理论 [444]，和通过提取脆弱性分析平衡性能优势与隐私问题的记忆机制 [1041, 122, 123]。

Memory Access Patterns and Structures LLMs表现出与人类认知过程具有显著相似性的特征性记忆访问模式，在回忆时表现出明显的首位效应和近因效应

information lists [477]. Memory retrieval operates through sequential access (retrieving content in consecutive order) and random access (accessing information from arbitrary points without processing preceding content) [1387]. Memory persistence studies employ recognition experiments, recall experiments, and retention experiments to quantify information accessibility duration and retrieval conditions [810], with cognitive psychology concepts like semantic and episodic memory integration improving LLM information synthesis capabilities [240].

Memory organization encompasses diverse structural approaches including textual-form storage (complete and recent agent-environment interactions, retrieved historical interactions, external knowledge), knowledge representation structures (chunks, knowledge triples, atomic facts, summaries, mixed approaches), hierarchical systems with library-enhanced reasoning components, and functional patterns organized by tasks, temporal relevance, or semantic relationships [1329, 1290, 1027]. Core memory operations include encoding (transforming textual information into latent space embeddings), retrieval (accessing relevant information based on semantic relevance, importance, and recency), reflection (extracting higher-level insights), summarization (condensing texts while highlighting critical points), utilization (integrating memory components for unified outputs), forgetting (selective information discarding), truncation (formatting within token limitations), and judgment (assessing information importance for storage prioritization) [1331]. These structures offer varying trade-offs between comprehensiveness, retrieval efficiency, and computational requirements.

5.2.2. Memory-Enhanced Agents

Memory systems fundamentally transform LLMs from stateless pattern processors into sophisticated agents capable of persistent learning and adaptation across extended interactions [1259]. Memory-enhanced agents leverage both short-term memory (facilitating real-time responses and immediate context awareness) and long-term memory (supporting deeper understanding and knowledge application over extended periods) to adapt to changing environments, learn from experiences, and make informed decisions requiring persistent information access [1259].

Agent Architecture Integration Contemporary LLM agents employ memory systems analogous to computer memory hierarchies, with short-term memory functioning as primary storage for contextual understanding within context windows, while long-term memory serves as persistent storage for extended information retention [770]. From object-oriented perspectives, AI systems generate personal memories related to individual users and system memories containing intermediate task results [1167]. Structured frameworks like MemOS classify memory into Parametric Memory (knowledge encoded in model weights), Activation Memory, and Plaintext Memory, with parametric memory representing long-term knowledge embedded within feedforward and attention layers enabling zero-shot generation [637].

Memory integration frameworks have evolved to address LLM limitations through sophisticated architectures. The Self-Controlled Memory (SCM) framework enhances long-term memory through LLM-based agent backbones, memory streams, and memory controllers managing updates and utilization [649]. The REMEMBERER framework equips LLMs with experience memory exploiting past episodes across task goals, enabling success/failure learning without parameter fine-tuning through verbal reinforcement and self-reflective feedback mechanisms [1299]. Advanced systems like MemLLM implement structured read-write memory modules addressing challenges in memorizing rare events, updating information, and preventing hallucinations [779]. Autonomous agents leveraging LLMs rely on four essential components—perception, memory, planning, and action—working together to enable environmental perception, interaction recall,

信息列表 [477]. 内存检索通过顺序访问（按连续顺序检索内容）和随机访问（从任意点访问信息而不处理前序内容）进行 [1387]. 内存持久性研究采用识别实验、回忆实验和保持实验来量化信息可访问持续时间及检索条件 [810], 与认知心理学概念（如语义和情景记忆整合）结合，提升LLM信息合成能力 [240]。

内存组织涵盖多种结构方法，包括文本-形式存储（完整的和最近的智能体-环境交互、检索的历史交互、外部知识）、知识表示结构（块、知识三元组、原子事实、摘要、混合方法）、具有库增强推理组件的分层系统，以及按任务、时间相关性或语义关系组织的功能模式 [1329, 1290, 1027]。核心内存操作包括编码（将文本信息转换为潜在空间嵌入）、检索（基于语义相关性、重要性和时效性访问相关信息）、反思（提取高级洞察）、摘要（压缩文本并突出关键点）、利用（整合内存组件以生成统一输出）、遗忘（选择性信息丢弃）、截断（在标记限制内进行格式化）和判断（评估信息重要性以进行存储优先级排序）[1331]。这些结构在全面性、检索效率和计算需求之间提供了不同的权衡。

5.2.2. 基于记忆的智能体

记忆系统从根本上将大型语言模型从无状态的模式处理器转变为能够进行持久学习和适应的复杂智能体[1259]。基于记忆的智能体利用短期记忆（促进实时响应和即时上下文感知）和长期记忆（支持在长时间内进行更深入的理解和知识应用）来适应变化的环境、从经验中学习，并做出需要持久信息访问的明智决策 [1259]。

智能体架构集成 当代大型语言模型智能体采用类似于计算机内存层次的记忆系统，短期记忆作为上下文窗口内情境理解的临时存储，而长期记忆则作为持久存储用于信息保留 [770]。从面向对象的角度来看，AI系统生成与单个用户相关的个人记忆和包含中间任务结果的系统记忆 [1167]。MemOS等结构化框架将记忆分为参数化记忆（知识编码在模型权重中）、激活记忆和纯文本记忆，其中参数化记忆表示嵌入在前馈和注意力层中的长期知识，支持零样本生成 [637]。

记忆集成框架通过复杂的架构进化以解决大语言模型的局限性。自控记忆（SCM）框架通过基于大语言模型的代理骨干、记忆流和记忆控制器来增强长期记忆，这些控制器管理更新和利用 [649]。REMEMBERER框架为LLM配备了经验记忆，通过利用跨任务目标的过去事件，使LLM能够在不进行参数微调的情况下通过语言强化和自我反思反馈机制实现成功/失败学习 [1299]。像MemLLM这样的高级系统实现了结构化的读写内存模块，以解决记忆罕见事件、更新信息和防止幻觉的挑战 [779]。利用LLM的自主代理依赖于四个基本组件——感知、记忆、规划和行动——协同工作以实现环境感知、交互回忆，

Model	Textual Form				Parametric Form	
	Complete	Recent	Retrieved	External	Fine-tuning	Editing
Core Memory Systems						
MemoryBank [1363]	×	×	✓	×	×	×
RET-LLM [778]	×	×	✓	×	×	×
ChatDB [423]	×	×	✓	×	×	×
TiM [683]	×	×	✓	×	×	×
Voyager [1078]	×	×	✓	×	×	×
MemGPT [814]	×	✓	✓	×	×	×
RecMind [1115]	✓	×	×	×	×	×
Retroformer [1249]	✓	×	×	✓	✓	×
ExpeL [1337]	✓	×	✓	✓	×	×
Synapse [1357]	×	×	✓	×	×	×
Agent-Based Systems						
ChatDev [855]	✓	×	×	×	×	×
InteRecAgent [450]	×	✓	✓	✓	×	×
TPTU [909, 554]	✓	×	×	✓	×	×
MetaGPT [409]	✓	×	×	×	×	×
S^3 [301]	×	×	✓	×	×	×
Mem0 [169]	×	×	✓	×	×	×
Advanced Memory Architectures						
Larimar [198]	×	✓	✓	×	×	✓
EM-LLM [286]	×	✓	✓	×	×	×
Controllable Working Memory [597]	✓	✓	✓	×	✓	×
Working Memory Hub [355]	✓	✓	✓	✓	×	×
Recent and Emerging Systems						
LLM-based Opinion Dynamics [175]	×	×	✓	×	×	×
Memory Sandbox [456]	×	×	✓	×	×	✓
A-MEM [1203]	×	×	✓	×	×	✓
MemEngine [1331]	×	×	✓	✓	×	×
HIAGENT [429]	×	✓	✓	×	×	×
MemInsight [917]	×	×	✓	✓	×	×
Memory Sharing (MS) [302]	×	×	✓	✓	×	×
MemoRAG [860]	✓	×	✓	✓	✓	×
Echo [694]	✓	✓	✓	✓	✓	×

Table 6: Extended from [1329]: Memory implementation patterns. ✓= Adopted, ×= Not Adopted

and real-time planning and execution [614, 38].

Real-World Applications Memory-enhanced LLM agents have demonstrated transformative impact across diverse application domains. In conversational AI, memory systems enable more natural, human-like interactions by recalling past experiences and user preferences to deliver personalized, context-aware responses. Commercial implementations include Charlie Mnemonic (combining Long-Term, Short-Term, and episodic memory using GPT-4), Google Gemini (leveraging long-term memory for personalized experiences across Google's ecosystem), and ChatGPT Memory (remembering conversations across sessions) [578]. User simulation applications employ LLM-powered conversational agents mimicking human behavior for cost-effective dialogue system evaluation, adapting flexibly across open-domain dialogues, task-oriented interactions, and conversational recommendation [204], with systems like Memory Sandbox enabling user control over conversational memories through data object manipulation [455].

模型	文本形式				参数形式	
	完整	最新	检索	External	Fine-tuning	Editing
核心记忆系统						
MemoryBank [1363]	×	×	✓	×	×	×
RET-LLM [778]	×	×	✓	×	×	×
ChatDB [423]	×	×	✓	×	×	×
TiM [683]	×	×	✓	×	×	×
Voyager [1078]	×	×	✓	×	×	×
MemGPT [814]	×	✓	✓	×	×	×
RecMind [1115]	✓	×	×	×	×	×
Retroformer [1249]	✓	×	×	✓	✓	×
ExpeL [1337]	✓	×	✓	✓	×	×
Synapse [1357]	×	×	✓	×	×	×
基于代理的系统						
ChatDev [855]	✓	×	×	×	×	×
InteRecAgent [450]	×	✓	✓	✓	×	×
TPTU [909, 554]	✓	×	×	✓	×	×
MetaGPT [409]	✓	×	×	×	×	×
S^3 [301]	×	×	✓	×	×	×
Mem0 [169]	×	×	✓	×	×	×
高级内存架构						
Larimar [198]	×	✓	✓	×	×	✓
EM-LLM [286]	×	✓	✓	×	×	×
可控工作内存 [597]	✓	✓	✓	×	✓	×
工作内存中心 [355]	✓	✓	✓	✓	×	×
近期及新兴系统						
基于LLM的意见动力学 [175]	×	×	✓	×	×	×
内存沙盒 [456]	×	×	✓	×	×	✓
A-MEM [1203]	×	×	✓	×	×	✓
MemEngine [1331]	×	×	✓	✓	×	×
HIAGENT [429]	×	✓	✓	×	×	×
MemInsight [917]	×	×	✓	✓	×	×
内存共享 (MS) [302]	×	×	✓	✓	×	×
MemoRAG [860]	✓	×	✓	✓	✓	×
Echo [694]	✓	✓	✓	✓	✓	×

Table 6: Extended from [1329]: Memory implementation patterns. ✓= Adopted, ×= Not Adopted

以及实时规划和执行 [614, 38].

现实世界应用 具有记忆增强功能的 LLM 代理在不同应用领域展现了变革性的影响。在对话式 AI 中，记忆系统能够通过回忆过去的经验和用户偏好来提供个性化、上下文感知的响应，从而实现更自然、更接近人类的交互。商业应用包括 Charlie Mnemonic（结合长期、短期和情景记忆使用 GPT-4）、Google Gemini（利用长期记忆在 Google 生态系统中提供个性化体验）以及 ChatGPT Memory（跨会话记忆对话） [578]。用户模拟应用采用 LLM 驱动的对话代理模仿人类行为，以经济高效的方式评估对话系统，灵活适应开放域对话、面向任务的交互和对话式推荐 [204],，系统如 Memory Sandbox 通过数据对象操作使用户能够控制对话记忆 [455]。

Task-oriented agents utilize memory to perform complex autonomous operations with minimal human intervention, employing LLMs as controllers extended through multimodal perception, tool utilization, and external memory [1160]. Applications span recommendation systems (RecMind providing personalized recommendations through planning and external knowledge, InteRecAgent employing LLMs with recommender models as tools), autonomous driving (DiLu instilling human-like knowledge through reasoning, reflection, and memory), scientific research (ChemCrow automating chemical synthesis design and execution), and social simulation (generative agents exhibiting believable behavior through memory storage and synthesis) [1019, 647, 92, 825]. Proactive conversational agents address challenges in strategic dialogue scenarios requiring goal-oriented conversation steering through prompt-based policy planning methods and AI feedback generation based on dialogue history [204, 203].

Personalized assistant applications leverage memory to maintain coherent long-term relationships with users, with memory components serving as structured repositories storing contextually relevant information including user preferences and historical interactions [438]. Domain-specific implementations include healthcare assistants employing memory coordination for medical interactions [1316, 1307], recommendation agents leveraging external knowledge bases [1316, 1293], educational agents providing context-aware support through memory-enabled progress tracking [647], and specialized frameworks like MARK enhancing personalized AI assistants through user preference memory [299].

Memory Technologies and Integration Methods Memory technology evolution addresses fundamental context window limitations through RAG, which combines parametric and non-parametric memory for language generation using pre-trained seq2seq models and dense vector indices [1209, 591]. This approach enables access to information beyond parameter storage without requiring retraining, significantly extending knowledge capabilities. Advanced memory mechanisms including vector databases and retrieval-augmented generation enable vast information storage with quick relevant data access, incorporating short-term contextual memory and long-term external storage [38, 367, 1184, 507].

Non-parametric approaches maintain frozen LLM parameters while leveraging external resources like RAG to enrich task contexts [934]. Systems like Reflexion implement verbal reinforcement through self-reflective feedback in episodic memory buffers, while REMEMBERER incorporates persistent experience memory enabling learning from past successes and failures. Advanced architectures like MemoryBank enable memory retrieval, continuous evolution through updates, and personality adaptation by integrating previous interaction information [1202, 1362].

Specialized memory architectures address particular agent requirements through sophisticated organization and retrieval mechanisms. While early systems required predefined storage structures and retrieval timing, newer systems like Mem0 incorporate graph databases following RAG principles for more effective memory organization and relevance-based retrieval [1202]. Commercial and open-source implementations including OpenAI ChatGPT Memory, Apple Personal Context, mem0, and MemoryScope demonstrate widespread adoption of memory systems for enhanced personalization capabilities [1167]. Tool-augmentation paradigms validate effectiveness in complex task decomposition while leveraging world interaction tools, with memory-enhanced agents becoming central to modern AI systems performing complex tasks through natural language integration of planning, tool use, memory, and multi-step reasoning [247, 356, 1091, 34].

5.2.3. Evaluation and Challenges

Memory evaluation frameworks have emerged as critical components for systematically assessing LLM agent capabilities across multiple dimensions, reflecting the multifaceted nature of memory in intelligent systems.

面向任务的代理利用记忆来执行复杂的自主操作，仅需少量人工干预，并采用LLM作为通过多模态感知、工具使用和外部记忆扩展的控制器 [1160]。应用涵盖推荐系统（RecMind通过规划和外部知识提供个性化推荐，InteRecAgent将LLM作为工具与推荐模型结合使用）、自动驾驶（DiLu通过推理、反思和记忆灌输类人知识）、科学研究（ChemCrow自动化化学合成设计和执行）、以及社交模拟（生成式代理通过记忆存储和合成展现逼真行为）[1019, 647, 92, 825]。主动式对话代理通过基于提示的策略规划方法和基于对话历史的AI反馈生成，解决在需要目标导向对话引导的战略对话场景中的挑战 [204, 203]。

个性化助手应用利用记忆来维持与用户连贯的长期关系，记忆组件作为结构化存储库，存储与上下文相关的信息，包括用户偏好和历史交互 [438]。特定领域的实现包括采用记忆协调进行医疗交互的健康护理助手 [1316, 1307]，利用外部知识库的推荐代理 [1316, 1293]，通过记忆增强的进度跟踪提供上下文感知支持的教育代理 [647]，以及通过用户偏好记忆增强个性化AI助手的专用框架MARK [299]。

记忆技术与集成方法 记忆技术演进通过RAG解决了基本上下文窗口限制，它结合了参数和非参数记忆，用于使用预训练的seq2seq模型和密集向量索引 [1209, 591]进行语言生成。这种方法能够在不要求重新训练的情况下访问参数存储之外的information，显著扩展了知识能力。先进的记忆机制包括向量数据库和检索增强生成，能够存储大量信息并快速访问相关数据，结合短期上下文记忆和长期外部存储 [38, 367, 1184, 507]。

非参数方法在冻结LLM参数的同时，利用外部资源如RAG来丰富任务上下文 [934]。像Reflexion这样的系统通过自我反思反馈在情景记忆缓冲区中实现语言强化，而REMEMBERER则包含持久经验记忆，能够从过去的成功和失败中学习。先进的架构如MemoryBank通过集成先前交互信息实现记忆检索、通过更新实现持续进化以及通过个性适应 [1202, 1362]。

专用内存架构通过复杂的组织和检索机制来解决特定代理的需求。早期的系统需要预定义的存储结构和检索时间，而新的系统如Mem0则遵循RAG原则采用图数据库，以实现更有效的内存组织和基于相关性的检索 [1202]。包括OpenAI ChatGPT Memory、Apple Personal Context、mem0和MemoryScope的商业和开源实现展示了内存系统在增强个性化能力方面的广泛应用 [1167]。工具增强范式在复杂任务分解中验证了其有效性，同时利用世界交互工具，内存增强代理已成为现代AI系统通过自然语言集成规划、工具使用、记忆和多步推理来执行复杂任务的核心 [247, 356, 1091, 34]。

5.2.3. 评估与挑战

内存评估框架已成为系统地评估LLM代理能力的关键组成部分，反映了智能系统中记忆的多方面特性。

These comprehensive evaluation approaches reveal significant challenges while pointing toward promising research directions that could unlock new capabilities for memory-enhanced agents.

Evaluation Frameworks and Metrics Contemporary memory evaluation employs specialized metrics extending beyond traditional NLP performance indicators to capture nuanced memory functionality aspects [1330]. Effectiveness metrics focus on factual information storage and utilization through accuracy measures (correctness of responses based on historical messages) and recall@5 indicators (percentage of relevant messages retrieved within top-5 results). Efficiency metrics examine temporal aspects through response time (duration for information retrieval and utilization) and adaptation time (period required for new information storage) [1330].

Extensive benchmarks such as LongMemEval assess five fundamental long-term memory capabilities: information extraction, temporal reasoning, multi-session reasoning, knowledge updates, and abstention through 500 carefully selected questions, demonstrating 30% accuracy degradation in commercial assistants throughout prolonged interactions, while automated memory evaluation frameworks facilitate thorough assessment extending beyond passkey search methodologies [1171]. Dedicated frameworks target episodic memory via benchmarks assessing temporally-situated experiences, with research demonstrating that cutting-edge models including GPT-4, Claude variants, and Llama 3.1 encounter difficulties with episodic memory challenges involving interconnected events or intricate spatio-temporal associations even in comparatively brief contexts [457]. Contemporary LLM benchmarks predominantly concentrate on assessing models' retention of factual information and semantic relationships while substantially overlooking episodic memory assessment—the capacity to contextualize memories with temporal and spatial occurrence details [841].

Task-specific evaluations encompass long-context passage retrieval (locating specific paragraphs within extended contexts), long-context summarization (developing comprehensive understanding for concise summaries), NarrativeQA (answering questions based on lengthy narratives), and specialized benchmarks like MADail-Bench evaluating both passive and proactive memory recall in conversational contexts with novel dimensions including memory injection, emotional support proficiency, and intimacy assessment [1329, 1380, 550, 386]. Additional task-specific frameworks include QMSum for meeting summarization, QuALITY for reading comprehension, DialSim for dialogue-based QA requiring spatiotemporal memory, and MEMENTO for personalized embodied agent evaluation using two-stage processes to assess memory utilization in physical environment tasks [1380, 566].

Current Limitations and Challenges Memory evaluation faces substantial challenges limiting effective assessment of capabilities. Fundamental limitations include absence of consistent, rigorous methodologies for assessing memory performance, particularly regarding generalization beyond training data [284]. The lack of standardized benchmarks specifically designed for long-term memory evaluation represents another significant obstacle, with existing frameworks often failing to capture the full spectrum of memory capabilities needed for human-like intelligence [1071].

Architectural constraints significantly complicate evaluation efforts, as most contemporary LLM-based agents operate in fundamentally stateless manners, treating interactions independently without truly accumulating knowledge incrementally over time [1355, 1354], despite advances in working memory through attentional tagging mechanisms enabling flexible memory representation control [864]. This limitation prevents genuine lifelong learning assessment—a cornerstone of human-level intelligence involving continuous knowledge acquisition, retention, and reuse across diverse contexts and extended time horizons.

这些综合评估方法揭示了重大挑战，同时指出了有望解锁记忆增强智能体新能力的有前景的研究方向。

评估框架和指标 当代记忆评估采用扩展了传统NLP性能指标的专用指标，以捕捉精细的记忆功能方面[1330]。有效性指标侧重于通过准确性度量（基于历史消息的响应正确性）和召回@5指标（在顶部5个结果中检索到的相关消息的百分比）来关注事实信息的存储和利用。效率指标通过响应时间（信息检索和利用的持续时间）和适应时间（存储新信息所需的时间）来检查时间方面 [1330]。

如LongMemEval等广泛的基准评估了五个基本长期记忆能力：信息提取、时间推理、多会话推理、知识更新和通过500个精心选择的问题进行的回避，显示在整个长时间交互过程中商业助手的准确率降低了30%，而自动记忆评估框架促进了超越密钥搜索方法的全面评估 [1171]。专用框架通过评估时间定位体验的基准来针对情景记忆，研究表明包括GPT-4、Claude变体和Llama 3.1在内的尖端模型即使在相对较短的上下文中也难以处理涉及相互关联事件或复杂时空关联的情景记忆挑战 [457]。当代LLM基准主要集中于评估模型保留事实信息和语义关系的能力，而大幅忽略了情景记忆评估——即根据时间和空间发生细节对记忆进行上下文化的能力 [841]。

任务特定的评估包括长上下文段落检索（在扩展的上下文中定位特定段落）、长上下文摘要（为简洁摘要开发全面的理解）、叙事QA（基于长篇叙述回答问题）以及MADail-Bench等专门基准，后者评估对话环境中被动和主动记忆召回，并包含记忆注入、情感支持熟练度和亲密关系评估等新维度[1329, 1380, 550, 386]。其他任务特定的框架包括用于会议摘要的QMSum、用于阅读理解的QuALITY、需要时空记忆的DialSim对话式QA，以及使用两阶段过程评估物理环境任务中记忆利用率的MEMENTO个性化具身智能体评估 [1380, 566]。

当前局限与挑战 记忆评估面临重大挑战，限制了能力的有效评估。基本局限包括缺乏评估记忆性能的一致、严格的方法，尤其是在训练数据之外的泛化方面 [284]。缺乏专门为长期记忆评估设计的标准化基准是另一个重大障碍，现有框架往往无法捕捉到实现类人智能所需的记忆能力全谱 [1071]。

架构约束显著增加了评估工作的复杂性，因为大多数当代基于LLM的代理以根本无状态的方式运行，将交互独立处理，而没有真正随着时间的推移逐步积累知识 [1355, 1354]，尽管通过注意力标记机制在工作记忆方面取得了进步，从而实现了灵活的内存表示控制 [864]。这种限制阻碍了真正的终身学习评估——这是人类水平智能的核心，涉及在多样化的环境和延长的时间范围内持续获取、保留和重用知识。

Methodological issues arise when isolating memory-specific performance from other intelligence aspects, challenging determination of whether failures stem from inadequate memory mechanisms or reasoning limitations [284]. Dynamic memory usage in real-world applications poses evaluation challenges, as controlled laboratory tests inadequately capture memory system performance in complex scenarios where information relevance changes unpredictably [1071].

Optimization Strategies and Future Research Directions Memory optimization encompasses diverse techniques enhancing utilization while minimizing computational overhead and maximizing efficiency. Biologically-inspired forgetting mechanisms provide effective optimization approaches, with frameworks like MemoryBank implementing Ebbinghaus forgetting curves to selectively preserve and discard information based on temporal factors and significance [1362]. Reflection-based optimization through systems like Reflexion enables performance assessment through integrated evaluation and self-reflection, creating dual feedback systems refining memory and behavior through continuous learning [300].

Hierarchical memory structures optimize information organization through multi-level formats enabling efficient retrieval, demonstrated by Experience-based Hierarchical Control frameworks with rapid memory access modules [862], memory consolidation processes through bidirectional fast-slow variable interactions [63], and Adaptive Cross-Attention Networks dynamically ranking memories based on query relevance [406].

Future research directions encompass hybrid memory frameworks combining parametric precision with non-parametric efficiency [934], automated feedback mechanisms for scalable response evaluation [885], multi-agent memory systems enabling collaborative learning through shared external memories [302], enhanced metadata learning with knowledge graph integration [888, 382], domain-specific memory architectures for specialized applications [501], cognitive-inspired optimization incorporating memory consolidation during inactive periods [752], and parameter-efficient memory updates through techniques like Low-Rank Adaptation for efficient knowledge integration [424, 252]. These developments promise advancing memory-enhanced LLM agents toward sophisticated, human-like cognitive capabilities while addressing computational and architectural limitations, with applications extending to long-term robotic planning, real-world decision-making systems, and collaborative AI assistants through streaming learning scenarios and continuous feedback integration [1150, 1336, 1269].

5.3. Tool-Integrated Reasoning

Tool-Integrated Reasoning transforms language models from passive text generators into active world interactors capable of dynamic tool utilization and environmental manipulation. This implementation enables models to transcend their inherent limitations through function calling mechanisms, integrated reasoning frameworks, and sophisticated environment interaction capabilities.

5.3.1. Function Calling Mechanisms

Function calling transforms LLMs from generative models into interactive agents through structured output generation leveraging functions' abstraction mechanism, enabling external tool manipulation and access to current, domain-specific information for complex problem-solving [5, 663, 331, 874, 58, 517, 1104].

Evolution began with Toolformer's self-supervised approach demonstrating autonomous API learning, inspiring ReAct's "thought-action-observation" cycle, progressing through specialized models like Gorilla and comprehensive frameworks including ToolLLM, RestGPT, with OpenAI's JSON standardization, while

在将特定于内存的性能与其他智能方面隔离时会出现方法论问题，这挑战了确定故障是否源于内存机制不足或推理限制的判断 [284]。现实世界应用中的动态内存使用提出了评估挑战，因为受控的实验室测试无法充分捕捉在信息相关性不可预测地变化的复杂场景中内存系统性能 [1071]。

优化策略和未来研究方向内存优化包含多种技术，旨在提高利用率同时最小化计算开销并最大化效率。受生物学启发的遗忘机制提供了有效的优化方法，例如MemoryBank实现艾宾浩斯遗忘曲线，根据时间因素和重要性选择性地保留和丢弃信息 [1362]。通过Reflexion等系统进行的基于反射的优化，通过集成评估和自我反思进行性能评估，创建双重反馈系统，通过持续学习不断改进内存和行为 [300]。

分层内存结构通过多级格式优化信息组织，实现高效检索，例如基于经验的分层控制框架，具有快速内存访问模块 [862]，通过双向快慢变量交互进行记忆巩固过程[63]，以及动态根据查询相关性对记忆进行排序的自适应交叉注意力网络 [406]。

未来的研究方向包括结合参数化精度与非参数化效率的混合内存框架 [934]，用于可扩展响应评估的自动反馈机制 [885]，通过共享外部内存实现协作学习的多智能体内存系统 [302]，与知识图谱集成以增强元数据学习的增强元数据学习 [888, 382]，用于专业应用的特定领域内存架构 [501]，结合记忆巩固以在非活动期间进行认知启发优化的优化 [752]，以及通过低秩自适应等技术进行参数高效内存更新以实现高效知识集成的更新 [424, 252]。这些发展有望将内存增强的 LLM 智能体推向复杂的、类似人类的认知能力，同时解决计算和架构限制，应用扩展到长期机器人规划、现实世界决策系统和协作式 AI 助手，通过流式学习场景和持续反馈集成 [1150, 1336, 1269]。

5.3. 工具集成推理

工具集成推理将语言模型从被动的文本生成器转变为能够动态利用工具和环境交互的主动世界交互者。这种实现使模型能够通过函数调用机制、集成推理框架和复杂的环境交互能力超越其固有限制。

5.3.1. 函数调用机制

函数调用通过利用函数的抽象机制生成结构化输出，将大型语言模型从生成模型转变为交互式代理，从而实现外部工具的操作和对当前、特定领域信息的访问，以解决复杂问题 [5, 663, 331, 874, 58, 517, 1104]。

进化始于 Toolformer 的自监督方法，展示了自主 API 学习，启发了 ReAct 的“思考-行动-观察”循环，通过像 Gorilla 这样的专用模型和包括 ToolLLM、RestGPT 在内的综合框架发展，同时

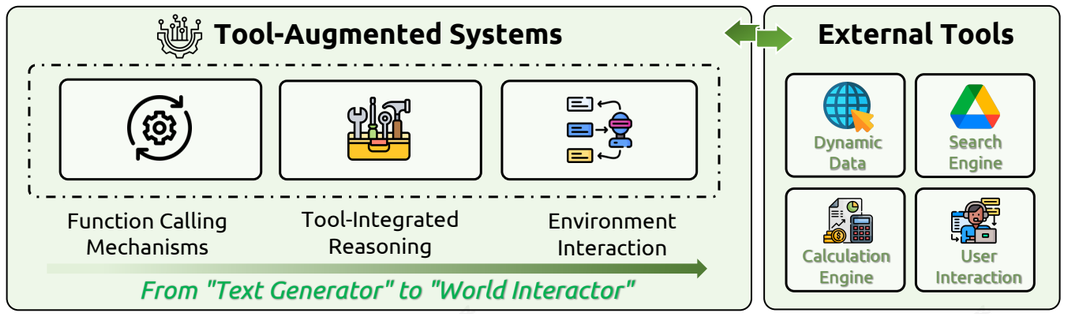

Figure 6: Tool-Augmented Systems Framework: Evolution from text generators to world interactors through function calling mechanisms, tool-integrated reasoning, and environment interaction capabilities.

advanced systems like Chameleon enabled multimodal question answering and TaskMatrix.AI managed AI models across domains [931, 248, 648, 541, 915, 866, 867, 709, 653, 945].

Technical implementation involves fine-tuning (dominant method providing stable capabilities via extensive API training but requiring significant resources) and prompt engineering (flexible, resource-efficient but unstable), with approaches like "Reverse Chain" enabling API operation via prompts, addressing challenges in large tool management [388, 5, 1323, 785, 144, 250].

Core process encompasses intent recognition, function selection, parameter-value-pair mapping, function execution, and response generation, with modern implementations utilizing structured LLM outputs for external program interaction, while tools include diverse interfaces (digital systems, scratch pads, user interactions, other LLMs, developer code), requiring complex navigation of tool selection, argument formulation, and result parsing [1259, 663, 1132, 189, 952, 584, 902].

Training Methodologies and Data Systems Training methodologies evolved from basic prompt-based approaches to sophisticated multi-task learning frameworks, with fine-tuning on specialized datasets through systems like ToolLLM and Granite-20B-FunctionCalling, beginning with synthetic single-tool data followed by human annotations [388, 5, 353, 771, 1226].

Data generation strategies include Weaver's GPT-4-based environment synthesis, APIGen's hierarchical verification pipelines (format checking, function execution, semantic verification), generating 60,000+ high-quality entries across thousands of APIs [1104, 1177, 1259, 1156, 65, 1393, 743].

Tool selection enhancement involves irrelevance-aware data augmentation, with Hammer's function masking techniques, oracle tool mixing for increased difficulty, tool intent detection synthesis for over-triggering mitigation, emphasizing high-quality data through stringent filtering and format verification [664, 10, 353, 467, 1291, 214].

Self-improvement paradigms reduce external supervision dependence through JOSH algorithm's sparse reward simulation environments and TTPA's token-level optimization with error-oriented scoring, demonstrating improvements while preserving general capabilities [573, 440, 362, 1262].

Sophisticated benchmarks include API-Bank (73 APIs, 314 dialogues), StableToolBench (API instability solutions), NesTools (nested tool evaluation), ToolHop (995 queries, 3,912 tools), addressing single-tool to

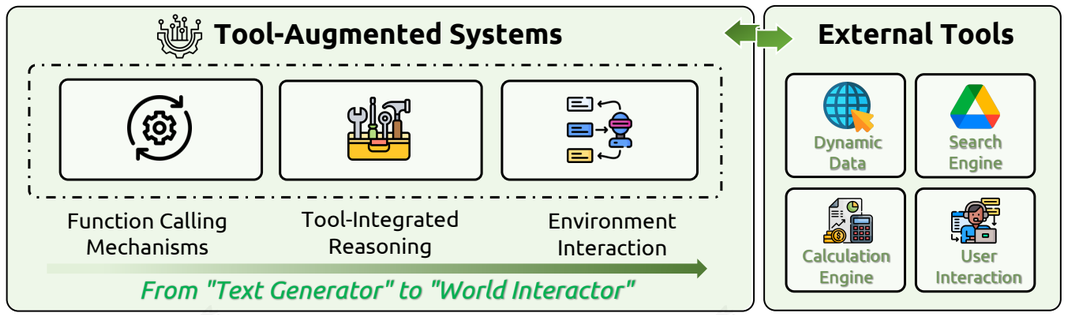

图 6：工具增强系统框架：通过函数调用机制、工具集成推理和环境交互能力，从文本生成器进化到世界交互器。

像Chameleon这样的高级系统支持多模态问答，而TaskMatrix.AI管理跨领域的AI模型 [931, 248, 648, 541, 915, 866, 867, 709, 653, 945]。

技术实现涉及微调（主要方法通过大量API训练提供稳定能力，但需要大量资源）和提示工程（灵活、资源高效但不稳定），如"反向链"等方法通过提示实现API操作，解决大型工具管理中的挑战 [388, 5, 1323, 785, 144, 250]。

核心过程包括意图识别、函数选择、参数-值对映射、函数执行和响应生成，现代实现利用结构化LLM输出来进行外部程序交互，工具包括多样化的接口（数字系统、草稿板、用户交互、其他LLMs、开发者代码），需要复杂地导航工具选择、参数制定和结果解析 [1259, 663, 1132, 189, 952, 584, 902]。

训练方法与数据系统 训练方法从基本的基于提示的方法发展到复杂的多任务学习框架，通过ToolLLM和Granite-20B-FunctionCalling等系统在专业数据集上进行微调，从合成单工具数据开始，然后进行人工标注 [388, 5, 353, 771, 1226]。

数据生成策略包括Weaver的基于GPT-4的环境合成、APIGen的分层验证管道（格式检查、函数执行、语义验证），生成60,000+高质量的条目跨越数千个API [1104, 1177, 1259, 1156, 65, 1393, 743]。

工具选择增强涉及无关性感知数据增强，包括Hammer的功能掩码技术、预言机工具混合以增加难度、工具意图检测合成以减少误触发，通过严格的过滤和格式验证强调高质量数据 [664, 10, 353, 467, 1291, 214]。

自我改进范式通过JOSH算法的稀疏奖励模拟环境以及TTPA的基于错误的评分的token级优化减少外部监督依赖，在保持通用能力的同时展示改进 [573, 440, 362, 1262]。

复杂的基准测试包括 API-Bank（73 个 API，314 个对话）、StableToolBench（API 不稳定性解决方案）、NesTools（嵌套工具评估）、ToolHop（995 个查询，3,912 个工具），解决单个工具到

multi-hop scenarios [615, 359, 373, 1255, 821, 987, 1248, 979].

5.3.2. Tool-Integrated Reasoning

Tool-Integrated Reasoning (TIR) represents a paradigmatic advancement in Large Language Model capabilities, addressing fundamental limitations including outdated knowledge, calculation inaccuracy, and shallow reasoning by enabling dynamic interaction with external resources during the reasoning process [858]. Unlike traditional reasoning approaches that rely exclusively on internal model knowledge, TIR establishes a synergistic relationship where reasoning guides complex problem decomposition into manageable subtasks while specialized tools ensure accurate execution of each computational step [771]. This paradigm extends beyond conventional text-based reasoning by requiring models to autonomously select appropriate tools, interpret intermediate outputs, and adaptively refine their approach based on real-time feedback [858].

The evolution of TIR methodologies encompasses three primary implementation categories addressing distinct aspects of tool utilization optimization. Prompting-based methods guide models through carefully crafted instructions without additional training, exemplified by approaches that decompose mathematical problems into executable code while delegating computation to Python interpreters [152, 595]. Supervised fine-tuning approaches teach tool usage through imitation learning, with systems like ToRA focusing on mathematical problem-solving by integrating natural language reasoning with computational libraries and symbolic solvers [341]. Reinforcement learning methods optimize tool-use behavior through outcome-driven rewards, though current implementations often prioritize final correctness without considering efficiency, potentially leading to cognitive offloading phenomena where models over-rely on external tools [223].

In operational terms, TIR-based agents serve as intelligent orchestrators that systematically interweave cognitive processing with external resource engagement to achieve targeted outcomes [1087]. This mechanism requires the harmonious integration of intrinsic reasoning capabilities and extrinsic tool utilization for progressive knowledge synthesis toward objective fulfillment, where the agent's execution pathway is formally characterized as a structured sequence of tool activations coupled with corresponding information assimilation events [1087]. Emerging developments have established Agentic Reasoning architectures that amplify language model intelligence by incorporating autonomous tool-deploying agents, fluidly orchestrating web-based information retrieval, computational processing, and layered reasoning-memory integration to tackle sophisticated challenges necessitating comprehensive research and cascaded logical analysis [1153].

Implementation Frameworks and Paradigms Single-tool frameworks established foundational principles of tool-integrated reasoning through specialized implementations targeting specific computational domains. Program-Aided Language Models (PAL) pioneered problem decomposition strategies by generating executable code while delegating mathematical computations to Python interpreters [305]. ToolFormer demonstrated that language models could learn external API usage with minimal demonstrations, incorporating calculators, search engines, and diverse tools to enhance computational capabilities [931]. ToRA advanced mathematical reasoning by integrating natural language processing with computational libraries and symbolic solvers, while ReTool applied reinforcement learning to optimize code interpreter usage, demonstrating improvements in self-correction patterns [341, 1311, 965]. Self-Edit utilizes execution results of generated code to improve code quality for competitive programming tasks, employing a fault-aware code editor to correct errors based on test case results [1309].

Multi-tool coordination systems address the complexity of orchestrating heterogeneous tools within integrated reasoning architectures. ReAct pioneered the interleaving of reasoning traces with task-specific actions, enabling models to think and act complementarily where reasoning supports plan tracking while

多跳场景 [615, 359, 373, 1255, 821, 987, 1248, 979]。

5.3.2. 工具集成推理

工具集成推理（TIR）代表了大型语言模型能力的范式性进步，通过在推理过程中启用与外部资源的动态交互，解决了包括过时知识、计算不准确性和浅层推理在内的基本限制 [858]。与完全依赖内部模型知识的传统推理方法不同，TIR建立了一种协同关系，其中推理指导将复杂问题分解为可管理的子任务，而专业工具确保每个计算步骤的准确执行 [771]。这种范式通过要求模型自主选择适当工具、解释中间输出并根据实时反馈自适应地调整其方法，超越了传统的基于文本的推理 [858]。

TIR方法论的演变涵盖了三种主要实施类别，分别针对工具利用优化的不同方面。提示方法通过精心设计的指令引导模型，无需额外训练，例如将数学问题分解为可执行代码并将计算委托给Python解释器 [152, 595]。监督微调方法通过模仿学习教授工具使用，如ToRA专注于通过将自然语言推理与计算库和符号求解器集成来解决数学问题 [341]。强化学习方法通过结果驱动的奖励优化工具使用行为，尽管当前实现通常优先考虑最终正确性而不考虑效率，可能导致认知卸载现象，即模型过度依赖外部工具 [223]。

在操作层面，基于TIR的智能体充当智能协调器，系统地交织认知处理与外部资源参与，以实现目标结果 [1087]。这种机制要求内在推理能力与外部工具利用的和谐集成，以逐步合成知识并实现目标，其中智能体的执行路径被正式描述为一系列结构化的工具激活序列，伴随着相应的信息吸收事件 [1087]。新兴发展已经建立了Agentic Reasoning架构，通过整合自主部署的智能体来增强语言模型的智能，灵活地协调基于网络的信息检索、计算处理和分层推理-记忆集成，以应对需要全面研究和级联逻辑分析的综合挑战 [1153]。

实现框架和范式 单工具框架通过针对特定计算领域的专门实现，建立了工具集成推理的基础原则。程序辅助语言模型(PAL)通过生成可执行代码而将数学计算委托给Python解释器，开创了问题分解策略 [305]。ToolFormer证明了语言模型可以学习外部API使用，通过最小量的演示结合计算器、搜索引擎和多样化工具来增强计算能力 [931]。ToRA通过将自然语言处理与计算库和符号求解器集成，提升了数学推理能力，而ReTool应用强化学习来优化代码解释器的使用，展示了自我纠正模式的改进 [341, 1311, 965]。Self-Edit利用生成的代码的执行结果来提高代码质量，用于竞赛编程任务，采用故障感知代码编辑器根据测试用例结果纠正错误 [1309]。

多工具协调系统解决了在集成推理架构中协调异构工具的复杂性。ReAct开创了推理轨迹与特定任务动作的交错，使模型能够互补地思考和行动，其中推理支持计划跟踪，而

actions interface with external information sources [1245]. Chameleon introduced plug-and-play compositional reasoning by synthesizing programs combining vision models, search engines, and Python functions with an LLM-based planner core [709]. AutoTools established automated frameworks transforming raw tool documentation into executable functions, reducing manual engineering requirements in tool integration [419, 952]. Chain-of-Agents (CoA) trains models to decode reasoning chains with abstract placeholders, subsequently calling domain-specific tools to fill knowledge gaps [594, 1327].

Agent-based frameworks represent the most sophisticated evolution of TIR systems, moving beyond static prompting approaches to create autonomous and adaptive AI systems. Unlike conventional tool-use that follows rigid patterns, agent models learn to couple Chain-of-Thought (CoT) and Chain-of-Action (CoA) patterns into their core behavior, resulting in stronger logical coherence and natural transitions between reasoning and action [1328]. These systems build upon foundational agent architectures including reactive systems that map perceptions directly to actions, deliberative systems implementing Belief-Desire-Intention (BDI) models, and hybrid architectures combining multiple subsystems in hierarchical structures [728].

Method	Tool Categories							
	Search & Retrieval	Computation & Code Execution	Knowledge Base & QA	APIs & External Services	Multimodal Tools	Language Processing	Interactive Environments	Domain-Specific Tools
ReAct [1247]	✓		✓				✓	
Toolformer [931]	✓	✓	✓			✓		✓
ToolkenGPT [378]	✓	✓	✓	✓			✓	
ToolLLM [867]	✓	✓	✓	✓	✓	✓	✓	✓
ToRA [341]		✓						
PAL [303]		✓						
HuggingGPT [945]				✓	✓			
GPT4Tools [1225]					✓			
CRITIC [340]	✓	✓	✓					
Chain of Code [595]		✓						
TRICE [863]	✓	✓	✓			✓		
TP-LLaMA [149]	✓	✓	✓	✓	✓	✓	✓	✓
AlignToolLLaMA [161]	✓	✓	✓	✓	✓	✓	✓	✓
ReTool [270]		✓						
Tool-Star [221]	✓	✓						
ARTIST [965]		✓						
Ego-R1 [1038]					✓			
VTool-R1 [1155]					✓			
KG-Agent [487]			✓					✓
CACTUS [755]								✓
MuMath-Code [1265]		✓						
ToRL [621]		✓						
MetaTool [452]	✓	✓	✓	✓				
ToolEyes [1253]				✓				✓
Graph-CoT [495]			✓					✓
ToolRL [858]	✓	✓	✓	✓				
LATS [1364]	✓							✓

Table 7: Tool-augmented language model architectures: Comparison of multiple methods across 8 tool categories including search, computation, knowledge bases, APIs, multimodal, language tools, interactive environments, and domain-specific applications.

5.3.3. Agent-Environment Interaction

Reinforcement learning approaches have emerged as superior alternatives to prompting-based methods and supervised fine-tuning for tool integration, enabling models to autonomously discover optimal tool usage strategies through exploration and outcome-driven rewards [223]. ReTool exemplifies this advancement

动作接口与外部信息源交互 [1245]. Chameleon通过合成结合视觉模型、搜索引擎和Python函数的基于LLM的规划器核心，引入了即插即用的组合推理 [709]. AutoTools建立了将原始工具文档转换为可执行函数的自动化框架，减少了工具集成的手动工程需求[419, 952]. Chain-of-Agents (CoA) 训练模型解码具有抽象占位符的推理链，随后调用特定领域的工具来填补知识空白 [594, 1327].

基于代理的框架代表了TIR系统的最复杂进化，超越了静态提示方法，创建了自主和自适应的AI系统。与遵循刚性模式的传统工具使用不同，代理模型学习将思维链（CoT）和行动链（CoA）模式耦合到其核心行为中，从而产生更强的逻辑一致性和推理与行动之间的自然过渡 [1328]。这些系统基于基础代理架构，包括将感知直接映射到动作的反射系统、实现信念-欲望-意图（BDI）模型的推理系统，以及结合多个子系统在分层结构中的混合架构 [728]。

方法	工具类别							
	搜索 & 检索	计算 & 代码执行	知识库 & QA	APIs & 外部服务	多模态 工具	语言 处理	Interactive Environments	领域- Specific Tools
ReAct [1247]	✓		✓				✓	
Toolformer [931]	✓	✓	✓			✓		✓
ToolkenGPT [378]	✓	✓	✓	✓			✓	
ToolLLM [867]	✓	✓	✓	✓	✓	✓	✓	✓
ToRA [341]		✓						
PAL [303]		✓						
HuggingGPT [945]				✓	✓			
GPT4Tools [1225]					✓			
CRITIC [340]	✓	✓	✓					
Chain of Code [595]		✓						
TRICE [863]	✓	✓	✓			✓		
TP-LLaMA [149]	✓	✓	✓	✓	✓	✓	✓	✓
AlignToolLLaMA [161]	✓	✓	✓	✓	✓	✓	✓	✓
ReTool [270]		✓						
Tool-Star [221]	✓	✓						
ARTIST [965]		✓						
Ego-R1 [1038]					✓			
VTool-R1 [1155]					✓			
KG-Agent [487]			✓					✓
CACTUS [755]								✓
MuMath-Code [1265]		✓						
ToRL [621]		✓						
MetaTool [452]	✓	✓	✓	✓				
ToolEyes [1253]				✓				✓
Graph-CoT [495]			✓					✓
ToolRL [858]	✓	✓	✓	✓				
LATS [1364]	✓							✓

表 7：工具增强的语言模型架构：比较 8 个工具类别（包括搜索、计算、知识库、API、多模态、语言工具、交互式环境以及特定领域应用）中的多种方法。

5.3.3. Agent-Environment Interaction

强化学习方法已成为工具集成的首选替代方案，超越了基于提示的方法和监督微调，使模型能够通过探索和结果驱动的奖励自主发现最优工具使用策略 [223]。ReTool 体现了这一进步

by focusing on code interpreter optimization for mathematical reasoning, achieving 67.0% accuracy on AIME2024 benchmarks after only 400 training steps, substantially outperforming text-based RL baselines reaching 40.0% accuracy with extensive training [270]. This demonstrates that explicitly modeling tool use within decision processes enhances both reasoning capabilities and training efficiency.

Search-augmented reasoning systems represent innovative integrations of information retrieval directly into reasoning processes through specialized learning environments. The Search-R1 framework trains models to make dynamic decisions about when to search and what queries to generate during multi-step reasoning tasks, unlike traditional retrieval-augmented generation systems [976]. The architecture employs specialized token systems structuring reasoning and search processes, where models learn to generate reasoning steps interspersed with explicit search actions triggered through tokens that encapsulate generated queries [648].

Multi-turn and customizable tool invocation frameworks address the complexity of coordinating multiple heterogeneous tools during reasoning processes. Recent developments include frameworks like VisTA that use reinforcement learning to enable visual agents to dynamically explore, select, and combine tools from diverse libraries based on empirical performance [454]. ReVeal demonstrates self-evolving code agents via iterative generation-verification processes [506]. In multimodal domains, systems like VideoAgent employ vision-language foundation models as tools for translating and retrieving visual information, achieving impressive performance on video understanding benchmarks [1108, 254].

Evaluation and Applications Comprehensive evaluation of tool-integrated reasoning systems requires specialized benchmarks that measure tool-integrated capabilities rather than general model performance. MCP-RADAR provides a standardized evaluation framework employing strictly objective metrics derived from quantifiable performance data, with extensible design spanning software engineering, mathematical reasoning, and general problem-solving domains [310]. The framework visualizes performance through radar charts highlighting model strengths and weaknesses across multiple dimensions, enabling systematic comparison of tool-integrated language models regardless of implementation mechanisms.

Real-world evaluation approaches reveal significant performance gaps between current systems and human-level capabilities, providing crucial insights into practical limitations and optimization opportunities. The General Tool Agents (GTA) benchmark addresses limitations in existing evaluations by featuring real human-written queries with implicit tool-use requirements, evaluation platforms with deployed tools across perception, operation, logic, and creativity categories, and authentic multimodal inputs including images and code snippets [1090]. Results demonstrate substantial challenges for current LLMs, with GPT-4 completing less than 50

Function calling enabled sophisticated multi-agent systems where multiple LLM agents collaborate through coordinated tool use and task decomposition, with MAS leveraging collective intelligence through parallel processing, information sharing, and adaptive role assignment, while LLM integration enhanced capabilities in planning, specialization, and task decomposition through frameworks like DyLAN, MAD, and MetaGPT [239, 903, 344, 140, 625]. Advanced multi-agent function calling employs sophisticated orchestration mechanisms decomposing complex tasks into manageable subtasks, with fundamental approaches involving splitting reward machines into parallel execution units, each agent maintaining individual reward machines, local state spaces, and propositions, while adaptive orchestration enables dynamic agent selection based on context, responses, and status reports [39, 1048, 691, 117].

通过专注于数学推理的代码解释器优化，在仅400个训练步骤后就在AIME2024基准测试中达到67.0%的准确率，大幅优于需要大量训练才能达到40.0%准确率的基于文本的RL基线 [270]。这表明在决策过程中显式建模工具使用可以增强推理能力和训练效率。

搜索增强推理系统代表了通过专门的学习环境将信息检索直接集成到推理过程中的创新整合。Search-R1框架训练模型在多步推理任务中动态决定何时搜索以及生成什么查询，这与传统的检索增强生成系统不同 [976]。该架构采用专门构建的标记系统来结构化推理和搜索过程，其中模型学习生成推理步骤，这些步骤中穿插着通过封装生成查询的标记触发的显式搜索动作 [648]。

多轮和可定制的工具调用框架解决了在推理过程中协调多个异构工具的复杂性。最近的进展包括使用强化学习来使视觉代理能够根据经验性能动态探索、选择和组合来自不同库的工具的框架，如VisTA [454]。ReVeal通过迭代生成-验证过程展示了自我演化的代码代理 [506]。在多模态领域，像VideoAgent这样的系统使用视觉-语言基础模型作为翻译和检索视觉信息的工具，在视频理解基准测试中取得了令人印象深刻的性能 [1108, 254]。

评估与应用 对工具集成推理系统的全面评估需要专门的基准测试，这些基准测试衡量的是工具集成的能力，而不是通用模型性能。MCP-RADAR提供了一个标准化的评估框架，该框架采用严格客观的指标，这些指标源自可量化的性能数据，其设计具有可扩展性，涵盖软件工程、数学推理和通用问题解决领域 [310]。该框架通过雷达图可视化性能，突出模型在多个维度上的优势和劣势，从而能够不受实现机制的限制，对工具集成的语言模型进行系统比较。

现实世界的评估方法揭示了当前系统与人类水平能力之间的显著性能差距，为实际限制和优化机会提供了关键见解。通用工具代理（GTA）基准测试通过采用人类编写的真实查询（带有隐含的工具使用要求）、部署了感知、操作、逻辑和创造力类别的工具的评估平台，以及包括图像和代码片段的真实多模态输入 [1090]，解决了现有评估的局限性。结果表明，当前的大型语言模型面临重大挑战，GPT-4完成的不到50

函数调用使复杂的多代理系统得以实现，其中多个 LLM 代理通过协调的工具使用和任务分解进行协作，多代理系统（MAS）通过并行处理、信息共享和自适应角色分配利用集体智能，而 LLM 集成通过 DyLAN、MAD 和 MetaGPT 等框架增强了规划、专业化和任务分解的能力 [239, 903, 344, 140, 625]。高级多代理函数调用采用复杂的编排机制，将复杂任务分解为可管理的子任务，基本方法涉及将奖励机拆分为并行执行单元，每个代理维护独立的奖励机、本地状态空间和命题，而自适应编排能够根据上下文、响应和状态报告动态选择代理 [39, 1048, 691, 117]。

5.4. Multi-Agent Systems

Multi-Agent Systems represent the pinnacle of collaborative intelligence, enabling multiple autonomous agents to coordinate and communicate for solving complex problems beyond individual agent capabilities. This implementation focuses on sophisticated communication protocols, orchestration mechanisms, and coordination strategies that enable seamless collaboration across diverse agent architectures.

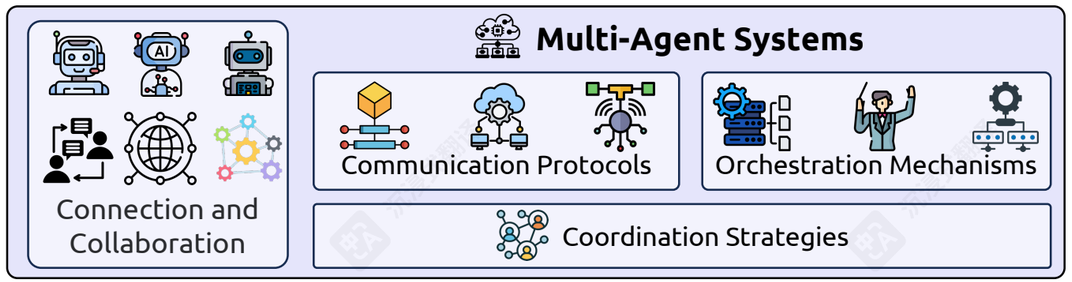

Figure 7: Multi-Agent Systems Framework: Overview of communication protocols, orchestration mechanisms, and coordination strategies for collaborative AI agent systems.

5.4.1. Communication Protocols

Agent communication systems originate from the Knowledge Sharing Effort of the early 1990s, establishing foundational principles for autonomous entity coordination through standardized languages addressing interoperability challenges [369, 93]. KQML emerged as the pioneering Agent Communication Language, introducing multi-layered architecture separating content, message, and communication layers while employing speech act theory [369, 82, 657, 280]. FIPA ACL enhanced this foundation through semantic frameworks based on modal logic, feasibility preconditions, and rational effects [1146, 369, 82].

Interoperability requirements necessitate semantic-level communication capabilities enabling cross-platform agent understanding without extensive pre-communication setup, addressing increasing heterogeneity through ontology-based protocol formalization and Semantic Web technologies, while incorporating security mechanisms against communication vulnerabilities [480, 66, 443, 481, 786, 1055].

Contemporary Protocol Ecosystem Contemporary standardized protocols address fragmentation challenges hindering LLM agent collaboration [1235, 1128, 408]. MCP functions as "USB-C for AI," standardizing agent-environment interactions through JSON-RPC client-server interfaces, enabling hundreds of servers across diverse domains while introducing security vulnerabilities [926, 246, 616, 266, 15, 257, 922, 1094, 370, 1185, 297, 1008, 713, 269].

A2A standardizes peer-to-peer communication through capability-based Agent Cards enabling task delegation and secure collaboration via JSON-based lifecycle models [616, 246, 926]. ACP provides general-purpose RESTful HTTP communication supporting multipart messages and synchronous/asynchronous interactions with discovery, delegation, and orchestration features [277, 246].

ANP extends interoperability to open internet through W3C decentralized identifiers and JSON-LD graphs, with emerging protocols AGNTCY and Agora diversifying standardization ecosystems [246, 679, 1128].

5.4. 多智能体系统

多智能体系统代表了协作智能的顶峰，使多个自主智能体能够协调和通信，以解决超出单个智能体能力的复杂问题。此实现重点在于复杂的通信协议、编排机制和协调策略，这些策略使跨不同智能体架构的协作无缝进行。

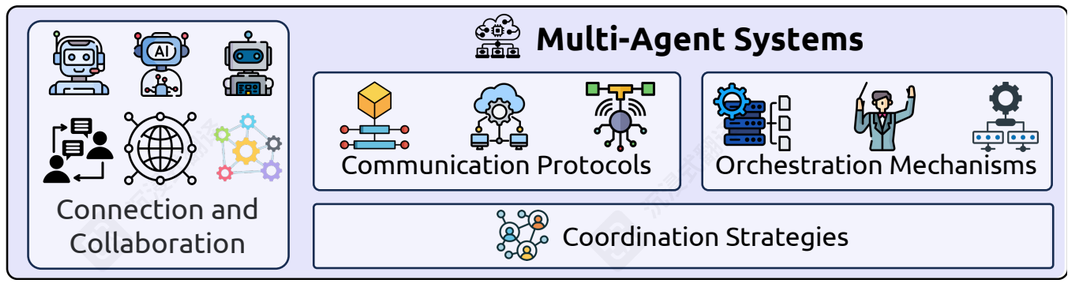

图7：多智能体系统框架：协作人工智能智能体系统的通信协议、编排机制和协调策略概述。

5.4.1. 通信协议

智能体通信系统源于20世纪90年代初的知识共享努力，通过标准化的语言解决互操作性挑战，为自主实体协调建立了基础原则 [369, 93]。KQML作为开创性的智能体通信语言出现，引入了分层架构，将内容、消息和通信层分离，同时采用言语行为理论 [369, 82, 657, 280]。FIPA ACL通过基于模态逻辑、可行性前提和理性效应的语义框架增强了这一基础 [1146, 369, 82]。

互操作性需求需要语义级通信能力，以实现跨平台代理的无需大量预通信设置的理解，通过基于本体的协议形式化和语义网技术解决日益增长的异构性，同时结合安全机制以防范通信漏洞 [480, 66, 443, 481, 786, 1055]。

当代协议生态系统 当代标准化协议解决了阻碍LLM代理协作的碎片化挑战 [1235, 1128, 408]。MCP充当"AI的USB-C"，通过JSON-RPC客户端-服务器接口标准化代理-环境交互，使数百个服务器能够在不同领域运行，同时引入安全漏洞 [926, 246, 616, 266, 15, 257, 922, 1094,370, 1185, 297, 1008, 713, 269]。

A2A通过基于能力的代理卡标准化点对点通信，实现任务委托和基于JSON的生命周期模型的安全协作 [616, 246, 926]。ACP提供通用RESTful HTTP通信，支持多部分消息和同步/异步交互，具有发现、委托和编排功能 [277, 246]。

ANP 通过 W3C 去中心化标识符和 JSON-LD 图扩展了互操作性至开放互联网，新兴协议 AGNTCY 和 Agora 正在多样化标准化生态系统 [246, 679, 1128]。

Progressive layering strategy: MCP provides tool access, ACP enables message exchange, A2A supports peer interaction, ANP extends network interoperability [1007, 926].

LLM-Enhanced Communication Frameworks LLMs transform agent communication through sophisticated natural language processing enabling unprecedented context sensitivity across academic and industrial applications spanning social science, natural science, and engineering domains [486, 684, 498, 1091, 1170, 1127, 896, 1052, 871]. Enhanced systems demonstrate cognitive synergy through specialized knowledge bases, planning, memory, and introspection capabilities, supporting cooperative, debate-oriented, and competitive communication paradigms [486, 356].

Communication structures encompass layered hierarchical organization, decentralized peer-to-peer networks, centralized coordination, and shared message pool architectures, complemented by sequential exchanges, universal language interfaces, and message-passing strategies [356, 1240, 1210, 167, 396, 485, 537, 659, 793, 941].

Framework implementations support comprehensive ecosystems: AutoGen enables dynamic response generation, MetaGPT provides shared message pools, CAMEL offers integrated orchestration, CrewAI facilitates adaptation, with reinforcement learning integration enhancing reward redesign, action selection, and policy interpretation [184, 38, 119, 996, 224, 865, 927, 950, 1264]. Human-agent communication introduces complex interaction landscapes through flexible participation and cognitive diversity, with agents inferring communicator properties and mirroring human communicative intentions [1399, 34, 669].

5.4.2. Orchestration Mechanisms

Orchestration mechanisms constitute the critical coordination infrastructure for multi-agent systems, managing agent selection, context distribution, and interaction flow control [894], enabling effective cooperation among human and non-human actors through user input processing, contextual distribution, and optimal agent selection based on capability assessment and response evaluation [53], while managing message flow, ensuring task progression, and addressing task deviations [171]. Advanced orchestration frameworks incorporate intent recognition, contextual memory maintenance, and task dispatching components for intelligent coordination across domain-specific agents, with the Swarm Agent framework utilizing real-time outputs to direct tool invocations while addressing limitations in static tool registries and bespoke communication frameworks [808, 263, 246].

Contemporary orchestration strategies exhibit distinct operational paradigms: a priori orchestration determines agent selection through pre-execution analysis of user input and agent capabilities, while posterior orchestration distributes inputs to multiple agents simultaneously, utilizing confidence metrics and response quality assessment as demonstrated by the 3S orchestrator framework [893]; function-based orchestration emphasizes agent selection from available pools, contextual information management, and conversation flow control [54]; component-based orchestration employs dynamic planning processes where orchestrators arrange components in logical sequences based on user instructions, utilizing LLMs as component orchestration tools to generate workflows with embedded orchestration logic [675].

Emergent orchestration paradigms include puppeteer-style orchestration featuring centralized orchestrators that dynamically direct agents in response to evolving task states through reinforcement learning-based adaptive sequencing and prioritization, and serialized orchestration addressing collaboration topology complexity by unfolding collaboration graphs into reasoning sequences guided by topological traversal, enabling orchestrators to select single agents at each step based on global system state and task specifications [194].

渐进式分层策略：MCP提供工具访问，ACP实现消息交换，A2A支持对等交互，ANP扩展网络互操作性 [1007, 926]。

LLM增强型通信框架 LLM通过复杂的自然语言处理转换代理通信，实现前所未有的上下文敏感性，涵盖社会科学、自然科学和工程领域的学术和工业应用 [486, 684, 498, 1091, 1170,1127, 896, 1052, 871]。增强系统通过专业知识库、规划、记忆和内省能力展示认知协同，支持合作、辩论导向和竞争性通信范式 [486, 356]。

通信结构包括分层层次组织、去中心化对等网络、集中式协调和共享消息池架构，辅以顺序交换、通用语言接口和消息传递策略 [356, 1240, 1210, 167, 396, 485,537, 659, 793, 941]。

框架实现支持全面的生态系统：AutoGen 支持动态响应生成，MetaGPT 提供共享消息池，CAMEl 提供集成编排，CrewAI 促进适应，通过强化学习集成增强奖励重新设计、动作选择和政策解释 [184, 38, 119, 996, 224, 865, 927, 950, 1264]。人机通信通过灵活的参与和认知多样性引入复杂的交互场景，代理推断通信者属性并镜像人类的交际意图 [1399, 34, 669]。

5.4.2. 编排机制

编排机制构成多智能体系统的关键协调基础设施，管理代理选择、上下文分发和交互流控制 [894], ，通过用户输入处理、上下文分发以及基于能力评估和响应评估的最优代理选择，实现人类和非人类参与者之间的有效合作 [53], ，同时管理消息流，确保任务进展，并处理任务偏差 [171]。高级编排框架包含意图识别、上下文记忆维护和任务调度组件，用于跨特定领域代理的智能协调，Swarm Agent 框架利用实时输出来指导工具调用，同时解决静态工具注册表和定制通信框架的限制 [808, 263, 246]。

当代编排策略表现出不同的操作范式：先验编排通过预执行分析用户输入和代理能力来确定代理选择，而后验编排同时将输入分配给多个代理，利用置信度指标和响应质量评估，如3S编排框架 [893]所示；基于功能的编排强调从可用池中选择代理、上下文信息管理和对话流程控制 [54]；基于组件的编排采用动态规划过程，编排器根据用户指令将组件按逻辑顺序排列，利用LLMs作为组件编排工具生成嵌入编排逻辑的工作流 [675]。

涌现的编排范式包括类似提线木偶式的编排，其特点是集中式编排器通过基于强化学习的自适应序列化和优先级排序动态指导代理，以响应不断变化的任务状态，以及串行编排通过将协作图展开为拓扑遍历指导的推理序列来解决协作拓扑复杂性，使编排器能够根据全局系统状态和任务规范在每个步骤选择单个代理 [194]。

Context Management and Environmental Adaptation Context serves as the foundational element guiding agent actions and interactions within orchestrated systems, supporting operational mode diversity while maintaining application individuality and task execution sequencing through global state maintenance that enables orchestration systems to track task execution progress across distributed nodes, providing agents with contextual awareness necessary for effective subtask performance within broader workflow contexts [26]. Session-based context refinement defines collaborative scope boundaries, facilitating event-driven orchestration where agents can enter and exit dynamically, create output streams, and contribute to shared session streams, with configurable sessions enabling agent inclusion based on user input or autonomous decision-making to create adaptable systems responsive to changing task requirements [513].

Well-designed interaction structures and task orchestration mechanisms underscore context's critical role in scalable multi-agent collaboration. Systems adapt communication patterns and agent roles to contextual requirements, supporting dynamic collaboration tailored to specific task demands through complex task decomposition and suitable agent assignment for subtask execution [1128]. This contextual adaptation encompasses both organizational and operational dimensions, enabling systems to maintain coherence while accommodating environmental variability and evolving user requirements.

5.4.3. Coordination Strategies

Multi-agent orchestration encounters significant challenges in maintaining transactional integrity across complex workflows, with contemporary frameworks including LangGraph, AutoGen, and CAMEL demonstrating insufficient transaction support: LangGraph provides basic state management while lacking atomicity guarantees and systematic compensation mechanisms, AutoGen prioritizes flexible agent interactions without adequate compensatory action management potentially resulting in inconsistent system states following partial failures, and validation limitations emerge as many frameworks rely exclusively on large language models' inherent self-validation capabilities without implementing independent validation procedures, exposing systems to reasoning errors, hallucinations, and inter-agent inconsistencies [128].

Context handling failures compound these challenges as agents struggle with long-term context maintenance encompassing both episodic and semantic information [210, 1113], while central orchestrator topologies introduce non-deterministic, runtime-dependent execution paths that enhance adaptability while complicating anomaly detection, requiring dynamic graph reconstruction rather than simple path matching [390], and environmental misconfigurations and LLM hallucinations can distract agentic systems, with poor recovery leading to goal deviation that becomes amplified in multi-agent setups with distributed subtasks [210, 1091].

Inter-agent dependency opacity presents additional concerns as agents may operate on inconsistent assumptions or conflicting data without explicit constraints or validation layers, necessitating anomaly detection incorporating reasoning over orchestration intent and planning coherence [390], while addressing these challenges requires comprehensive solutions such as the SagaLLM framework providing transaction support, independent validation procedures, and robust context preservation mechanisms [128], and approaches like CodeAct integrating Python interpreters with LLM agents to enable code action execution and dynamic revision capabilities through multi-turn interactions [1113].

Applications and Performance Implications Agent and context orchestration demonstrates practical utility across diverse application domains: healthcare applications employ context-switching mechanisms within specialized agent-based architectures performing information retrieval, question answering, and decision support, utilizing supervisory agents to interpret input features and assign subtasks to specialized

上下文管理与环境适应 上下文作为指导智能体在编排系统中的行为和交互的基础元素，通过全局状态维护支持操作模式多样性，同时保持应用程序的个体性和任务执行顺序，使编排系统能够跨分布式节点跟踪任务执行进度，为智能体提供必要的上下文感知能力，以便在更广泛的流程上下文中有效执行子任务[26]。基于会话的上下文细化定义了协作范围边界，促进了事件驱动的编排，其中智能体可以动态进入和退出，创建输出流，并为共享会话流做出贡献，可配置的会话允许根据用户输入或自主决策来包含智能体，以创建能够适应不断变化的任务需求的灵活系统 [513]。

精心设计的交互结构和任务编排机制突出了上下文在可扩展的多智能体协作中的关键作用。系统根据上下文需求调整通信模式和智能体角色，通过复杂的任务分解和合适的智能体分配来执行子任务，支持针对特定任务需求的动态协作 [1128]。这种上下文适应涵盖了组织和操作两个维度，使系统能够在适应环境变化和不断变化的用户需求的同时保持一致性。

5.4.3. 协调策略

多智能体编排在跨复杂工作流维护事务完整性方面面临重大挑战，当代框架包括 LangGraph、AutoGen 和 CAMEL 展示了不足的事务支持：LangGraph 提供基本状态管理但缺乏原子性保证和系统性的补偿机制，AutoGen 优先考虑灵活的智能体交互而缺乏充分的补偿动作管理，可能导致部分故障后系统状态不一致，并且验证限制凸显，因为许多框架仅依赖大型语言模型的固有自验证能力而未实现独立的验证程序，使系统暴露于推理错误、幻觉和智能体间不一致 [128]。

上下文处理失败加剧了这些挑战，因为智能体在维护涵盖情景和语义信息的长期上下文方面存在困难 [210, 1113], ，而中央编排器拓扑引入了非确定性、运行时依赖的执行路径，增强了适应性但使异常检测复杂化，需要动态图重构而非简单的路径匹配 [390], ，以及环境配置错误和 LLM 幻觉会分散智能体系统，恢复不佳会导致目标偏离，在具有分布式子任务的多智能体设置中这种偏离会被放大 [210, 1091]。

智能体间依赖的透明度带来了额外的挑战，因为智能体可能在缺乏显式约束或验证层的情况下基于不一致的假设或冲突数据运行，这需要结合编排意图和规划一致性进行推理的异常检测 [390], ，而解决这些挑战需要综合解决方案，例如提供事务支持、独立验证程序和强大的上下文保留机制的SagaLLM框架 [128], ，以及将Python解释器与LLM智能体集成的CodeAct等方法，以通过多轮交互实现代码执行和动态修订能力 [1113]。

应用与性能影响 智能体和上下文编排在多种应用领域展示了实用价值：医疗保健应用在专门的基于智能体的架构中采用上下文切换机制，执行信息检索、问答和决策支持，利用监督智能体解释输入特征并将子任务分配给专业智能体

agents based on clinical query type, user background, and data modality requirements [613, 754, 1051]; network management applications leverage context-aware orchestration to address complexity challenges by equipping Points of Access with agents dedicated to unique contexts, enabling efficient network dynamics management through context-specific action sets including available service instances and network paths [958].

Business process management and simulation represent significant application areas through platforms like AgentSimulator, enabling process behavior discovery and simulation in orchestrated and autonomous settings where orchestrated behavior follows global control-flow patterns with activity selection dependent on previous activities and agent assignment based on capabilities and availability, while autonomous behavior operates through local control-flow and handover patterns acknowledging agent autonomy in collaborative work [543].

Performance implications indicate that well-designed orchestration improves system effectiveness by leveraging distinct agent capabilities, with research demonstrating that human users frequently struggle with effective agent selection from available sets while automated orchestration enhances overall performance [72], motivating frameworks that learn agent capabilities online and orchestrate multiple agents under real-world constraints including cost, capability requirements, and operational limitations, with autonomy levels varying across implementations where some systems exhibit pronounced autonomy within designated phases, demonstrating adaptability in action management corresponding to task specificity and reaching Level 2 autonomy through contextual resource utilization [460].

6. Evaluation

The evaluation of context-engineered systems presents unprecedented challenges that transcend traditional language model assessment paradigms. These systems exhibit complex, multi-component architectures with dynamic, context-dependent behaviors requiring comprehensive evaluation frameworks that assess component-level diagnostics, task-based performance, and overall system robustness [835, 1132].

The heterogeneous nature of context engineering components—spanning retrieval mechanisms, memory systems, reasoning chains, and multi-agent coordination—demands evaluation methodologies that can capture both individual component effectiveness and emergent system-level behaviors [310, 931].

6.1. Evaluation Frameworks and Methodologies

This subsection presents comprehensive approaches for evaluating both individual components and integrated systems in context engineering.

6.1.1. Component-Level Assessment

Intrinsic evaluation focuses on the performance of individual components in isolation, providing foundational insights into system capabilities and failure modes.

For **prompt engineering** components, evaluation encompasses prompt effectiveness measurement through semantic similarity metrics, response quality assessment, and robustness testing across diverse input variations. Current approaches reveal brittleness and robustness challenges in prompt design, necessitating more sophisticated evaluation frameworks that can assess contextual calibration and adaptive prompt optimization [1132, 663].

基于临床查询类型、用户背景和数据模态要求的代理 [613, 754, 1051]；网络管理应用程序利用上下文感知编排来应对复杂性挑战，通过为接入点配备专门针对独特上下文的代理，并通过上下文特定的动作集（包括可用服务实例和网络路径）实现高效的网络动态管理[958]。

业务流程管理和模拟代表重要的应用领域，通过AgentSimulator等平台，在编排和自主环境中实现过程行为发现和模拟，其中编排行为遵循全局控制流模式，活动选择依赖于先前活动，代理分配基于能力和可用性，而自主行为通过本地控制流和交接模式运行，承认代理在协作工作中的自主性 [543]。

性能影响表明，精心设计的编排通过利用不同的代理能力提高了系统有效性，研究表明人类用户经常难以从可用集中选择有效的代理，而自动化编排提高了整体性能[72], ，这促使框架在线学习代理能力并在现实世界约束下编排多个代理，包括成本、能力要求和运营限制，自主程度在不同的实现中有所不同，其中一些系统在指定阶段表现出明显的自主性，在动作管理中表现出适应性，对应于任务的特定性，并通过上下文资源利用达到Level 2自主性 [460]。

6. 评估

对上下文工程系统的评估面临着前所未有的挑战，这些挑战超越了传统的语言模型评估范式。这些系统表现出复杂的多组件架构和动态的上下文相关行为，需要全面的评估框架，该框架可以评估组件级诊断、基于任务的性能和整体系统鲁棒性 [835, 1132]。

上下文工程组件的异构性——涵盖检索机制、记忆系统、推理链和多智能体协调——要求评估方法能够捕捉单个组件的有效性和涌现的系统级行为 [310, 931]。

6.1. 评估框架和方法

本小节介绍了评估上下文工程中单个组件和集成系统的全面方法。

6.1.1. 组件级评估

内在评估关注单个组件在隔离状态下的性能，为系统能力和故障模式提供基础性见解。

对于 **提示工程** 组件，评估涵盖通过语义相似度指标测量提示效果、响应质量评估以及在多样化输入变化中的鲁棒性测试。当前方法揭示了提示设计的脆弱性和鲁棒性挑战，需要更复杂的评估框架，该框架能够评估上下文校准和自适应提示优化 [1132, 663]。

Long context processing evaluation requires specialized metrics addressing information retention, positional bias, and reasoning coherence across extended sequences. The "needle in a haystack" evaluation paradigm tests models' ability to retrieve specific information embedded within long contexts, while multi-document reasoning tasks assess synthesis capabilities across multiple information sources. Position interpolation techniques and ultra-long sequence processing methods face significant computational challenges that limit practical evaluation scenarios [731, 295].

Self-contextualization mechanisms undergo evaluation through meta-learning assessments, adaptation speed measurements, and consistency analysis across multiple iterations. Self-refinement frameworks including Self-Refine, Reflexion, and N-CRITICS demonstrate substantial performance improvements, with GPT-4 achieving approximately 20% improvement through iterative self-refinement processes [735, 956, 789]. Multi-dimensional feedback mechanisms and ensemble-based evaluation approaches provide comprehensive assessment of autonomous evolution capabilities [577, 704].

Structured and relational data integration evaluation examines accuracy in knowledge graph traversal, table comprehension, and database query generation. However, current evaluation frameworks face significant limitations in assessing structural reasoning capabilities, with high-quality structured training data development presenting ongoing challenges. LSTM-based models demonstrate increased errors when sequential and structural information conflict, highlighting the need for more sophisticated benchmarks testing structural understanding [763, 668, 163].

6.1.2. System-Level Integration Assessment

Extrinsic evaluation measures end-to-end performance on downstream tasks, providing holistic assessments of system utility through comprehensive benchmarks spanning question answering, reasoning, and real-world applications.

System-level evaluation must capture emergent behaviors arising from component interactions, including synergistic effects where combined components exceed individual performance and potential interference patterns where component integration degrades overall effectiveness [835, 1132].

Retrieval-Augmented Generation evaluation encompasses both retrieval quality and generation effectiveness through comprehensive metrics addressing precision, recall, relevance, and factual accuracy. Agentic RAG systems introduce additional complexity requiring evaluation of task decomposition accuracy, multi-plan selection effectiveness, and memory-augmented planning capabilities. Self-reflection mechanisms demonstrate iterative improvement through feedback loops, with MemoryBank implementations incorporating Ebbinghaus Forgetting Curve principles for enhanced memory evaluation [438, 162, 1362, 1183, 41].

Memory systems evaluation encounters substantial difficulties stemming from the absence of standardized assessment frameworks and the inherently stateless characteristics of contemporary LLMs. LongMemEval offers 500 carefully curated questions that evaluate fundamental capabilities encompassing information extraction, temporal reasoning, multi-session reasoning, and knowledge updates. Commercial AI assistants exhibit 30% accuracy degradation throughout extended interactions, underscoring significant deficiencies in memory persistence and retrieval effectiveness [1330, 1171, 457, 841, 386]. Dedicated benchmarks such as NarrativeQA, QMSum, QuALITY, and MEMENTO tackle episodic memory evaluation challenges [550, 566].

Tool-integrated reasoning systems require comprehensive evaluation covering the entire interaction trajectory, including tool selection accuracy, parameter extraction precision, execution success rates, and error recovery capabilities. The MCP-RADAR framework provides standardized evaluation employing objective metrics for software engineering and mathematical reasoning domains. Real-world evaluation reveals

长上下文处理 评估需要专门的指标来处理信息保留、位置偏差和跨长序列的推理连贯性。"大海捞针"评估范式测试模型检索长上下文中嵌入的特定信息的能力，而多文档推理任务评估跨多个信息源的合成能力。位置插值技术和超长序列处理方法面临显著的计算挑战，限制了实际评估场景 [731, 295]。

自上下文化 机制通过元学习评估、适应速度测量和多轮次迭代中的一致性分析进行评估。包括 Self-Refine、Reflexion 和 N-CRITICS 在内的自优化框架表现出显著的性能提升，GPT-4 通过迭代自优化过程实现了约 20% 的性能提升 [735, 956, 789]。多维反馈机制和基于集成的方法为自主进化能力提供全面评估 [577, 704]。

结构化和关系数据集成 评估检查知识图谱遍历、表格理解和数据库查询生成的准确性。然而，当前的评估框架在评估结构推理能力方面面临重大限制，高质量结构化训练数据开发持续存在挑战。基于 LSTM 的模型在顺序和结构信息冲突时表现出更高的错误率，突显了需要更复杂的基准来测试结构理解的必要性 [763, 668, 163]。

6.1.2. 系统级集成评估

外部评估衡量下游任务的端到端性能，通过涵盖问答、推理和实际应用的全面基准测试，提供系统实用性的整体评估。

系统级评估必须捕捉组件交互产生的涌现行为，包括组合组件超过个体性能的协同效应以及组件集成导致整体效果退化的潜在干扰模式 [835, 1132]。

检索增强生成 评估通过综合指标涵盖检索质量和生成效果，包括精确度、召回率、相关性和事实准确性。代理式 RAG 系统引入了额外的复杂性，需要评估任务分解的准确性、多计划选择的有效性和记忆增强规划能力。自我反思机制通过反馈循环展示迭代改进，MemoryBank 实现结合了艾宾浩斯遗忘曲线原理以增强记忆评估 [438, 162, 1362, 1183, 41]。

记忆系统 评估由于缺乏标准化的评估框架和当代 LLMs 的固有状态lessness 特性而面临重大困难。LongMemEval 提供了 500 个精心策划的问题，评估涵盖信息提取、时间推理、多会话推理和知识更新的基本能力。商业 AI 助手在整个长时间交互中准确率下降 30%，突显了记忆持久性和检索效果的显著缺陷 [1330, 1171, 457, 841, 386]。专用基准测试如 NarrativeQA、QMSum、QuALITY 和 MEMENTO 解决了情景记忆评估挑战 [550, 566]。

工具集成推理系统 需要全面评估整个交互轨迹，包括工具选择准确性、参数提取精度、执行成功率以及错误恢复能力。MCP-RADAR框架提供了针对软件工程和数学推理领域的标准化评估，采用客观指标进行。实际评估显示

significant performance gaps, with GPT-4 completing less than 50% of tasks in the GTA benchmark, compared to human performance of 92% [310, 1090, 126, 931]. Advanced benchmarks including BFCL (2,000 testing cases), T-Eval (553 tool-use cases), API-Bank (73 APIs, 314 dialogues), and ToolHop (995 queries, 3,912 tools) address multi-turn interactions and nested tool calling scenarios [259, 359, 373, 1255, 157, 829].

Multi-agent systems evaluation captures communication effectiveness, coordination efficiency, and collective outcome quality through specialized metrics addressing protocol adherence, task decomposition accuracy, and emergent collaborative behaviors. Contemporary orchestration frameworks including LangGraph, AutoGen, and CAMEL demonstrate insufficient transaction support, with validation limitations emerging as systems rely exclusively on LLM self-validation capabilities without independent validation procedures. Context handling failures compound challenges as agents struggle with long-term context maintenance encompassing both episodic and semantic information [128, 390, 893].

6.2. Benchmark Datasets and Evaluation Paradigms

This subsection reviews specialized benchmarks and evaluation paradigms designed for assessing context engineering system performance.

6.2.1. Foundational Component Benchmarks

Long context processing evaluation employs specialized benchmark suites designed to test information retention, reasoning, and synthesis across extended sequences. Current benchmarks face significant computational complexity challenges, with $O(n^2)$ scaling limitations in attention mechanisms creating substantial memory constraints for ultra-long sequences. Position interpolation and extension techniques require sophisticated evaluation frameworks that can assess both computational efficiency and reasoning quality across varying sequence lengths [731, 295, 1227].

Advanced architectures including LongMamba and specialized position encoding methods demonstrate promising directions for long context processing, though evaluation reveals persistent challenges in maintaining coherence across extended sequences. The development of sliding attention mechanisms and memory-efficient implementations requires comprehensive benchmarks that can assess both computational tractability and task performance [1258, 347].

Structured and relational data integration benchmarks encompass diverse knowledge representation formats and reasoning patterns. However, current evaluation frameworks face limitations in assessing structural reasoning capabilities, with the development of high-quality structured training data presenting ongoing challenges. Evaluation must address the fundamental tension between sequential and structural information processing, particularly in scenarios where these information types conflict [763, 668, 163].

6.2.2. System Implementation Benchmarks

Retrieval-Augmented Generation evaluation leverages comprehensive benchmark suites addressing diverse retrieval and generation challenges. Modular RAG architectures demonstrate enhanced flexibility through specialized modules for retrieval, augmentation, and generation, enabling fine-grained evaluation of individual components and their interactions. Graph-enhanced RAG systems incorporating GraphRAG and LightRAG demonstrate improved performance in complex reasoning scenarios, though evaluation frameworks must address the additional complexity of graph traversal and multi-hop reasoning assessment [312, 965, 360].

Agentic RAG systems introduce sophisticated planning and reflection mechanisms requiring evaluation

显著的性能差距，与人类92%的性能相比，GPT-4在GTA基准测试中完成的任务不到50% [310, 1090, 126, 931]。高级基准测试包括BFCL（2,000个测试用例）、T-Eval（553个工具使用用例）、API-Bank（73个API，314个对话）和ToolHop（995个查询，3,912个工具）解决了多轮交互和嵌套工具调用场景 [259, 359, 373, 1255, 157, 829]。

多智能体系统评估通过专门指标捕捉通信有效性、协调效率和集体结果质量，这些指标针对协议遵守、任务分解准确性和涌现式协作行为。当代编排框架包括LangGraph、AutoGen和CAMEL在事务支持方面不足，随着系统完全依赖LLM自我验证能力而不进行独立验证程序，验证限制问题凸显。上下文处理失败加剧了挑战，因为智能体在维护包含情景和语义信息的长期上下文方面存在困难 [128, 390, 893]。

6.2. 基准数据集和评估范式

本小节回顾了为评估上下文工程系统性能而设计的专用基准和评估范式。

6.2.1. 基础组件基准

长上下文处理评估采用专门设计的基准套件，用于测试跨扩展序列的信息保留、推理和综合能力。当前基准面临显著的计算复杂度挑战，注意力机制中的 $O(n_2)$ 缩放限制为超长序列创造了巨大的内存约束。位置插值和扩展技术需要能够评估不同序列长度下计算效率和推理质量的复杂评估框架 [731, 295, 1227]。

包括 LongMamba 在内的先进架构和专门的位置编码方法为长上下文处理展示了有前景的方向，尽管评估揭示了在跨扩展序列保持连贯性方面持续存在的挑战。滑动注意力机制和内存高效实现的开发需要能够评估计算可行性和任务性能的全面基准 [1258, 347]。

结构化和关系数据集成基准涵盖了多样的知识表示格式和推理模式。然而，当前的评估框架在评估结构化推理能力方面存在局限性，高质量结构化训练数据的开发持续存在挑战。评估必须解决序列化信息处理和结构化信息处理之间的基本张力，特别是在这些信息类型冲突的场景中 [763, 668, 163]。

6.2.2. 系统实现基准测试

检索增强生成评估利用全面的基准测试套件来解决多样化的检索和生成挑战。模块化 RAG 架构通过专门的检索、增强和生成模块展示了增强的灵活性，能够对单个组件及其交互进行细粒度评估。图增强 RAG 系统结合 GraphRAG 和 LightRAG 在复杂推理场景中表现出改进的性能，但评估框架必须解决图遍历和多跳推理评估的额外复杂性 [312, 965, 360]。

代理式 RAG 系统引入了复杂的规划和反思机制，需要评估

of task decomposition accuracy, multi-plan selection effectiveness, and iterative refinement capabilities. Real-time and streaming RAG applications present unique evaluation challenges in assessing both latency and accuracy under dynamic information conditions [438, 162, 1183].

Tool-integrated reasoning system evaluation employs comprehensive benchmarks spanning diverse tool usage scenarios and complexity levels. The Berkeley Function Calling Leaderboard (BFCL) provides 2,000 testing cases with step-by-step and end-to-end assessments measuring call accuracy, pass rates, and win rates across increasingly complex scenarios. T-Eval contributes 553 tool-use cases testing multi-turn interactions and nested tool calling capabilities [259, 1380, 829]. Advanced benchmarks including StableToolBench address API instability challenges, while NesTools evaluates nested tool scenarios and ToolHop assesses multi-hop tool usage across 995 queries and 3,912 tools [359, 373, 1255].

Web agent evaluation frameworks including WebArena and Mind2Web provide comprehensive assessment across thousands of tasks spanning 137 websites, revealing significant performance gaps in current LLM capabilities for complex web interactions. VideoWebArena extends evaluation to multimodal agents, while Deep Research Bench and DeepShop address specialized evaluation for research and shopping agents respectively [1368, 202, 87, 476].

Multi-agent system evaluation employs specialized frameworks addressing coordination, communication, and collective intelligence. However, current frameworks face significant challenges in transactional integrity across complex workflows, with many systems lacking adequate compensation mechanisms for partial failures. Orchestration evaluation must address context management, coordination strategy effectiveness, and the ability to maintain system coherence under varying operational conditions [128, 893].

Release Date	Open Source	Method / Model	Success Rate (%)	Source
2025-02	×	IBM CUGA	61.7	[747]
2025-01	×	OpenAI Operator	58.1	[807]
2024-08	×	Jace.AI	57.1	[470]
2024-12	×	ScribeAgent + GPT-4o	53.0	[942]
2025-01	✓	AgentSymbiotic	52.1	[1314]
2025-01	✓	Learn-by-Interact	48.0	[990]
2024-10	✓	AgentOccam-Judge	45.7	[1222]
2024-08	×	WebPilot	37.2	[1322]
2024-10	✓	GUI-API Hybrid Agent	35.8	[980]
2024-09	✓	Agent Workflow Memory	35.5	[1135]
2024-04	✓	SteP	33.5	[971]
2025-06	✓	TTI	26.1	[943]
2024-04	✓	BrowserGym + GPT-4	23.5	[234]

Table 8: WebArena [1368] Leaderboard: Top performing models with their success rates and availability status.

6.3. Evaluation Challenges and Emerging Paradigms

This subsection identifies current limitations in evaluation methodologies and explores emerging approaches for more effective assessment.

任务分解准确性、多计划选择有效性以及迭代优化能力。实时和流式RAG应用在动态信息条件下评估延迟和准确性的独特挑战 [438, 162, 1183]。

工具集成推理系统评估采用涵盖不同工具使用场景和复杂程度的综合基准。伯克利函数调用排行榜（BFCL）提供2,000个测试用例，进行逐步和端到端的评估，测量在日益复杂的场景中调用准确性、通过率和胜率。T-Eval贡献553个工具使用用例，测试多轮交互和嵌套工具调用能力 [259, 1380, 829]。高级基准包括StableToolBench，解决API不稳定性挑战，而NesTools评估嵌套工具场景，ToolHop评估跨995个查询和3,912个工具的多跳工具使用 [359, 373, 1255]。

包括WebArena和Mind2Web的网页代理评估框架在跨越137个网站的数千个任务中提供综合评估，揭示当前LLM在复杂网页交互方面的显著性能差距。VideoWebArena将评估扩展到多模态代理，而Deep Research Bench和DeepShop分别针对研究和购物代理进行专业评估 [1368, 202, 87, 476]。

多智能体系统评估采用专门框架处理协调、通信和集体智能问题。然而，当前框架在复杂工作流中的事务完整性方面面临重大挑战，许多系统缺乏对部分故障的充分补偿机制。编排评估必须解决上下文管理、协调策略有效性以及在不同运行条件下维持系统一致性的能力 [128, 893]。

发布日期	开源方法/模型		成功率 (%)	来源
2025-02	×	IBM CUGA	61.7	[747]
2025-01	×	OpenAI Operator	58.1	[807]
2024-08	×	Jace.AI	57.1	[470]
2024-12	×	ScribeAgent + GPT-4o	53.0	[942]
2025-01	✓	AgentSymbiotic	52.1	[1314]
2025-01	✓	Learn-by-Interact	48.0	[990]
2024-10	✓	AgentOccam-Judge	45.7	[1222]
2024-08	×	WebPilot	37.2	[1322]
2024-10	✓	GUI-API 混合代理	35.8	[980]
2024-09	✓	代理工作流内存	35.5	[1135]
2024-04	✓	SteP	33.5	[971]
2025-06	✓	TTI	26.1	[943]
2024-04	✓	BrowserGym + GPT-4	23.5	[234]

表 8：WebArena [1368] 排行榜：表现最佳的模型及其成功率和使用状态。

6.3. 评估挑战与新兴范式

本小节确定了评估方法中的当前局限性，并探讨了更有效评估的新兴方法。

6.3.1. Methodological Limitations and Biases

Traditional evaluation metrics prove fundamentally inadequate for capturing the nuanced, dynamic behaviors exhibited by context-engineered systems. Static metrics like BLEU, ROUGE, and perplexity, originally designed for simpler text generation tasks, fail to assess complex reasoning chains, multi-step interactions, and emergent system behaviors. The inherent complexity and interdependencies of multi-component systems create attribution challenges where isolating failures and identifying root causes becomes computationally and methodologically intractable. Future metrics must evolve to capture not just task success, but the quality and robustness of the underlying reasoning process, especially in scenarios requiring compositional generalization and creative problem-solving [835, 1132].

Memory system evaluation faces particular challenges due to the lack of standardized benchmarks and the stateless nature of current LLMs. Automated memory testing frameworks must address the isolation problem where different memory testing stages cannot be effectively separated, leading to unreliable assessment results. Commercial AI assistants demonstrate significant performance degradation during sustained interactions, with accuracy drops of up to 30% highlighting critical gaps in current evaluation methodologies and pointing to the need for longitudinal evaluation frameworks that track memory fidelity over time [1330, 1171, 457].

Tool-integrated reasoning system evaluation reveals substantial performance gaps between current systems and human-level capabilities. The GAIA benchmark demonstrates that while humans achieve 92% accuracy on general assistant tasks, advanced models like GPT-4 achieve only 15% accuracy, indicating fundamental limitations in current evaluation frameworks and system capabilities [772, 1090, 126]. Evaluation frameworks must address the complexity of multi-tool coordination, error recovery, and adaptive tool selection across diverse operational contexts [310, 931].

6.3.2. Emerging Evaluation Paradigms

Self-refinement evaluation paradigms leverage iterative improvement mechanisms to assess system capabilities across multiple refinement cycles. Frameworks including Self-Refine, Reflexion, and N-CRITICS demonstrate substantial performance improvements through multi-dimensional feedback and ensemble-based evaluation approaches. GPT-4 achieves approximately 20% improvement through self-refinement processes, highlighting the importance of evaluating systems across multiple iteration cycles rather than single-shot assessments. However, a key future challenge lies in evaluating the meta-learning capability itself—not just whether the system improves, but how efficiently and robustly it learns to refine its strategies over time [735, 956, 789, 577].

Multi-aspect feedback evaluation incorporates diverse feedback dimensions including correctness, relevance, clarity, and robustness, providing comprehensive assessment of system outputs. Self-rewarding mechanisms enable autonomous evolution and meta-learning assessment, allowing systems to develop increasingly sophisticated evaluation criteria through iterative refinement [704].

Criticism-guided evaluation employs specialized critic models to provide detailed feedback on system outputs, enabling fine-grained assessment of reasoning quality, factual accuracy, and logical consistency. These approaches address the limitations of traditional metrics by providing contextual, content-aware evaluation that can adapt to diverse task requirements and output formats [789, 577].

Orchestration evaluation frameworks address the unique challenges of multi-agent coordination by incorporating transactional integrity assessment, context management evaluation, and coordination strategy effectiveness measurement. Advanced frameworks including SagaLLM provide transaction support and

6.3.1. 方法论局限性和偏差

传统的评估指标无法充分捕捉上下文工程系统所展现的细致、动态的行为。静态指标如BLEU、ROUGE和困惑度，最初是为更简单的文本生成任务设计的，无法评估复杂的推理链、多步交互和涌现的系统行为。多组件系统的固有复杂性和相互依赖性导致了归因挑战，将故障隔离并识别根本原因变得在计算和方法上难以处理。未来的指标必须进化，不仅要捕捉任务成功，还要评估底层推理过程的质量和鲁棒性，特别是在需要组合泛化和创造性解决问题的场景中 [835, 1132]。

由于缺乏标准化的基准和当前LLM的无状态特性，内存系统评估面临特殊挑战。自动内存测试框架必须解决隔离问题，即不同的内存测试阶段无法有效分离，导致评估结果不可靠。商业AI助手在持续交互期间性能显著下降，准确率下降高达30%，这突出了当前评估方法论的严重缺陷，并指出了需要纵向评估框架来跟踪内存保真度随时间变化的必要性 [1330, 1171, 457]。

工具集成推理系统评估揭示了当前系统与人类水平能力之间的巨大性能差距。GAIA基准表明，虽然人类在通用助手任务上达到92%的准确率，但像GPT-4这样的高级模型仅达到15%的准确率，这表明当前评估框架和系统能力存在根本性局限性 [772, 1090, 126]。评估框架必须解决多工具协调的复杂性、错误恢复以及在多样化操作环境中的自适应工具选择 [310, 931]。

6.3.2. 新兴评估范式

自我完善评估范式利用迭代改进机制来评估系统在多个完善周期中的能力。包括Self-Refine、Reflexion和N-CRITICS在内的框架通过多维反馈和基于集成的方法展示了显著的性能提升。GPT-4通过自我完善过程实现了约20%的改进，突出了在多个迭代周期中评估系统而非单次评估的重要性。然而，一个关键的未来挑战在于评估元学习能力本身——不仅系统是否改进，而且它如何高效且稳健地学习随着时间的推移完善其策略 [735, 956, 789, 577]。

多方面反馈评估包含多种反馈维度，包括正确性、相关性、清晰度和鲁棒性，为系统输出提供全面评估。自我奖励机制实现自主进化和元学习评估，允许系统通过迭代优化开发日益复杂的评估标准 [704]。

批评引导评估采用专门的批评模型为系统输出提供详细反馈，实现推理质量、事实准确性和逻辑一致性的细粒度评估。这些方法通过提供上下文感知、内容感知的评估来弥补传统指标的局限性，能够适应多样化的任务要求和输出格式 [789, 577]。

编排评估框架通过包含事务完整性评估、上下文管理评估和协调策略有效性测量来解决多智能体协调的独特挑战。包括 SagaLLM 的高级框架提供事务支持并

independent validation procedures to address the limitations of systems that rely exclusively on LLM self-validation capabilities [128, 390].

6.3.3. Safety and Robustness Assessment

Safety-oriented evaluation incorporates comprehensive robustness testing, adversarial attack resistance, and alignment assessment to ensure responsible development of context-engineered systems. Particular attention must be paid to the evaluation of agentic systems that can operate autonomously across extended periods, as these systems present unique safety challenges that traditional evaluation frameworks cannot adequately address [965, 360].

Robustness evaluation must assess system performance under distribution shifts, input perturbations, and adversarial conditions through comprehensive stress testing protocols. Multi-agent systems face additional challenges in coordination failure scenarios, where partial system failures can cascade through the entire agent network. Evaluation frameworks must address graceful degradation strategies, error recovery protocols, and the ability to maintain system functionality under adverse conditions. Beyond predefined failure modes, future evaluation must grapple with assessing resilience to "unknown unknowns"—emergent and unpredictable failure cascades in highly complex, autonomous multi-agent systems [128, 390].

Alignment evaluation measures system adherence to intended behaviors, value consistency, and beneficial outcome optimization through specialized assessment frameworks. Context engineering systems present unique alignment challenges due to their dynamic adaptation capabilities and complex interaction patterns across multiple components. Long-term evaluation must assess whether systems maintain beneficial behaviors as they adapt and evolve through extended operational periods [893].

Looking ahead, the evaluation of context-engineered systems requires a paradigm shift from static benchmarks to dynamic, holistic assessments. Future frameworks must move beyond measuring task success to evaluating compositional generalization for novel problems and tracking long-term autonomy in interactive environments. The development of 'living' benchmarks that co-evolve with AI capabilities, alongside the integration of socio-technical and economic metrics, will be critical for ensuring these advanced systems are not only powerful but also reliable, efficient, and aligned with human values in real-world applications [310, 1368, 1330].

The evaluation landscape for context-engineered systems continues evolving rapidly as new architectures, capabilities, and applications emerge. Future evaluation paradigms must address increasing system complexity while providing reliable, comprehensive, and actionable insights for system improvement and deployment decisions. The integration of multiple evaluation approaches—from component-level assessment to system-wide robustness testing—represents a critical research priority for ensuring the reliable deployment of context-engineered systems in real-world applications [835, 1132].

7. Future Directions and Open Challenges

Context Engineering stands at a critical inflection point where foundational advances converge with emerging application demands, creating unprecedented opportunities for innovation while revealing fundamental challenges that require sustained research efforts across multiple dimensions [835, 1132].

As the field transitions from isolated component development toward integrated system architectures, the complexity of research challenges grows exponentially, demanding interdisciplinary approaches that bridge theoretical computer science, practical system engineering, and domain-specific expertise [310, 931].

独立验证程序以解决依赖 LLM 自我验证能力的系统局限性 [128, 390]。

6.3.3. 安全性和鲁棒性评估

面向安全的评估包含全面的鲁棒性测试、对抗性攻击抵抗和对齐评估，以确保上下文工程系统的负责任开发。必须特别关注对能够在长时间内自主运行的代理系统的评估，因为这些系统提出了传统评估框架无法充分解决的独特安全挑战 [965, 360]。

鲁棒性评估必须通过全面的压力测试协议，在分布偏移、输入扰动和对抗性条件下评估系统性能。多代理系统在协调失败场景中面临额外挑战，其中部分系统故障可能通过整个代理网络级联。评估框架必须解决优雅降级策略、错误恢复协议以及在不利条件下保持系统功能的能力。除了预定义的故障模式之外，未来的评估必须应对评估对“未知未知”的弹性——高度复杂、自主多代理系统中涌现的不可预测的故障级联 [128, 390]。

对齐评估衡量系统对预期行为的遵循程度、价值一致性以及通过专业评估框架优化有益结果。由于上下文工程系统具有动态适应能力和跨多个组件的复杂交互模式，因此它们面临着独特的对齐挑战。长期评估必须评估系统在通过扩展运行周期进行适应和演变时是否保持有益行为 [893]。

展望未来，对上下文工程系统的评估需要从静态基准测试转向动态、整体评估。未来的框架必须超越衡量任务成功，转而评估新问题的组合泛化能力，并跟踪交互环境中的长期自主性。开发与AI能力共同演变的“动态”基准测试，以及整合社会技术和经济指标，对于确保这些先进系统不仅功能强大，而且可靠、高效，并在实际应用中与人类价值观保持一致至关重要 [310, 1368, 1330]。

上下文工程系统的评估领域随着新架构、能力和应用的不断涌现而迅速发展。未来的评估范式必须在解决系统日益复杂性的同时，为系统改进和部署决策提供可靠、全面和可操作的见解。从组件级评估到系统级鲁棒性测试的多种评估方法的整合，是确保上下文工程系统在实际应用中可靠部署的关键研究优先事项 [835, 1132]。

7. 未来方向和开放挑战

上下文工程正处在一个关键的拐点，基础性进展与新兴应用需求交汇，为创新创造了前所未有的机遇，同时揭示了需要在多个维度上持续研究的基本挑战 [835, 1132]。

随着该领域从孤立组件开发转向集成系统架构，研究挑战的复杂性呈指数级增长，需要跨学科方法来连接理论计算机科学、实际系统工程和特定领域专业知识 [310, 931]。

This section systematically examines key research directions and open challenges that will define the evolution of Context Engineering over the coming decade.

7.1. Foundational Research Challenges

This subsection examines core theoretical and computational challenges that must be addressed to advance context engineering systems beyond current limitations.

7.1.1. Theoretical Foundations and Unified Frameworks

Context Engineering currently operates without unified theoretical foundations that connect disparate techniques and provide principled design guidelines, representing a critical research gap that limits systematic progress and optimal system development.

The absence of mathematical frameworks characterizing context engineering capabilities, limitations, and optimal design principles across different architectural configurations impedes both fundamental understanding and practical optimization [1132, 663, 835, 310].

Information-theoretic analysis of context engineering systems requires comprehensive investigation into optimal context allocation strategies, information redundancy quantification, and fundamental compression limits within context windows. Current approaches lack principled methods for determining optimal context composition, leading to suboptimal resource utilization and performance degradation. Research must establish mathematical bounds on context efficiency, develop optimization algorithms for context selection, and create theoretical frameworks for predicting system behavior across varying context configurations [731, 295].

Compositional understanding of context engineering systems demands formal models describing how individual components interact, interfere, and synergize within integrated architectures. The emergence of complex behaviors from component interactions requires systematic investigation through both empirical studies and theoretical modeling approaches. Multi-agent orchestration presents particular challenges in developing mathematical frameworks for predicting coordination effectiveness and emergent collaborative behaviors [128, 893].

7.1.2. Scaling Laws and Computational Efficiency

The fundamental asymmetry between LLMs' remarkable comprehension capabilities and their pronounced generation limitations represents one of the most critical challenges in Context Engineering research.

This comprehension-generation gap manifests across multiple dimensions including long-form output coherence, factual consistency maintenance, and planning sophistication, requiring investigation into whether limitations stem from architectural constraints, training methodologies, or fundamental computational boundaries [835, 1132].

Long-form generation capabilities demand systematic investigation into planning mechanisms that can maintain coherence across thousands of tokens while preserving factual accuracy and logical consistency. Current systems exhibit significant performance degradation in extended generation tasks, highlighting the need for architectural innovations beyond traditional transformer paradigms. State space models including Mamba demonstrate potential for more efficient long sequence processing through linear scaling properties, though current implementations require substantial development to match transformer performance across diverse tasks [731, 1258, 347, 216].

本节系统地考察了将定义未来十年 ContextEngineering 进化的关键研究方向和开放性挑战。

7.1. 基础研究挑战

本小节考察了必须解决的核心理论和计算挑战，以推动上下文工程系统超越当前限制。

7.1.1. 理论基础和统一框架

上下文工程目前缺乏统一的理论基础，无法连接不同的技术并提供原则性的设计指南，这代表了一个关键的研究差距，限制了系统性进展和最佳系统开发。

缺乏表征上下文工程能力、限制和不同架构配置下最佳设计原则的数学框架，阻碍了基本理解和实际优化 [1132, 663, 835, 310]。

对上下文工程系统的信息论分析需要对最优上下文分配策略、信息冗余量化以及上下文窗口内的基本压缩极限进行深入调查。当前方法缺乏确定最优上下文组成的原理性方法，导致资源利用不足和性能下降。研究必须建立上下文效率的数学界限，开发上下文选择的优化算法，并创建预测系统行为随不同上下文配置变化的理论框架[731, 295]。

对上下文工程系统的组合理解需要正式模型描述单个组件如何交互、干扰和协同工作在集成架构中。从组件交互中出现的复杂行为需要通过实证研究和理论建模方法进行系统调查。多智能体编排在开发预测协调有效性和涌现协作行为的数学框架方面提出了特殊挑战 [128, 893]。

7.1.2. 扩展定律和计算效率

LLM的卓越理解能力与其明显的生成限制之间的基本不对称性是上下文工程研究中最关键的挑战之一。

这种理解-生成差距体现在多个维度，包括长文本输出的连贯性、事实一致性维护和规划复杂性，需要调查限制是否源于架构约束、训练方法或基本计算边界 [835, 1132]。

长文本生成能力需要系统地研究能够维持数千个token的连贯性，同时保持事实准确性和逻辑一致性的规划机制。当前系统在扩展生成任务中表现出显著的性能下降，突出了超越传统Transformer范式的架构创新需求。状态空间模型（包括Mamba）通过线性缩放特性展示了在更高效的长序列处理方面的潜力，尽管当前的实现需要大量开发才能在多样化任务中匹配Transformer的性能 [731, 1258, 347, 216]。

Context scaling efficiency faces fundamental computational challenges, with current attention mechanisms scaling quadratically ($O(n^2)$) with sequence length, creating prohibitive memory and computational requirements for ultra-long sequences. Sliding attention mechanisms and memory-efficient implementations represent promising directions, though significant research is needed to address both computational tractability and reasoning quality preservation [295, 1227, 347]. Position interpolation and extension techniques require advancement to handle sequences exceeding current architectural limitations while maintaining positional understanding and coherence.

7.1.3. Multi-Modal Integration and Representation

The integration of diverse modalities within context engineering systems presents fundamental challenges in representation learning, cross-modal reasoning, and unified architectural design. Current approaches typically employ modality-specific encoders with limited cross-modal interaction, failing to capture the rich interdependencies that characterize sophisticated multi-modal understanding. VideoWebArena demonstrates the complexity of multimodal agent evaluation, revealing substantial performance gaps in current systems when processing video, audio, and text simultaneously [476].

Beyond these sensory modalities, context engineering must also handle more abstract forms of information such as graphs, whose structural semantics are not directly interpretable by language models. Capturing the high-level meaning encoded in graph structures introduces unique challenges, including aligning graph representations with language model embeddings and expressing graph topology efficiently. Recent efforts like GraphGPT [1024] and GraphRAG [244] attempt to bridge this gap through cross-modal alignment strategies, while others explore converting graphs into natural language descriptions to facilitate model understanding [262, 319]. Bi et al. [75] further propose a divide-and-conquer approach to encode text-attributed heterogeneous networks, addressing context length limitations and enabling effective link prediction. Graph reasoning thus emerges as a core difficulty in context engineering, requiring models to navigate complex relational structures beyond raw modalities.

Temporal reasoning across multi-modal contexts requires sophisticated architectures capable of tracking object persistence, causal relationships, and temporal dynamics across extended sequences. Web agent frameworks including WebArena showcase the challenges of maintaining coherent understanding across complex multi-step interactions involving diverse modalities and dynamic content. Current systems demonstrate significant limitations in coordinating multi-modal information processing with action planning and execution [1368, 202].

Cross-modal alignment and consistency present ongoing challenges in ensuring that information extracted from different modalities remains factually consistent and semantically coherent. Deep Research Bench evaluation reveals that current multi-modal agents struggle with complex research tasks requiring synthesis across textual, visual, and structured data sources, highlighting the need for more sophisticated alignment mechanisms [87].

7.2. Technical Innovation Opportunities

This subsection explores emerging technical approaches and architectural innovations that promise to enhance context engineering capabilities.

上下文扩展效率面临根本性的计算挑战，当前的注意力机制随序列长度呈二次方（$O(n_2)$）扩展，为超长序列带来了过高的内存和计算需求。滑动注意力机制和内存高效的实现方式代表了有前景的方向，尽管需要大量研究来应对计算可行性和推理质量保留的挑战 [295, 1227, 347]。位置插值和扩展技术需要进一步发展，以处理超出当前架构限制的序列，同时保持位置理解和连贯性。

7.1.3. 多模态集成与表示

在上下文工程系统中集成多种模态在表示学习、跨模态推理和统一架构设计方面面临根本性挑战。当前方法通常采用特定模态的编码器，跨模态交互有限，未能捕捉到复杂多模态理解的丰富相互依赖关系。VideoWebArena展示了多模态智能体评估的复杂性，揭示了当前系统在同时处理视频、音频和文本时存在的巨大性能差距 [476]。

除了这些感觉模态之外，上下文工程还必须处理图等更抽象的信息形式，其结构语义不能直接被语言模型解释。捕捉图中编码的高级含义引入了独特的挑战，包括将图表示与语言模型嵌入对齐以及高效表达图拓扑。近期努力如GraphGPT [1024] 和GraphRAG [244] 通过跨模态对齐策略试图弥合这一差距，而其他人则探索将图转换为自然语言描述以促进模型理解 [262, 319]。Bi等人 [75] 进一步提出了一种分而治之的方法来编码文本属性异构网络，解决了上下文长度限制并实现了有效的链接预测。因此，图推理成为上下文工程的核心难点，要求模型在原始模态之外导航复杂的关联结构。

跨多模态上下文的时间推理需要能够跟踪对象持久性、因果关系和跨长序列的时间动态的复杂架构。Web代理框架，包括WebArena，展示了在涉及多种模态和动态内容的复杂多步交互中保持连贯理解的挑战。当前系统在协调多模态信息处理与行动规划和执行方面表现出显著局限性 [1368, 202]。

跨模态对齐和一致性在确保从不同模态中提取的信息保持事实一致性和语义连贯性方面仍然存在持续挑战。Deep Research Bench评估表明，当前的多模态代理在需要跨文本、视觉和结构化数据源进行综合的复杂研究任务中存在困难，突出了对更复杂对齐机制的需求 [87]。

7.2. 技术创新机会

本小节探讨了新兴技术方法和架构创新，这些方法有望增强上下文工程能力。

7.2.1. Next-Generation Architectures

Architectural innovations beyond traditional transformer paradigms offer promising directions for addressing current limitations in context engineering systems. State space models including LongMamba demonstrate potential for more efficient long sequence processing through linear scaling properties and improved memory utilization, though current implementations require substantial development to match transformer performance across diverse tasks [1258, 731]. Specialized position encoding methods and parameter-efficient architectures present opportunities for scaling to ultra-long sequences while maintaining computational tractability [347, 295].

Memory-augmented architectures require advancement beyond current external memory mechanisms to enable more sophisticated long-term memory organization, hierarchical memory structures, and adaptive memory management strategies. MemoryBank implementations incorporating Ebbinghaus Forgetting Curve principles demonstrate promising approaches to memory persistence, though significant research is needed to address the fundamental stateless nature of current LLMs [1362, 1330, 1171, 813, 1202]. The development of episodic memory systems capable of maintaining coherent long-term context across extended interactions represents a critical architectural challenge [457, 841, 393].

Modular and compositional architectures enable flexible system construction through specialized component integration while maintaining overall system coherence. Modular RAG architectures demonstrate enhanced flexibility through specialized modules for retrieval, augmentation, and generation, enabling fine-grained optimization of individual components. Graph-enhanced approaches including GraphRAG and LightRAG showcase the potential for integrating structured knowledge representation with neural processing [312, 965, 360].

7.2.2. Advanced Reasoning and Planning

Context engineering systems require enhanced reasoning capabilities spanning causal reasoning, counterfactual thinking, temporal reasoning, and analogical reasoning across extended contexts. Current systems demonstrate limited capacity for sophisticated reasoning patterns that require integration of multiple evidence sources, consideration of alternative scenarios, and maintenance of logical consistency across complex inference chains [1132, 835].

Multi-step planning and execution capabilities represent critical advancement areas enabling systems to decompose complex tasks, formulate execution strategies, monitor progress, and adapt plans based on intermediate results. Agentic RAG systems demonstrate sophisticated planning and reflection mechanisms requiring integration of task decomposition, multi-plan selection, and iterative refinement capabilities. However, current implementations face significant challenges in maintaining coherence across extended planning horizons and adapting to dynamic information conditions [438, 162, 1183].

Tool-integrated reasoning represents a paradigmatic advancement requiring dynamic interaction with external resources during reasoning processes. The GAIA benchmark demonstrates substantial performance gaps, with human achievement of 92% accuracy compared to advanced models achieving only 15%, highlighting fundamental limitations in current reasoning and planning capabilities [772, 1090, 126]. Advanced tool integration must address autonomous tool selection, parameter extraction, multi-tool coordination, and error recovery across diverse operational contexts [310, 931].

7.2.1. 下一代架构

超越传统 Transformer 范式的架构创新为解决当前上下文工程系统的局限性提供了有前景的方向。包含 LongMamba 的状态空间模型通过线性缩放特性和改进的内存利用率，展示了在更高效长序列处理方面的潜力，尽管当前的实现需要大量开发才能在多样化任务中匹配 Transformer 的性能 [1258, 731]。专门的位置编码方法和参数高效的架构为扩展到超长序列同时保持计算可行性提供了机会 [347, 295]。

内存增强架构需要在当前外部内存机制的基础上取得进步，以实现更复杂的长期记忆组织、分层内存结构和自适应内存管理策略。结合艾宾浩斯遗忘曲线原理的 MemoryBank 实现展示了内存持久化的有前景的方法，尽管需要大量研究来解决当前 LLM 的基本无状态特性 [1362, 1330, 1171, 813, 1202]。开发能够维持跨长时间交互的连贯长期上下文的情景记忆系统代表了一个关键的架构挑战 [457, 841, 393]。

模块化和组合式架构通过专业组件集成实现灵活的系统构建，同时保持整体系统一致性。模块化 RAG 架构通过检索、增强和生成等专用模块展示了增强的灵活性，实现了对单个组件的细粒度优化。图增强方法，包括 GraphRAG 和 LightRAG，展示了将结构化知识表示与神经处理集成的潜力 [312, 965, 360]。

7.2.2. 高级推理和规划

上下文工程系统需要增强的推理能力，涵盖因果推理、反事实思考、时间推理和类比推理等跨扩展上下文的能力。当前系统在需要整合多个证据来源、考虑替代场景以及维护复杂推理链中逻辑一致性的复杂推理模式方面表现出有限的容量 [1132, 835]。

多步规划和执行能力代表了关键的进步领域，使系统能够分解复杂任务、制定执行策略、监控进度并根据中间结果调整计划。代理式 RAG 系统展示了复杂的规划和反思机制，需要集成任务分解、多计划选择和迭代优化能力。然而，当前实现面临在扩展规划范围内保持一致性和适应动态信息条件方面的重大挑战 [438, 162, 1183]。

工具集成推理代表了一种范式上的进步，需要在推理过程中与外部资源进行动态交互。GAIA基准测试显示了显著的性能差距，人类达到了92%的准确率，而先进模型仅达到15%，突出了当前推理和规划能力的根本局限性 [772, 1090, 126]。高级工具集成必须解决自主工具选择、参数提取、多工具协调和跨不同操作环境进行错误恢复的问题 [310, 931]。

7.2.3. Complex Context Organization and Solving Graph Problems

Graph reasoning represents a fundamental challenge in context engineering, requiring systems to navigate complex structural relationships while maintaining semantic understanding across interconnected elements. Recent advances in graph-language model integration demonstrate multiple paradigms: specialized architectural approaches that incorporate graph-specific components and text-based encoding strategies that transform graph structures into natural language representations [1085, 1023].

Architectural integration approaches include GraphGPT, which employs dual-stage instruction tuning aligning graph structural information with language tokens via self-supervised graph matching [1023, 741]. This framework introduces specialized GraphTokens refined through Graph Instruction Tuning and utilizes a lightweight graph-text alignment projector for transitioning between textual and structural processing modalities [1270, 274]. Building upon instruction-tuning paradigms, GraphWiz extends this approach by incorporating DPO to enhance reasoning reliability, achieving 65% average accuracy across diverse graph tasks and significantly outperforming GPT-4's 43.8% [145]. Chain-of-thought distillation mechanisms enhance step-by-step reasoning performance [1138, 1391]. RL presents another promising direction, as demonstrated by G1, which trains LLMs on synthetic graph-theoretic tasks using the Erdős dataset comprising 50 diverse tasks, achieving strong zero-shot generalization with a 3B parameter model outperforming significantly larger models [357].

Text-based encoding approaches transform graph structures into natural language descriptions using few-shot prompting and chain-of-thought reasoning without architectural modifications [262, 192]. These methods introduce diverse graph description templates contextualizing structural elements through multiple semantic interpretations [936, 716]. Recent work investigates the impact of graph description ordering on LLM performance, revealing that sequential presentation significantly influences model comprehension and reasoning accuracy [319]. Benchmark evaluations have expanded with GraphArena, offering both polynomial-time tasks and NP-complete challenges with a rigorous evaluation framework that classifies outputs as correct, suboptimal, hallucinatory, or missing [1025]. Combined with existing benchmarks like NLGraph and GraphDO, these evaluations reveal substantial performance disparities between simple connectivity problems and complex tasks like maximum flow computation [1085, 895, 319].

Current implementations face challenges in scaling to large structures, maintaining consistency across multi-hop relationships, and generalizing to novel topologies, with text-based approaches offering interpretability at reduced structural precision while specialized architectures provide enhanced performance through increased complexity [889, 1100]. Emerging hybrid approaches including InstructGraph and GraphAdapter attempt to bridge these paradigms through structured format verbalizers and GNN-based adapters, though limitations persist in handling dynamic structures and temporal evolution of relationships [261]. Looking forward, broad connection paradigms that organize information through associative networks rather than fragmented searches, spreading outward from central nodes to discover potential connections between entities, may represent the next generation of RAG systems for complex context organization [131].

7.2.4. Intelligent Context Assembly and Optimization

Automated context engineering systems capable of intelligently assembling contexts from available components represent a critical research frontier requiring development of context optimization algorithms, adaptive selection strategies, and learned assembly functions. Current approaches rely heavily on heuristic methods and domain-specific engineering, limiting scalability and optimality across diverse applications [1132, 663].

7.2.3. 复杂上下文组织与图问题求解

图推理代表了上下文工程中的一个基本挑战，要求系统在维护互联元素之间语义理解的同时，能够导航复杂的结构关系。近年来，图-语言模型集成方面的进展展示了多种范式：包含特定图组件的专用架构方法，以及将图结构转换为自然语言表示的基于文本的编码策略 [1085, 1023]。

架构集成方法包括GraphGPT，它采用双阶段指令微调，通过自监督图匹配将图结构信息与语言标记对齐 [1023, 741]。该框架引入了通过图指令微调精炼的专用GraphTokens，并利用轻量级的图-文本对齐投影器在文本处理和结构处理模态之间进行转换 [1270, 274]。基于指令微调范式，GraphWiz通过结合DPO增强了推理可靠性，在多样化的图任务中实现了65%的平均准确率，显著优于GPT-4的43.8% [145]。思维链蒸馏机制提升了逐步推理性能 [1138, 1391]。强化学习（RL）是另一个有前景的方向，如G1所示，它使用包含50个多样化任务的Erdős数据集在合成图论任务上训练LLM，使用3B参数模型实现了强大的零样本泛化，显著优于更大的模型 [357]。

基于文本的编码方法通过少量样本提示和思维链推理将图结构转换为自然语言描述，而无需进行架构修改 [262, 192]。这些方法引入了多样的图描述模板，通过多种语义解释来 contextualizing 结构元素 [936, 716]。最近的研究调查了图描述顺序对 LLM 性能的影响，揭示顺序呈现显著影响模型理解和推理准确性 [319]。基准评估随着 GraphArena 的扩展而扩展，提供多项式时间任务和 NP 完全挑战，并采用严格的评估框架将输出分类为正确、次优、幻觉或缺失 [1025]。结合现有的 NLGraph 和 GraphDO 等基准，这些评估揭示了简单连通性问题与复杂任务（如最大流计算）之间的性能差异 [1085, 895, 319]。

当前的实现面临扩展到大型结构的挑战，跨多跳关系保持一致性，以及泛化到新颖拓扑结构的挑战，基于文本的方法在降低结构精度的情况下提供可解释性，而专用架构通过增加复杂性提供增强性能 [889, 1100]。新兴的混合方法，包括 InstructGraph 和 GraphAdapter，试图通过结构化格式语言化器和基于 GNN 的适配器来弥合这些范式，尽管在处理动态结构和关系的时序演变方面仍存在局限性 [261]。展望未来，通过关联网络组织信息的广泛连接范式，而不是碎片化搜索，从中心节点向外扩散以发现实体之间的潜在连接，可能代表复杂上下文组织的下一代 RAG 系统 [131]。

7.2.4. 智能上下文组装与优化

能够从可用组件中智能组装上下文的自动化上下文工程系统代表一个关键的研究前沿，需要开发上下文优化算法、自适应选择策略和学习的组装函数。当前方法严重依赖启发式方法和特定领域的工程，限制了在不同应用中的可扩展性和最优性[1132, 663]。

Self-refinement mechanisms demonstrate substantial potential for intelligent context optimization through iterative improvement processes. Self-Refine, Reflexion, and N-CRITICS frameworks achieve significant performance improvements, with GPT-4 demonstrating approximately 20% improvement through iterative refinement. However, these approaches require advancement in optimization strategies for autonomous evolution and meta-learning across diverse contexts [735, 956, 789, 577].

Multi-dimensional feedback mechanisms incorporating diverse feedback dimensions including correctness, relevance, clarity, and robustness provide promising directions for context optimization. Self-rewarding mechanisms enable autonomous evolution capabilities, though research must address fundamental questions about optimal adaptation rates, stability-plasticity trade-offs, and preservation of beneficial adaptations across varying operational conditions [704].

7.3. Application-Driven Research Directions

This subsection addresses research challenges arising from real-world deployment requirements and domain-specific applications.

7.3.1. Domain Specialization and Adaptation

Context engineering systems require sophisticated specialization mechanisms for diverse domains including healthcare, legal analysis, scientific research, education, and engineering applications, each presenting unique requirements for knowledge integration, reasoning patterns, safety considerations, and regulatory compliance. Domain-specific optimization demands investigation into transfer learning strategies, domain adaptation techniques, and specialized training paradigms that preserve general capabilities while enhancing domain-specific performance [1132, 663].

Scientific research applications require sophisticated reasoning capabilities over complex technical content, mathematical expressions, experimental data, and theoretical frameworks while maintaining rigorous accuracy standards. Deep Research Bench evaluation reveals significant challenges in current systems' ability to conduct complex research tasks requiring synthesis across multiple information sources and reasoning over technical content. Research must address integration of symbolic reasoning with neural approaches and incorporation of domain-specific knowledge bases [87].

Healthcare applications demand comprehensive safety evaluation frameworks, regulatory compliance mechanisms, privacy protection protocols, and integration with existing clinical workflows while maintaining interpretability and auditability requirements. Medical context engineering must address challenges in handling sensitive information, ensuring clinical accuracy, supporting diagnostic reasoning, and maintaining patient privacy across complex healthcare ecosystems. Current evaluation frameworks reveal substantial gaps in medical reasoning capabilities and safety assessment methodologies [386].

7.3.2. Large-Scale Multi-Agent Coordination

Scaling multi-agent context engineering systems to hundreds or thousands of participating agents requires development of distributed coordination mechanisms, efficient communication protocols, and hierarchical management structures that maintain system coherence while enabling local autonomy. Research must address fundamental challenges in distributed consensus, fault tolerance, and emergent behavior prediction in large-scale agent populations [239, 140].

Communication protocol standardization represents a critical research frontier, with emerging protocols

自省机制通过迭代改进过程展示了在智能上下文优化方面的巨大潜力。Self-Refine、Reflexion 和 N-CRITICS 框架实现了显著的性能提升，GPT-4 通过迭代改进实现了约 20% 的提升。然而，这些方法需要优化策略的进步，以实现自主进化和跨不同上下文的元学习 [735, 956, 789, 577]。

包含正确性、相关性、清晰性和鲁棒性等多种反馈维度的多维反馈机制为上下文优化提供了有前景的方向。自奖励机制使能自主进化能力，尽管研究必须解决关于最优适应率、稳定性-可塑性权衡以及在不同操作条件下保留有益适应性的基本问题 [704]。

7.3. Application-Driven Research Directions

This 本小节探讨了由实际部署需求和专业领域应用引发的研究挑战。 -
spec

7.3.1. 领域专业化和适应

上下文工程系统需要对包括医疗保健、法律分析、科学研究、教育和工程应用在内的不同领域进行复杂的专业化机制，每个领域都提出了独特的知识集成、推理模式、安全考虑和监管合规要求。领域特定优化需要研究迁移学习策略、领域适应技术和专门的训练范式，这些范式在保留通用能力的同时增强了领域特定性能 [1132, 663]。

科学研究应用需要复杂的推理能力来处理复杂的技术内容、数学表达式、实验数据和理论框架，同时保持严格的准确性标准。深度研究基准评估揭示了当前系统在执行需要跨多个信息源综合和推理技术内容的研究任务方面的重大挑战。研究必须解决符号推理与神经方法的集成以及领域特定知识库的整合问题 [87]。

医疗保健应用需要全面的安全评估框架、监管合规机制、隐私保护协议，并与现有的临床工作流程集成，同时保持可解释性和可审计性要求。医疗上下文工程必须解决处理敏感信息、确保临床准确性、支持诊断推理以及在整个复杂的医疗保健生态系统中维护患者隐私的挑战。当前的评估框架揭示了在医疗推理能力和安全评估方法方面的巨大差距 [386]。

7.3.2. Large-Scale Multi-Agent Coordination

将多智能体上下文工程系统扩展到数百或数千个参与智能体需要开发分布式协调机制、高效的通信协议和分层管理结构，这些结构在保持系统一致性的同时能够实现局部自主性。研究必须解决大规模智能体群体中分布式共识、容错性和涌现行为预测的基本挑战 [239, 140]。

通信协议标准化代表了一个关键的研究前沿，新兴协议

including MCP (“USB-C for AI”), A2A (Agent-to-Agent), ACP (Agent Communication Protocol), and ANP (Agent Network Protocol) demonstrating the need for unified frameworks enabling interoperability across diverse agent ecosystems. However, current implementations face security vulnerabilities and scalability limitations that must be addressed for large-scale deployment [37, 1007, 462, 1, 246, 926, 616].

Orchestration challenges including transactional integrity, context management, and coordination strategy effectiveness represent significant obstacles to large-scale multi-agent deployment. Contemporary frameworks including LangGraph, AutoGen, and CAMEL demonstrate insufficient transaction support and validation limitations, requiring systems that rely exclusively on LLM self-validation capabilities. Advanced coordination frameworks must address compensation mechanisms for partial failures and maintain system coherence under varying operational conditions [128, 390, 893].

7.3.3. Human-AI Collaboration and Integration

Sophisticated human-AI collaboration frameworks require deep understanding of human cognitive processes, communication preferences, trust dynamics, and collaboration patterns to enable effective hybrid teams that leverage complementary strengths. Research must investigate optimal task allocation strategies, communication protocols, and shared mental model development between humans and AI systems [1132, 835].

Web agent evaluation frameworks reveal significant challenges in human-AI collaboration, particularly in complex task scenarios requiring sustained interaction and coordination. WebArena and Mind2Web demonstrate that current systems struggle with multi-step interactions across diverse websites, highlighting fundamental gaps in collaborative task execution. Advanced interfaces require investigation into context-aware adaptation and personalization mechanisms that enhance human-AI team performance [1368, 202].

Trust calibration and transparency mechanisms represent critical research areas for ensuring appropriate human reliance on AI systems while maintaining human agency and decision-making authority. Evaluation frameworks must address explanation generation, uncertainty communication, and confidence calibration to support informed human decision-making in collaborative scenarios. The substantial performance gaps revealed by benchmarks like GAIA underscore the importance of developing systems that can effectively communicate their limitations and capabilities [772, 1090].

7.4. Deployment and Societal Impact Considerations

This subsection examines critical considerations for deploying context engineering systems at scale while ensuring responsible and beneficial outcomes.

7.4.1. Scalability and Production Deployment

Production deployment of context engineering systems requires addressing scalability challenges across multiple dimensions including computational resource management, latency optimization, throughput maximization, and cost efficiency while maintaining consistent performance across diverse operational conditions. The $O(n^2)$ scaling limitation of current attention mechanisms creates substantial barriers to deploying ultra-long context systems in production environments, necessitating advancement in memory-efficient architectures and sliding attention mechanisms [295, 1227].

Reliability and fault tolerance mechanisms become critical as context engineering systems assume increasingly important roles in decision-making processes across domains. Multi-agent orchestration frameworks

包括MCP（“USB-C用于AI”）、A2A（Agent-to-Agent）、ACP（Agent通信协议）和ANP（Agent网络协议），展示了统一框架的需求，该框架能够实现不同智能体生态系统之间的互操作性。然而，当前的实现面临安全漏洞和可扩展性限制，这些问题必须得到解决才能进行大规模部署 [37, 1007, 462, 1, 246, 926, 616]。

编排挑战包括事务完整性、上下文管理和协调策略有效性，这些挑战代表了大规模多智能体部署的重大障碍。当代框架包括LangGraph、AutoGen和CAMEL展示了事务支持不足和验证限制，需要依赖仅基于LLM自我验证能力的系统。高级协调框架必须解决部分失败情况下的补偿机制，并在不同的运行条件下保持系统一致性 [128, 390, 893]。

7.3.3. 人类-人工智能协作与集成

复杂的智能体-人类协作框架需要深入理解人类认知过程、沟通偏好、信任动态和协作模式，以实现有效利用互补优势的混合团队。研究必须调查人类和人工智能系统之间的最佳任务分配策略、沟通协议和共享心智模型开发[1132, 835]。

Web agent evaluation frameworks揭示出人类-AI协作中的重大挑战，特别是在需要持续交互和协调的复杂任务场景中。WebArena和Mind2Web表明当前系统在跨不同网站的多步交互中存在困难，突显了协作任务执行中的基本差距。高级界面需要研究上下文感知的适应和个性化机制，以提升人类-AI团队性能 [1368, 202]。

信任校准和透明机制代表确保人类适当依赖AI系统，同时保持人类自主性和决策权的关键研究领域。评估框架必须解决解释生成、不确定性沟通和置信度校准，以支持人类在协作场景中的知情决策。像GAIA这样的基准揭示的性能差距强调了开发能够有效沟通其局限性和能力的系统的重要性 [772, 1090]。

7.4. 部署和社会影响考虑

本小节探讨了在规模化部署上下文工程系统时，确保负责任和有益结果的关键考虑因素。

7.4.1. 可扩展性和生产部署

上下文工程系统的生产部署需要解决跨多个维度的可扩展性挑战，包括计算资源管理、延迟优化、吞吐量最大化和成本效率，同时保持在不同操作条件下的性能一致性。当前注意力机制的时间复杂度 $O(n_2)$ 限制了其扩展性，为在生产环境中部署超长上下文系统制造了重大障碍，迫切需要改进内存高效的架构和滑动注意力机制 [295, 1227]。

随着上下文工程系统在各个领域的决策过程中承担越来越重要的角色，可靠性和容错机制变得至关重要。多代理编排框架

face particular challenges in maintaining transactional integrity across complex workflows, with current systems lacking adequate compensation mechanisms for partial failures. Research must address graceful degradation strategies, error recovery protocols, and redundancy mechanisms that maintain system functionality under adverse conditions [128, 390].

Maintainability and evolution challenges require investigation into system versioning, backward compatibility, continuous integration protocols, and automated testing frameworks that support ongoing system improvement without disrupting deployed services. Memory system implementations face additional challenges due to the stateless nature of current LLMs and the lack of standardized benchmarks for long-term memory persistence and retrieval efficiency [1330, 1171].

7.4.2. Safety, Security, and Robustness

Comprehensive safety evaluation requires development of assessment frameworks that can identify potential failure modes, safety violations, and unintended behaviors across the full spectrum of context engineering system capabilities. Agentic systems present particular safety challenges due to their autonomous operation capabilities and complex interaction patterns across extended operational periods [965, 360].

Security considerations encompass protection against adversarial attacks, data poisoning, prompt injection, model extraction, and privacy violations while maintaining system functionality and usability. Multi-agent communication protocols including MCP, A2A, and ACP introduce security vulnerabilities that must be addressed while preserving interoperability and functionality. Research must develop defense mechanisms and detection systems that address evolving threat landscapes across distributed agent networks [246, 926].

Alignment and value specification challenges require investigation into methods for ensuring context engineering systems behave according to intended objectives while avoiding specification gaming, reward hacking, and goal misalignment. Context engineering systems present unique alignment challenges due to their dynamic adaptation capabilities and complex interaction patterns across multiple components. The substantial performance gaps revealed by evaluation frameworks underscore the importance of developing robust alignment mechanisms that can maintain beneficial behaviors as systems evolve and adapt [772, 128].

7.4.3. Ethical Considerations and Responsible Development

Bias mitigation and fairness evaluation require comprehensive assessment frameworks that can identify and address systematic biases across different demographic groups, application domains, and use cases while maintaining system performance and utility. Research must investigate bias sources in training data, model architectures, and deployment contexts while developing mitigation strategies that address root causes rather than symptoms [1132, 835].

Privacy protection mechanisms must address challenges in handling sensitive information, preventing data leakage, and maintaining user privacy while enabling beneficial system capabilities. Memory systems face particular privacy challenges due to their persistent information storage and retrieval capabilities, requiring advanced frameworks for secure memory management and selective forgetting mechanisms [1330, 457].

Transparency and accountability frameworks require development of explanation systems, audit mechanisms, and governance structures that enable responsible oversight of context engineering systems while supporting innovation and beneficial applications. The substantial performance gaps revealed by evaluation frameworks like GAIA (human 92% vs AI 15%) highlight the importance of transparent capability communication and appropriate expectation setting for deployed systems [772, 1090].

在维护复杂工作流中的事务完整性方面面临特殊挑战，当前系统缺乏对部分故障的充分补偿机制。研究必须解决优雅降级策略、错误恢复协议和冗余机制，这些机制能在不利条件下保持系统功能 [128, 390]。

可维护性和进化挑战需要研究系统版本控制、向后兼容性、持续集成协议和自动化测试框架，这些框架支持持续的系统改进而不中断已部署的服务。由于当前LLM的无状态特性以及长期内存持久性和检索效率缺乏标准化基准，内存系统实现面临额外的挑战 [1330, 1171]。

7.4.2. 安全性、安全性和鲁棒性

全面的安全评估需要开发评估框架，这些框架能够识别潜在故障模式、安全违规行为和非预期行为，涵盖上下文工程系统能力的整个范围。由于自主操作能力和长时间运行期间复杂的交互模式，代理系统带来了特殊的安全挑战 [965, 360]。

安全考虑涵盖防范对抗性攻击、数据中毒、提示注入、模型提取和隐私侵犯，同时保持系统功能性和可用性。包含MCP、A2A和ACP的多智能体通信协议引入了安全漏洞，必须在保持互操作性和功能性的同时加以解决。研究必须开发防御机制和检测系统，以应对分布式智能体网络中不断变化的威胁环境 [246, 926]。

对齐和价值规范挑战需要研究确保上下文工程系统根据预期目标运行的方法，同时避免规范博弈、奖励黑客和目标错位。由于上下文工程系统具有动态适应能力和跨多个组件的复杂交互模式，它们呈现出独特的对齐挑战。评估框架揭示的显著性能差距强调了开发能够维持有益行为随系统演进而适应的鲁棒对齐机制的重要性 [772, 128]。

7.4.3. 伦理考量与负责任开发

偏差缓解和公平性评估需要全面的评估框架，能够识别和解决跨不同人口群体、应用领域和使用场景的系统偏差，同时保持系统性能和效用。研究必须调查训练数据、模型架构和部署环境中的偏差来源，同时开发缓解策略以解决根本原因而非症状 [1132, 835]。

隐私保护机制必须解决处理敏感信息、防止数据泄露和维护用户隐私，同时启用有益的系统功能所面临的挑战。由于内存系统具有持久信息存储和检索能力，它们面临着特殊的隐私挑战，需要先进的框架来确保内存安全管理和选择性遗忘机制 [1330, 457]。

透明度和问责制框架要求开发解释系统、审计机制和治理结构，这些结构能够对上下文工程系统进行负责任的监督，同时支持创新和有益的应用。评估框架（如GAIA）揭示的巨大性能差距（人类92% vs AI 15%）突出了透明能力沟通和为部署系统设置适当期望的重要性 [772, 1090]。

The future of Context Engineering will be shaped by our ability to address these interconnected challenges through sustained, collaborative research efforts that bridge technical innovation with societal considerations.

Success will require continued investment in fundamental research, interdisciplinary collaboration, and responsible development practices that ensure context engineering systems remain beneficial, reliable, and aligned with human values as they become increasingly integrated into critical societal functions [835, 1132, 310].

8. Conclusion

This survey has presented the first comprehensive examination of Context Engineering as a formal discipline that systematically designs, optimizes, and manages information payloads for LLMs. Through our analysis of over 1400 research papers, we have established Context Engineering as a critical foundation for developing sophisticated AI systems that effectively integrate external knowledge, maintain persistent memory, and interact dynamically with complex environments.

Our primary contribution lies in introducing a unified taxonomic framework that organizes context engineering techniques into **Foundational Components** (Context Retrieval and Generation, Context Processing, and Context Management) and **System Implementations** (Retrieval-Augmented Generation, Memory Systems, Tool-Integrated Reasoning, and Multi-Agent Systems). This framework demonstrates how core technical capabilities integrate into sophisticated architectures addressing real-world requirements.

Through this systematic examination, we have identified several key insights. First, we observe a fundamental asymmetry between LLMs' remarkable capabilities in understanding complex contexts and their limitations in generating equally sophisticated outputs. This comprehension-generation gap represents one of the most critical challenges facing the field. Second, our analysis reveals increasingly sophisticated integration patterns where multiple techniques combine synergistically, creating capabilities that exceed their individual components. Third, we observe a clear trend toward modularity and compositionality, enabling flexible architectures adaptable to diverse applications while maintaining system coherence. The evaluation challenges we identified underscore the need for comprehensive assessment frameworks that capture the complex, dynamic behaviors exhibited by context-engineered systems. Traditional evaluation methodologies prove insufficient for systems that integrate multiple components, exhibit adaptive behaviors, and operate across extended time horizons. Our examination of future research directions reveals significant opportunities including developing next-generation architectures for efficient long context handling, creating intelligent context assembly systems, and advancing multi-agent coordination mechanisms. Key challenges span theoretical foundations, technical implementation, and practical deployment, including the lack of unified theoretical frameworks, scaling limitations, and safety considerations.

Looking toward the future, Context Engineering stands poised to play an increasingly central role in AI development as the field moves toward complex, multi-component systems. The interdisciplinary nature of Context Engineering necessitates collaborative research approaches spanning computer science, cognitive science, linguistics, and domain-specific expertise.

As LLMs continue to evolve, the fundamental insight underlying Context Engineering—that AI system performance is fundamentally determined by contextual information—will remain central to artificial intelligence development. This survey provides both a comprehensive snapshot of the current state and a roadmap for future research, establishing Context Engineering as a distinct discipline with its own principles, methodologies, and challenges to foster innovation and support responsible development of context-aware AI systems.

上下文工程的未来将取决于我们通过持续、协作的研究努力来应对这些相互关联的挑战的能力，这些努力将技术创新与社会考量相结合。

成功将需要持续投资于基础研究、跨学科合作和负责任的开发实践，以确保上下文工程系统保持有益、可靠，并与人类价值观保持一致，随着它们越来越多地集成到关键的社会功能中[835, 1132, 310]。

8. Conclusion

本调查首次全面考察了上下文工程作为一个正式学科，它系统地设计、优化和管理用于大型语言模型的信息有效载荷。通过我们对1400多篇研究论文的分析，我们确立了上下文工程是开发能够有效整合外部知识、保持持久记忆并与复杂环境动态交互的复杂人工智能系统的关键基础。

我们的主要贡献在于引入了一个统一的分类框架，该框架将上下文工程技术组织为**基础组件**(上下文检索和生成、上下文处理和上下文管理)和**系统实现**(检索增强生成、记忆系统、工具集成推理和多智能体系统)。该框架展示了核心技术能力如何集成到处理现实世界需求的复杂架构中。

通过这项系统性的考察，我们识别出几个关键见解。首先，我们观察到大型语言模型（LLMs）在理解复杂上下文方面的卓越能力与其在生成同等复杂输出方面的局限性之间存在着根本性的不对称性。这种理解-生成差距代表了该领域面临的最关键挑战之一。其次，我们的分析揭示了一种日益复杂的集成模式，其中多种技术协同结合，创造出超越其各个组成部分的能力。第三，我们观察到一种明显的模块化和组合化趋势，这使得灵活的架构能够适应不同的应用，同时保持系统的一致性。我们识别出的评估挑战强调了需要全面的评估框架，这些框架能够捕捉上下文工程系统所展现的复杂、动态行为。传统的评估方法对于集成多个组件、表现出适应性行为并在长期时间范围内运行的系统来说是不够的。我们对未来研究方向的分析揭示了重大机遇，包括开发用于高效长上下文处理的下一代架构、创建智能上下文组装系统以及推进多智能体协调机制。关键挑战涵盖理论基础、技术实现和实践部署，包括缺乏统一的理论框架、扩展性限制以及安全性考量。

展望未来，随着该领域向复杂、多组件系统发展，上下文工程将在人工智能发展中扮演日益核心的角色。上下文工程的跨学科性质要求采用跨越计算机科学、认知科学、语言学和特定领域专业知识的协作研究方法。

随着大型语言模型（LLMs）的持续发展，上下文工程的基本见解——即人工智能系统性能从根本上由上下文信息决定——将始终是人工智能发展的核心。本调查既提供了当前状态的全面快照，也为未来研究提供了路线图，将上下文工程确立为一个拥有自身原则、方法和挑战的独立学科，以促进创新并支持上下文感知人工智能系统的负责任开发。

Acknowledgments

This survey represents an ongoing effort to comprehensively map the rapidly evolving landscape of Context Engineering for Large Language Models. Given the dynamic nature of this field, with new developments emerging continuously, we acknowledge that despite our best efforts, some recent works or emerging trends may have been inadvertently overlooked or underrepresented. We welcome feedback from the research community to help improve future iterations of this work. We are grateful to the broader research community whose foundational contributions have made this survey possible. This work would not have been achievable without the invaluable support of both the research community and the open-source community, whose collaborative efforts in developing frameworks, tools, and resources have significantly advanced the field of Context Engineering.

References

[1] Anp-agent communication meta-protocol specification(draft). https://agent-network-protocol.com/specs/communication.html. [Online; accessed 17-July-2025].

[2] S. A. Automating human evaluation of dialogue systems. *North American Chapter of the Association for Computational Linguistics*, 2022.

[3] Samir Abdaljalil, Hasan Kurban, Khalid A. Qaraqe, and E. Serpedin. Theorem-of-thought: A multi-agent framework for abductive, deductive, and inductive reasoning in language models. arXiv preprint, 2025.

[4] Abdelrahman Abdallah, Bhawna Piryani, Jamshid Mozafari, Mohammed Ali, and Adam Jatowt. Rankify: A comprehensive python toolkit for retrieval, re-ranking, and retrieval-augmented generation, arXiv preprint arXiv:2502.02464, 2025. URL https://arxiv.org/abs/2502.02464v3.

[5] Ibrahim Abdelaziz, Kinjal Basu, Mayank Agarwal, Sadhana Kumaravel, Matt Stallone, Rameswar Panda, Yara Rizk, G. Bhargav, M. Crouse, Chulaka Gunasekara, S. Ikbal, Sachin Joshi, Hima P. Karanam, Vineet Kumar, Asim Munawar, S. Neelam, Dinesh Raghu, Udit Sharma, Adriana Meza Soria, Dheeraj Sreedhar, P. Venkateswaran, Merve Unuvar, David Cox, S. Roukos, Luis A. Lastras, and P. Kapanipathi. Granite-function calling model: Introducing function calling abilities via multi-task learning of granular tasks. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[6] D. Acharya, Karthigeyan Kuppan, and Divya Bhaskaracharya. Agentic ai: Autonomous intelligence for complex goals—a comprehensive survey. *IEEE Access*, 2025.

[7] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. *AAAI Conference on Artificial Intelligence*, 2018.

[8] Shantanu Acharya, Fei Jia, and Boris Ginsburg. Star attention: Efficient llm inference over long sequences, arXiv preprint arXiv:2411.17116, 2024. URL https://arxiv.org/abs/2411.17116v3.

[9] Emre Can Acikgoz, Jeremy Greer, Akul Datta, Ze Yang, William Zeng, Oussama Elachqar, Emmanouil Koukoumidis, Dilek Hakkani-Tur, and Gokhan Tur. Can a single model master both multi-turn conversations and tool use? coalm: A unified conversational agentic language model, arXiv preprint arXiv:2502.08820, 2025. URL https://arxiv.org/abs/2502.08820v3.

致谢

这项调查代表了一项持续的努力，旨在全面梳理大型语言模型上下文工程快速发展的领域。鉴于该领域的动态特性，新进展不断涌现，我们承认尽管我们已尽力，但仍可能无意中遗漏或未能充分代表某些近期工作或新兴趋势。我们欢迎研究社区提供反馈，以帮助改进这项工作的未来版本。我们感谢更广泛的研究社区，其基础性贡献使这项调查成为可能。没有研究社区和开源社区无价的支持，这项工作将无法实现，他们的协作努力在开发框架、工具和资源方面显著推动了上下文工程领域的发展。

参考文献

[1]Anp-agent 通信元协议规范(草案)。https://agent-network-protocol.com/specs/communication.html. [在线；访问于 17- 7-2025]。

[2] S. A. 自动化对话系统的评估。 北美洲计算语言学协会分会, 2022。

[3] Samir Abdaljalil, Hasan Kurban, Khalid A. Qaraqe, 和 E. Serpedin. 思维定理：一种用于语言模型中溯因、演绎和归纳推理的多智能体框架。arXiv 预印本，2025年。

[4] Abdelrahman Abdallah, Bhawna Piryani, Jamshid Mozafari, Mohammed Ali, 和 Adam Jatowt. Rankify：一个用于检索、重排序和检索增强生成的全面 Python 工具包，arXiv 预印本 arXiv:2502.02464, 2025年。URLhttps://arxiv.org/abs/2502.02464v3.

[5] Ibrahim Abdelaziz, Kinjal Basu, Mayank Agarwal, Sadhana Kumaravel, Matt Stallone, Rameswar Panda, Yara Rizk, G. Bhargav, M. Crouse, Chulaka Gunasekara, S. Ikbal, Sachin Joshi, Hima P. Karanam, Vineet Kumar, Asim Munawar, S. Neelam, Dinesh Raghu, Udit Sharma, Adriana Meza Soria, Dheeraj Sreedhar, P. Venkateswaran, Merve Unuvar, David Cox, S. Roukos, Luis A. Lastras, 和 P. Kapanipathi. Granite-function 调用模型：通过粒度任务的 多任务学习引入函数调用能力。自然语言处理经验方法会议, 2024年。

[6] D. Acharya、Karthigeyan Kuppan 和 Divya Bhaskaracharya。Agentic ai：用于复杂目标的自主智能——一项综合调查。*IEEE Access*, 2025。

[7] Manoj Acharya、Kushal Kafle 和 Christopher Kanan。Tallyqa：回答复杂计数 g 问题。 *AAAI* 人工智能会议, 2018。

[8] Shantanu Acharya、Fei Jia 和 Boris Ginsburg。Star attention：高效 LLM 推理在 r 长序列上，arXiv 预印本 arXiv:2411.17116, 2024。URL https://arxiv.org/abs/2411 . 17116v3。

[9] Emre Can Acikgoz, Jeremy Greer, Akul Datta, Ze Yang, William Zeng, Oussama Elachqar, Emmanouil Koukoumidis, Dilek Hakkani-Tur, and Gokhan Tur. Can a single model master both multi-turn conversations and tool use? coalm: A unified conversational agentic language model, arXiv preprint arXiv:2502.08820, 2025. URL https://arxiv.org/abs/2502.08820v3.

[10] Emre Can Acikgoz, Cheng Qian, Hongru Wang, Vardhan Dongre, Xiusi Chen, Heng Ji, Dilek Hakkani-Tur, and Gokhan Tur. A desideratum for conversational agents: Capabilities, challenges, and future directions, arXiv preprint arXiv:2504.16939, 2025. URL https://arxiv.org/abs/2504.16939v1.

[11] Anum Afzal, Juraj Vladika, Gentrit Fazlija, Andrei Staradubets, and Florian Matthes. Towards optimizing a retrieval augmented generation using large language model on academic data. *International Conference on Natural Language Processing and Information Retrieval*, 2024.

[12] Ankush Agarwal, Sakharam Gawade, A. Azad, and P. Bhattacharyya. Kitlm: Domain-specific knowledge integration into language models for question answering. *ICON*, 2023.

[13] Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. Large scale knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. arXiv preprint, 2020.

[14] Monica Agrawal, S. Hegselmann, Hunter Lang, Yoon Kim, and D. Sontag. Large language models are few-shot clinical information extractors. *Conference on Empirical Methods in Natural Language Processing*, 2022.

[15] Arash Ahmadi, S. Sharif, and Yaser Mohammadi Banadaki. Mcp bridge: A lightweight, llm-agnostic restful proxy for model context protocol servers, arXiv preprint arXiv:2504.08999, 2025. URL https://arxiv.org/abs/2504.08999v1.

[16] J. Ainslie, J. Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebr'on, and Sumit K. Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *Conference on Empirical Methods in Natural Language Processing*, 2023.

[17] Adel Al-Jumaily. Multi-agent system concepts theory and application phases. arXiv preprint, 2006.

[18] Faisal Al-Khateeb, Nolan Dey, Daria Soboleva, and Joel Hestness. Position interpolation improves alibi extrapolation. arXiv preprint, 2023.

[19] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, A. Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, O. Vinyals, Andrew Zisserman, and K. Simonyan. Flamingo: a visual language model for few-shot learning. *Neural Information Processing Systems*, 2022.

[20] Stefano V. Albrecht and P. Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 2017.

[21] Buthayna AlMulla, Maram Assi, and Safwat Hassan. Understanding the challenges and promises of developing generative ai apps: An empirical study, arXiv preprint arXiv:2506.16453, 2025. URL https://arxiv.org/abs/2506.16453v2.

[22] Reem S. Alsuhaibani, Christian D. Newman, M. J. Decker, Michael L. Collard, and Jonathan I. Maletic. On the naming of methods: A survey of professional developers. *International Conference on Software Engineering*, 2021.

[10] Emre Can Acikgoz, Cheng Qian, Hongru Wang, Vardhan Dongre, Xiusi Chen, Heng Ji, Dilek Hakkani-Tur, and Gokhan Tur. 对话代理的期望：能力、挑战和未来方向，arXiv 预印本 arXiv:2504.16939，2025。URLhttps://arxiv.org/abs/2504.16939v1.

[11] Anum Afzal, Juraj Vladika, Gentrit Fazlija, Andrei Staradubets 和 Florian Matthes. 面向使用大型语言模型在学术数据上优化检索增强生成. 国际自然语言处理与信息检索会议, 2024.

[12] Ankush Agarwal、Sakharam Gawade、A. Azad 和 P. Bhattacharyya。Kitlm：将特定领域知识集成到语言模型中以进行问答。*ICON*, 2023。

[13] Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 基于大规模知识图谱的合成语料库生成用于知识增强语言模型预训练. arXiv preprint, 2020.

[14] Monica Agrawal, S. Hegselmann, Hunter Lang, Yoon Kim, and D. Sontag. 大型语言模型是少样本临床信息提取器. 自然语言处理经验方法会议, 2022.

[15] Arash Ahmadi, S. Sharif, and Yaser Mohammadi Banadaki. Mcp bridge: 一种轻量级、与llm无关的模型上下文协议服务器代理, arXiv preprint arXiv:2504.08999, 2025. URLhttps://arxiv.org/abs/2504.08999v1.

[16] J. Ainslie, J. Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebr'on, and Sumit K. Sanghai. Gqa: 从多头检查点训练通用多查询Transformer模型. 自然语言处理经验方法会议, 2023.

[17] Adel Al-Jumaily. 多智能体系统概念、理论和应用阶段。arXiv 预印本，2006年。

[18] Faisal Al-Khateeb, Nolan Dey, Daria Soboleva, 和 Joel Hestness. 位置插值改进 alibi 外推。arXiv 预印本, 2023年。

[19] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, A. Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, O. Vinyals, Andrew Zisserman, 和 K. Simonyan. Flamingo：一种用于少样本学习的视觉语言模型。神经信息处理系统，2022年。

[20] Stefano V. Albrecht 和 P. Stone. 自主代理模拟其他代理：一项综合调查和开放问题. 人工智能, 2017.

[21] Buthayna AlMulla, Maram Assi, 和 Safwat Hassan. 理解开发生成式 AI 应用的挑战和机遇：一项实证研究，arXiv 预印本 arXiv:2506.16453，2025. URLhttps://arxiv.org/abs/2506.16453v2.

[22] Reem S. Alsuhaibani, Christian D. Newman, M.J. Decker, Michael L. Collard, 和 Jonathan I. Maletic. 关于方法命名：专业开发者的调查. 软件工程国际会议, 2021.

[23] Francesco Alzetta, P. Giorgini, A. Najjar, M. Schumacher, and Davide Calvaresi. In-time explainability in multi-agent systems: Challenges, opportunities, and roadmap. *EXTRAAMAS@AAMAS*, 2020.

[24] Kenza Amara, Lukas Klein, Carsten T. Lüth, Paul F. Jäger, Hendrik Strobelt, and Mennatallah El-Assady. Why context matters in vqa and reasoning: Semantic interventions for vlm input modalities, arXiv preprint arXiv:2410.01690v1, 2024. URL https://arxiv.org/abs/2410.01690v1.

[25] Xavier Amatriain. Prompt design and engineering: Introduction and advanced methods, arXiv preprint arXiv:2401.14423, 2024. URL https://arxiv.org/abs/2401.14423v4.

[26] Zahra Aminiranjbar, Jianan Tang, Qiudan Wang, Shubha Pant, and Mahesh Viswanathan. Dawn: Designing distributed agents in a worldwide network, arXiv preprint arXiv:2410.22339, 2024. URL https://arxiv.org/abs/2410.22339v3.

[27] Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. Why does the effective context length of llms fall short? *International Conference on Learning Representations*, 2024.

[28] Kaikai An, Fangkai Yang, Liqun Li, Junting Lu, Sitao Cheng, Shuzheng Si, Lu Wang, Pu Zhao, Lele Cao, Qingwei Lin, et al. Thread: A logic-based data organization paradigm for how-to question answering with retrieval augmented generation. *arXiv preprint arXiv:2406.13372*, 2024.

[29] Kaikai An, Fangkai Yang, Junting Lu, Liqun Li, Zhixing Ren, Hao Huang, Lu Wang, Pu Zhao, Yu Kang, Hua Ding, et al. Nissist: An incident mitigation copilot based on troubleshooting guides. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI 2024)*, pages 4471–4474, 2024.

[30] Kaikai An, Li Sheng, Ganqu Cui, Shuzheng Si, Ning Ding, Yu Cheng, and Baobao Chang. Ultraif: Advancing instruction following from the wild. pages 7930–7957, 2025.

[31] Sumin An, Junyoung Sung, Wonpyo Park, Chanjun Park, and Paul Hongsuck Seo. Lcirc: A recurrent compression approach for efficient long-form context and query dependent modeling in llms. *North American Chapter of the Association for Computational Linguistics*, 2025.

[32] Sotiris Anagnostidis, Dario Pavllo, Luca Biggio, Lorenzo Noci, Aurélien Lucchi, and Thomas Hofmann. Dynamic context pruning for efficient and interpretable autoregressive transformers. *Neural Information Processing Systems*, 2023.

[33] John R. Anderson, M. Matessa, and C. Lebiere. Act-r: A theory of higher level cognition and its relation to visual attention. *Hum. Comput. Interact.*, 1997.

[34] Jacob Andreas. Language models as agent models. *Conference on Empirical Methods in Natural Language Processing*, 2022.

[35] Leonardo Aniello, R. Baldoni, and Leonardo Querzoni. Adaptive online scheduling in storm. *Distributed Event-Based Systems*, 2013.

[36] Petr Anokhin, Nikita Semenov, Artyom Sorokin, Dmitry Evseev, M. Burtsev, and Evgeny Burnaev. Arigraph: Learning knowledge graph world models with episodic memory for llm agents, arXiv preprint arXiv:2407.04363, 2024. URL https://arxiv.org/abs/2407.04363v3.

[23] Francesco Alzetta, P. Giorgini, A. Najjar, M.Schumacher, and Davide Calvaresi. 基于时序的可解释性在多智能体系统中的挑战、机遇和路线图。 *EXTRAAMAS@AAMAS*, 2020。

[24] Kenza Amara, Lukas Klein, Carsten T. Lüth, Paul F. Jäger, Hendrik Strobelt, and Mennatallah El-Assady. 为什么上下文在vqa和推理中很重要：为vlm输入模态的语义干预，arXiv预印本 arXiv:2410.01690v1，2024。URLhttps://arxiv.org/abs/2410.01690v1.

[25] Xavier Amatriain. 提示设计和工程：介绍和高级方法，arXiv预印本 arXiv:2401.14423，2024。URLhttps://arxiv.org/abs/2401.14423v4.

[26] 扎赫拉·阿明拉尼贾尔，唐建安，王秋丹，潘舒巴，和马赫什·维什瓦纳坦。Dawn：在世界范围内网络中设计分布式代理，arXiv预印本arXiv:2410.22339，2024年。URLhttps://arxiv.org/abs/2410.22339v3。

陈欣，张俊，钟明，李雷， Gong Shansan，罗瑶，徐晶晶，孔令鹏. 为什么LLM的有效上下文长度不足？学习表征国际会议, 2024.

[28] 凯凯·安，方凯·杨，李立群，陆俊庭，程思涛，司舒正，王路，赵普，曹乐乐，林清伟，等. Thread: 一种基于逻辑的数据组织范式，用于检索增强生成式问答. *arXiv*预印本 *arXiv:2406.13372*, 2024.

[29] 开开·安，方开·杨，陆俊庭，李立群，任志兴，黄浩，王露，赵普，康宇，丁华，等. Nissist: 基于故障排除指南的事故缓解副驾驶. 在第27届欧洲人工智能会议论文集 *(ECAI 2024)*, 第4471–4474页, 2024.

[30] 开开安、盛李、崔甘泉、思舒正、丁宁、程宇和长宝宝。Ultraif：从野外推进指令跟随。第7930-7957页，2025年。

[31] Sumin An, Junyoung Sung, Wonpyo Park, Chanjun Park 和 Paul Hongsuck Seo. Lcirc: 一种用于高效长文本上下文和查询依赖建模的循环压缩方法，在大型语言模型中。北美计算语言学协会分会, 2025.

[32] Sotiris Anagnostidis, Dario Pavllo, Luca Biggio, Lorenzo Noci, Aurélien Lucchi, and Thomas Hofmann. 动态上下文剪枝：高效且可解释的自回归Transformer。 *NeuralInformationProcessing Systems*, 2023。

[33] John R. Anderson, M. Matessa, and C. Lebiere. Act-r: 一种高级认知理论及其与视觉注意力的关系。 *Hum. Comput. Interact.*, 1997。

[34] Jacob Andreas. 语言模型作为代理模型。 自然语言处理经验方法会议, 2022。

[35] Leonardo Aniello, R. Baldoni, and Leonardo Querzoni. Storm中的自适应在线调度。 分布式事件系统, 2013。

[36] Petr Anokhin, Nikita Semenov, Artyom Sorokin, Dmitry Evseev, M. Burtsev, and Evgeny Burnaev. Arigraph: 为LLM代理学习具有情景记忆的知识图谱世界模型，arXiv预印本 arXiv:2407.04363, 2024。URLhttps://arxiv.org/abs/2407.04363v3.

[37] Anthropic. Introducing the model context protocol, November 2024. URL https://www.anthropic.com/news/model-context-protocol. [Online; accessed 17-July-2025].

[38] RM Aratchige and Dr. Wmks Ilmini. Llms working in harmony: A survey on the technological aspects of building effective llm-based multi agent systems, arXiv preprint arXiv:2504.01963, 2025. URL https://arxiv.org/abs/2504.01963v1.

[39] Leo Ardon, Daniel Furelos-Blanco, and A. Russo. Learning reward machines in cooperative multi-agent tasks. *AAMAS Workshops*, 2023.

[40] K. Armeni, C. Honey, and Tal Linzen. Characterizing verbatim short-term memory in neural language models. *Conference on Computational Natural Language Learning*, 2022.

[41] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *International Conference on Learning Representations*, 2023.

[42] Hikaru Asano, Tadashi Kozuno, and Yukino Baba. Self iterative label refinement via robust unlabeled learning, arXiv preprint arXiv:2502.12565, 2025. URL https://arxiv.org/abs/2502.12565v1.

[43] Ben Athiwaratkun, Sujan Kumar Gonugondla, Sanjay Krishna Gouda, Haifeng Qian, Hantian Ding, Qing Sun, Jun Wang, Jiacheng Guo, Liangfu Chen, Parminder Bhatia, Ramesh Nallapati, Sudipta Sengupta, and Bing Xiang. Bifurcated attention: Accelerating massively parallel decoding with shared prefixes in llms, arXiv preprint arXiv:2403.08845, 2024. URL https://arxiv.org/abs/2403.08845v2.

[44] Avinash Ayalasomayajula, Rui Guo, Jingbo Zhou, Sujan Kumar Saha, and Farimah Farahmandi. Lasp: Llm assisted security property generation for soc verification. *Workshop on Machine Learning for CAD*, 2024.

[45] Simon A. Aytes, Jinheon Baek, and Sung Ju Hwang. Sketch-of-thought: Efficient llm reasoning with adaptive cognitive-inspired sketching. arXiv preprint, 2025.

[46] Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, A. Kazerouni, I. Rekik, and D. Merhof. Foundational models in medical imaging: A comprehensive survey and future vision, arXiv preprint arXiv:2310.18689, 2023. URL https://arxiv.org/abs/2310.18689v1.

[47] Gilbert Badaro, Mohammed Saeed, and Paolo Papotti. Transformers for tabular data representation: A survey of models and applications. *Transactions of the Association for Computational Linguistics*, 2023.

[48] Jinheon Baek, N. Chandrasekaran, Silviu Cucerzan, Allen Herring, and S. Jauhar. Knowledge-augmented large language models for personalized contextual query suggestion. *The Web Conference*, 2023.

[49] Tianyi Bai, Hao Liang, Binwang Wan, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, Conghui He, Binhang Yuan, and Wentao Zhang. A survey of multimodal large language model from a data-centric perspective, arXiv preprint arXiv:2405.16640v2, 2024. URL https://arxiv.org/abs/2405.16640v2.

[37] Anthropic. Introducing the model context protocol, November 2024. URL https://www.anthropic.com/news/model-context-protocol. [Online; accessed 17-July-2025].

[38] RM Aratchige 和 Dr. Wmks Ilmini. 协同工作的 LLM：关于构建有效基于 LLM 的多智能体系统的技术方面的调查，arXiv 预印本 arXiv:2504.01963，2025。URLhttps://arxiv.org/abs/2504.01963v1.

[39] Leo Ardon、Daniel Furelos-Blanco 和 A. Russo。在合作多智能体任务中学习奖励机。*AAMASWorkshops*，2023.

[40] K.Armeni、C.Honey和Tal Linzen。神经语言模型中的短时记忆特性。计算自然语言学习会议，2022。

[41] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, 和 Hannaneh Hajishirzi. Self-rag: 通过自我反思学习检索、生成和评论. 国际学习表征会议, 2023.

[42] Hikaru Asano, Tadashi Kozuno, and Yukino Baba. 自我迭代标签优化通过鲁棒无标签学习，arXiv 预印本 arXiv:2502.12565, 2025. URL https://arxiv.org/abs/2502.12565v1.

[43] Ben Athiwaratkun, Sujan Kumar Gonugondla, Sanjay Krishna Gouda, Haifeng Qian, Hantian Ding, Qing Sun, Jun Wang, Jiacheng Guo, Liangfu Chen, Parminder Bhatia, Ramesh Nallapati, Sudipta Sengupta, and Bing Xiang. 分叉注意力：通过共享前缀加速大规模并行解码在 llms 中，arXiv 预印本 arXiv:2403.08845, 2024. URL https://arxiv.org/abs/2403.08845v2.

[44] Avinash Ayalasomayajula, Rui Guo, JingboZhou, Sujan Kumar Saha, and Farimah Farahmandi. Lasp: 通过 llm 辅助生成用于 soc 验证的安全属性。 机器学习 *CAD* 工作坊, 2024。

[45] Simon A. Aytes, Jinheon Baek, 和 Sung Ju Hwang. 思维草图：基于自适应认知启发式草图的效率llm推理。arXiv预印本，2025年。

[46] Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, A. Kazerouni, I. Rekik, 和 D. Merhof. 医学影像中的基础模型：一项综合调查和未来展望，arXiv预印本 arXiv:2310.18689，2023年。URLhttps://arxiv.org/abs/2310.18689v1.

[47] Gilbert Badaro, Mohammed Saeed, 和 Paolo Papotti. 用于表格数据表示的Transformer：模型和应用调查。 计算语言学协会汇刊，2023年。

[48] Jinheon Baek, N. Chandrasekaran, Silviu Cucerzan, Allen Herring, 和 S. Jauhar. 知识增强型大型语言模型用于个性化上下文查询建议。 网络会议，2023年。

[49] 白天一，梁浩，万斌旺，杨凌，李伯舟，王奕帆，崔斌，何聪辉，袁彬航，张文涛. 基于数据中心的跨模态大语言模型综述，arXiv预印本arXiv:2405.16640v2，2024.
URLhttps://arxiv.org/abs/2405.16640v2.

[50] Yu Bai, Xiyuan Zou, Heyan Huang, Sanxing Chen, Marc-Antoine Rondeau, Yang Gao, and Jackie Chi Kit Cheung. Citrus: Chunked instruction-aware state eviction for long sequence modeling. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[51] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, arXiv preprint arXiv:2212.08073, 2022. URL https://arxiv.org/abs/2212.08073.

[52] Souhail Bakkali, Sanket Biswas, Zuheng Ming, Mickaël Coustaty, Marccal Rusinol, O. R. Terrades, and Josep Llad'os. Globaldoc: A cross-modal vision-language framework for real-world document image retrieval and classification. *IEEE Workshop/Winter Conference on Applications of Computer Vision*, 2023.

[53] Jayachandu Bandlamudi, K. Mukherjee, Prerna Agarwal, Sampath Dechu, Siyu Huo, Vatche Isahagian, Vinod Muthusamy, N. Purushothaman, and Renuka Sindhgatta. Towards hybrid automation by bootstrapping conversational interfaces for it operation tasks. *AAAI Conference on Artificial Intelligence*, 2023.

[54] Jayachandu Bandlamudi, Kushal Mukherjee, Prerna Agarwal, Ritwik Chaudhuri, R. Pimplikar, Sampath Dechu, Alex Straley, Anbumunee Ponniah, and Renuka Sindhgatta. Building conversational artifacts to enable digital assistant for apis and rpas. *AAAI Conference on Artificial Intelligence*, 2024.

[55] Keqin Bao, Jizhi Zhang, Xinyu Lin, Yang Zhang, Wenjie Wang, and Fuli Feng. Large language models for recommendation: Past, present, and future. *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

[56] Sara Di Bartolomeo, Giorgio Severi, V. Schetinger, and Cody Dunne. Ask and you shall receive (a graph drawing): Testing chatgpt's potential to apply graph layout algorithms. *Eurographics Conference on Visualization*, 2023.

[57] Saikat Barua. Exploring autonomous agents through the lens of large language models: A review, arXiv preprint arXiv:2404.04442, 2024. URL https://arxiv.org/abs/2404.04442v1.

[58] Kinjal Basu, Ibrahim Abdelaziz, Kelsey Bradford, M. Crouse, Kiran Kate, Sadhana Kumaravel, Saurabh Goyal, Asim Munawar, Yara Rizk, Xin Wang, Luis A. Lastras, and P. Kapanipathi. Nestful: A benchmark for evaluating llms on nested sequences of api calls, arXiv preprint arXiv:2409.03797, 2024. URL https://arxiv.org/abs/2409.03797v3.

[59] Amin Beheshti. Natural language-oriented programming (nlop): Towards democratizing software creation. *2024 IEEE International Conference on Software Services Engineering (SSE)*, 2024.

[60] Azadeh Beiranvand and S. M. Vahidipour. Integrating structural and semantic signals in text-attributed graphs with bigtex, arXiv preprint arXiv:2504.12474, 2025. URL https://arxiv.org/abs/2504.12474v2.

[50] 余白, 邹西原, 黄鹤岩, 陈三兴, Marc-Antoine Rondeau, 高杨, 和 Jackie Chi Kit Cheung. Citrus: 长序列建模的指令感知状态驱逐分块方法.自然语言处理经验方法会议, 2024.

[51] 白云涛, Kadavath Saurav, Kundu Sandipan, Askell Amanda, Kernion Jackson, Jones Andy, Chen Anna, Goldie Anna, Mirhoseini Azalia, McKinnon Cameron, Chen Carol, Olsson Catherine, Olah Christopher, Hernandez Danny, Drain Dawn, Ganguli Deep, Li Dustin, Tran-Johnson Eli, Perez Ethan, Kerr Jamie, Mueller Jared, Ladish Jeffrey, Landau Joshua, Ndousse Kamal, Lukosuite Kamile, Lovitt Liane, Sellitto Michael, Elhage Nelson, Schiefer Nicholas, Mercado Noemi, DasSarma Nova, Lasenby Robert, Larson Robin, Ringer Sam, Johnston Scott, Kravec Shauna, Showk Sheer El, Fort Stanislav, Lanham Tamera, Telleen-Lawton Timothy, Conerly Tom, Henighan Tom, Hume Tristan, Bowman Samuel R., Hatfield-Dodds Zac, Mann Ben, Amodei Dario, Joseph Nicholas, McCandlish Sam, Brown Tom, 和 Kaplan Jared. 宪政式人工智能：基于人工智能反馈的无害性, arXiv 预印本 arXiv:2212.08073, 2022. URLhttps://arxiv.org/abs/2212.08073.

[52] Souhail Bakkali, Sanket Biswas, Zuheng Ming, Mickaël Coustaty, Marccal Rusinol, O. R. Terrades, 和 Josep Llad'os. Globaldoc: 一个跨模态视觉语言框架，用于现实世界文档图像检索和分类。IEEE Workshop/Winter Conference on Applications of Computer Vision, 2023.

[53] Jayachandu Bandlamudi、K.Mukherjee、Prerna Agarwal、SampathDechu、SiyuHuo、Vatche Isahagian、Vinod Muthusamy、N. Purushothaman 和 Renuka Sindhgatta。通过引导式对话界面实现 IT 操作任务的混合自动化。AAAI 人工智能会议，2023。

[54] Jayachandu Bandlamudi、Kushal Mukherjee、Prerna Agarwal、Ritwik Chaudhuri、R. Pimplikar、Sampath Dechu、Alex Straley、Anbumunee Ponniah和Renuka Sindhgatta。构建对话式工具以支持API和RPA的数字助手。*AAAI*人工智能会议, 2024。

[55] 鲍科勤, 张继志, 林新宇, 张阳, 王文杰, 冯福丽. 推荐系统中的大型语言模型：过去、现在与未来. 年度国际*ACM SIGIR*信息检索研究与发展会议, 2024.

[56] Sara Di Bartolomeo, Giorgio Severi, V. Schetinger, and Cody Dunne. Ask and you shall receive (a graph drawing): 测试ChatGPT应用图布局算法的潜力. *Eurographics*可视化会议, 2023.

[57] Saikat Barua. 通过大型语言模型视角探索自主代理：综述, arXiv预印本 arXiv:2404.04442, 2024. URLhttps://arxiv.org/abs/2404.04442v1.

[58] Kinjal Basu, Ibrahim Abdelaziz, Kelsey Bradford, M. Crouse, Kiran Kate, Sadhana Kumaravel, Saurabh Goyal, Asim Munawar, Yara Rizk, Xin Wang, Luis A. Lastras, and P. Kapanipathi. Nestful: 一个用于评估LLM在API调用嵌套序列上的基准, arXiv预印本 arXiv:2409.03797, 2024. URLhttps://arxiv.org/abs/2409.03797v3.

[59] Amin Beheshti. 面向自然语言的编程 (nlop)：迈向软件创作的民主化。 *2024 IEEE International ConferenceonSoftware Services Engineering (SSE)*, 2024。

[60] Azadeh Beiranvand and S. M. Vahidipour. 在文本属性图 bigtex 中整合结构和语义信号, arXiv preprint arXiv:2504.12474, 2025。URL https://arxiv.org/abs/2504.12474v2.

[61] Assaf Ben-Kish, Itamar Zimerman, Shady Abu-Hussein, Nadav Cohen, Amir Globerson, Lior Wolf, and Raja Giryes. Decimamba: Exploring the length extrapolation potential of mamba. *International Conference on Learning Representations*, 2024.

[62] Assaf Ben-Kish, Itamar Zimerman, M. J. Mirza, James R. Glass, Leonid Karlinsky, and Raja Giryes. Overflow prevention enhances long-context recurrent llms. arXiv preprint, 2025.

[63] M. Benna and Stefano Fusi. Complex synapses as efficient memory systems. *BMC Neuroscience*, 2015.

[64] M. Benna and Stefano Fusi. Computational principles of biological memory, arXiv preprint arXiv:1507.07580, 2015. URL https://arxiv.org/abs/1507.07580v1.

[65] Shelly Bensal, Umar Jamil, Christopher Bryant, M. Russak, Kiran Kamble, Dmytro Mozolevskyi, Muayad Ali, and Waseem Alshikh. Reflect, retry, reward: Self-improving llms via reinforcement learning, arXiv preprint arXiv:2505.24726, 2025. URL https://arxiv.org/abs/2505.24726v1.

[66] Idoia Berges, J. Bermúdez, A. Goñi, and A. Illarramendi. Semantic web technology for agent communication protocols. *Extended Semantic Web Conference*, 2008.

[67] Gaurav Beri and Vaishnavi Srivastava. Advanced techniques in prompt engineering for large language models: A comprehensive study. *2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG)*, 2024.

[68] Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew R. Gormley. Unlimiformer: Long-range transformers with unlimited length input. *Neural Information Processing Systems*, 2023.

[69] Maciej Besta, Nils Blach, Aleš Kubíček, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, H. Niewiadomski, P. Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *AAAI Conference on Artificial Intelligence*, 2023.

[70] Gregor Betz and Kyle Richardson. Judgment aggregation, discursive dilemma and reflective equilibrium: Neural language models as self-improving doxastic agents. *Frontiers in Artificial Intelligence*, 2022.

[71] L. Bezalel, Eyal Orgad, and Amir Globerson. Teaching models to improve on tape. *AAAI Conference on Artificial Intelligence*, 2024.

[72] Umang Bhatt, Sanyam Kapoor, Mihir Upadhyay, Ilia Sucholutsky, Francesco Quinzan, Katherine M. Collins, Adrian Weller, Andrew Gordon Wilson, and Muhammad Bilal Zafar. When should we orchestrate multiple agents?, arXiv preprint arXiv:2503.13577, 2025. URL https://arxiv.org/abs/2503.13577v1.

[73] Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, et al. Context-dpo: Aligning language models for context-faithfulness. *ACL 2025*, 2024.

[74] Baolong Bi, Shenghua Liu, Lingrui Mei, Yiwei Wang, Pengliang Ji, and Xueqi Cheng. Decoding by contrasting knowledge: Enhancing llms' confidence on edited facts. *ACL 2025*, 2024.

[61] Assaf Ben-Kish, Itamar Zimerman, Shady Abu-Hussein, Nadav Cohen, Amir Globerson, Lior Wolf, and Raja Giryes. Decimamba: 探索 mamba 的长度外推潜力。 *InternationalConferenceon Learning Representations*, 2024。

[62] Assaf Ben-Kish, Itamar Zimerman, M. J. Mirza, James R. Glass, Leonid Karlinsky, and Raja Giryes. 防止溢出增强了长上下文循环 llms。 arXiv preprint, 2025。

[63] M. Benna and Stefano Fusi. 复杂突触作为高效的记忆系统。 *BMCNeuroscience*, 2015。

[64] M. Benna and Stefano Fusi. 生物记忆的计算原理，arXiv preprin arXiv:1507.07580, 2015。 t URLhttps://arxiv.org/abs/1507.07580v1.

[65] Shelly Bensal、Umar Jamil、Christopher Bryant、M. Russak、Kiran Kamble、Dmytro Mozolevskyi、Muayad Ali 和 Waseem Alshikh。反思、重试、奖励：通过强化学习实现自我改进的大型语言模型，arXiv 预印本 arXiv:2505.24726，2025 年。URLhttps://arxiv.org/abs/2505.24726v1.

[66] Idoia Berges, J. Bermúdez, A. Goñi, and A. Illarramendi. 语义网技术用于智能体通信协议。 扩展语义网会议，2008。

[67] Gaurav Beri和Vaishnavi Srivastava. 大型语言模型的提示工程高级技术：一项综合研究. *2024 IEEE第4届国际商业、工业与政府信息通信技术会议 (ICTBIG)*, 2024.

[68] Amanda Bertsch, Uri Alon, Graham Neubig, 和 Matthew R. Gormley. Unlimiformer: 带无限长度输入的长程 Transformer。 *NeuralInformationProcessingSystems*, 2023。

[69] Maciej Besta, Nils Blach, Aleš Kubíček, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, H. Niewiadomski, P. Nyczyk, 和 Torsten Hoefler. Graph of thoughts: 用大型语言模型解决复杂问题。 *AAAI Conference onArtificialIntelligence*, 2023。

[70] Gregor Betz 和 Kyle Richardson. Judgment aggregation, discursive dilemma 和 reflective equilibrium: 作为自我改进的断言代理的神经语言模型。 *Frontiers inArtificialIntelligence*, 2022。

[71] L. Bezalel, Eyal Orgad, 和 Amir Globerson. Teaching models to improve on tape。 *AAAI Conferenceon ArtificialIntelligence*, 2024。

[72] Umang Bhatt, Sanyam Kapoor, Mihir Upadhyay, Ilia Sucholutsky, Francesco Quinzan, Katherine M. Collins, Adrian Weller, Andrew Gordon Wilson, and Muhammad Bilal Zafar. 我们应该在什么时候协调多个智能体？ , arXiv preprint arXiv:2503.13577, 2025. URL https://arxiv.org/abs/2503.13577v1.

[73] 鲍龙飞, 黄少航, 王一伟, 杨天池, 张子涵, 黄海珍, 梅凌瑞, 方俊峰, 李泽豪, 魏福如, 等. Context-dpo: 对齐语言模型以实现上下文忠实度. *ACL2025*, 2024.

[74] 鲍龙飞, 刘胜华, 梅凌瑞, 王艺伟, 季鹏亮, 和程雪琪. 对比知识解码：增强编辑事实下大语言模型的置信度。*ACL 2025*, 2024.

[75] Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, and Xueqi Cheng. Lpnl: Scalable link prediction with large language models. *ACL 2024*, 2024.

[76] Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Hongcheng Gao, Junfeng Fang, and Xueqi Cheng. Struedit: Structured outputs enable the fast and accurate knowledge editing for large language models. 2024.

[77] Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Hongcheng Gao, Yilong Xu, and Xueqi Cheng. Adaptive token biaser: Knowledge editing via biasing key entities. *EMNLP 2024*, 2024.

[78] Baolong Bi, Shenghua Liu, Xingzhang Ren, Dayiheng Liu, Junyang Lin, Yiwei Wang, Lingrui Mei, Junfeng Fang, Jiafeng Guo, and Xueqi Cheng. Refinex: Learning to refine pre-training data at scale from expert-guided programs. 2025.

[79] Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Junfeng Fang, Hongcheng Gao, Shiyu Ni, and Xueqi Cheng. Is factuality enhancement a free lunch for llms? better factuality can lead to worse context-faithfulness. *ICLR 2025*, 2025.

[80] Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. Parameters vs. context: Fine-grained control of knowledge reliance in language models. 2025.

[81] Bin Bi, Chenliang Li, Chen Wu, Ming Yan, and Wei Wang. Palm: Pre-training an autoencoding&autoregressive language model for context-conditioned generation. *Conference on Empirical Methods in Natural Language Processing*, 2020.

[82] Dinh Doan Van Bien, David Lillis, and Rem W. Collier. Call graph profiling for multi agent systems. *Multi-Agent Logics, Languages, and Organisations Federated Workshops*, 2009.

[83] Jonas Bode, Bastian Pätzold, Raphael Memmesheimer, and Sven Behnke. A comparison of prompt engineering techniques for task planning and execution in service robotics. *IEEE-RAS International Conference on Humanoid Robots*, 2024.

[84] P. Bonzon. Grounding mental representations in a virtual multi-level functional framework. *Journal of Cognition*, 2023.

[85] Sebastian Borgeaud, A. Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, T. Hennigan, Saffron Huang, Lorenzo Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, G. Irving, O. Vinyals, Simon Osindero, K. Simonyan, Jack W. Rae, Erich Elsen, and L. Sifre. Improving language models by retrieving from trillions of tokens. *International Conference on Machine Learning*, 2021.

[86] Zalán Borsos, Raphaël Marinier, Damien Vincent, E. Kharitonov, O. Pietquin, Matthew Sharifi, Dominik Roblek, O. Teboul, David Grangier, M. Tagliasacchi, and Neil Zeghidour. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2022.

[87] FutureSearch Nikos I. Bosse, Jon Evans, Robert G. Gambee, Daniel Hnyk, Peter Muhlbacher, Lawrence Phillips, Dan Schwarz, and Jack Wildman. Deep research bench: Evaluating ai web research agents, arXiv preprint arXiv:2506.06287, 2025. URL https://arxiv.org/abs/2506.06287v1.

[75] 鲍龙飞, 刘胜华, 王怡伟, 梅凌瑞, 和程学祺. Lpnl: 基于大型语言模型的可扩展链接预测. *ACL 2024*, 2024.

[76] 鲍龙飞, 刘胜华, 王怡伟, 梅凌瑞, 高洪成, 方俊峰, 和程学祺. Struedit: 结构化输出使大型语言模型的知识编辑快速准确. 2024.

[77] 鲍龙飞, 刘胜华, 王怡伟, 梅凌瑞, 高洪成, 许一龙, 和程学祺. Adaptive token biaser: 通过偏置关键实体进行知识编辑. *EMNLP2024*, 2024.

[78] 鲍龙飞, 刘胜华, 任兴章, 刘代恒, 林俊阳, 王怡伟, 梅凌瑞, 方俊峰, 郭嘉峰, 和程学祺. Refinex: 从专家指导程序中大规模学习预训练数据. 2025.

[79] 鲍龙飞, 刘胜华, 王怡伟, 梅凌瑞, 方俊峰, 高洪成, 倪时宇, 和程学祺. 事实性增强对大语言模型是免费的午餐吗？更好的事实性可能导致更差的上下文忠实度. *ICLR2025*, 2025.

[80] 鲍龙飞, 刘胜华, 王怡伟, 许一龙, 方俊峰, 梅凌瑞, 和程学祺. 参数与上下文：语言模型中知识依赖的细粒度控制. 2025.

[81] 毕斌, 李晨亮, 吴晨, 阎明, 和王伟. Palm：为上下文条件生成预训练自动编码&自回归语言模型. 自然语言处理经验方法会议, 2020.

[82] 丁多安·范 Bien, David Lillis, 和 Rem W. Collier. 多智能体系统的调用图分析.多智能体逻辑、语言和组织联合研讨会, 2009.

[83] Jonas Bode, Bastian Pätzold, Raphael Memmesheimer, 和 Sven Behnke. 服务机器人中任务规划和执行提示工程的比较. *IEEE-RAS* 国际人形机器人会议, 2024.

[84] P. Bonzon. 在虚拟多级功能框架中将心理表征接地. 认知学杂志, 2023.

[85] Sebastian Borgeaud, A. Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, T. Hennigan, Saffron Huang, Lorenzo Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, G. Irving, O. Vinyals, Simon Osindero, K. Simonyan, Jack W. Rae, Erich Elsen, 和 L. Sifre. 通过从数十亿个 token 中检索来改进语言模型. 国际机器学习会议, 2021.

[86] Zalán Borsos, Raphaël Marinier, Damien Vincent, E. Kharitonov, O. Pietquin, Matthew Sharifi, Dominik Roblek, O. Teboul, David Grangier, M. Tagliasacchi, 和 Neil Zeghidour. Audiolm: 一种用于音频生成的语言建模方法。 *IEEE/ACM* 语音、音频和语言处理汇刊, 2022。

[87] FutureSearch Nikos I. Bosse, Jon Evans, Robert G. Gambee, Daniel Hnyk, Peter Muhlbacher, Lawrence Phillips, Dan Schwarz, 和 Jack Wildman. 深度研究平台：评估人工智能网络研究代理，arXiv 预印本 arXiv:2506.06287, 2025。 URLhttps://arxiv.org/abs/2506.06287v1.

[88] Vicent Botti. Agentic ai and multiagentic: Are we reinventing the wheel?, arXiv preprint arXiv:2506.01463, 2025. URL https://arxiv.org/abs/2506.01463v1.

[89] William Brach, Kristián Kostál, and Michal Ries. The effectiveness of large language models in transforming unstructured text to standardized formats. *IEEE Access*, 2025.

[90] C. Brainerd, C. F. Gomes, and K. Nakamura. Dual recollection in episodic memory. *Journal of experimental psychology. General*, 2015.

[91] Inês Bramão, Jiefeng Jiang, A. Wagner, and M. Johansson. Encoding contexts are incidentally reinstated during competitive retrieval and track the temporal dynamics of memory interference. *Cerebral Cortex*, 2022.

[92] Andrés M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and P. Schwaller. Augmenting large language models with chemistry tools. *Nat. Mac. Intell.*, 2023.

[93] Maricela Claudia Bravo and Martha Coronel. Aligning agent communication protocols - a pragmatic approach. *International Conference on Software and Data Technologies*, 2008.

[94] F. Brazier, B. Dunin-Keplicz, N. Jennings, and Jan Treur. Desire: Modelling multi-agent systems in a compositional formal framework. *International Journal of Cooperative Information Systems*, 1997.

[95] Lorenz Brehme, Thomas Ströhle, and Ruth Breu. Can llms be trusted for evaluating rag systems? a survey of methods and datasets, arXiv preprint arXiv:2504.20119, 2025. URL https://arxiv.org/abs/2504.20119v2.

[96] R. Breil, D. Delahaye, Laurent Lapasset, and E. Feron. Multi-agent systems to help managing air traffic structure. arXiv preprint, 2017.

[97] Alexander Brinkmann and Christian Bizer. Self-refinement strategies for llm-based product attribute value extraction. *Datenbanksysteme für Business, Technologie und Web*, 2025.

[98] D. Britz, M. Guan, and Minh-Thang Luong. Efficient attention using a fixed-size memory representation. *Conference on Empirical Methods in Natural Language Processing*, 2017.

[99] Adam W. Broitman and M. J. Kahana. Neural context reinstatement of recurring events. *bioRxiv*, 2024.

[100] Ethan A. Brooks, Logan Walls, Richard L. Lewis, and Satinder Singh. Large language models can implement policy iteration. *Neural Information Processing Systems*, 2022.

[101] Rodney A. Brooks. A robust layered control system for a mobile robot. *IEEE J. Robotics Autom.*, 1986.

[102] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. *Computer Vision and Pattern Recognition*, 2022.

[103] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. Henighan, R. Child, A. Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, I. Sutskever, and Dario Amodei. Language models are few-shot learners. *Neural Information Processing Systems*, 2020.

[88]Vicent Botti. 自主式AI和多自主式：我们是在重新发明轮子吗？, arXiv预印本 arXiv:2506.01463, 2025. URLhttps://arxiv.org/abs/2506.01463v1.

[89] 威廉·布拉赫、克里斯蒂安·科斯塔尔和米哈尔·里斯。大型语言模型在将非结构化文本转换为标准化格式方面的有效性。*IEEEAccess*, 2025.

[90] C. Brainerd, C. F. Gomes, and K. Nakamura. 双重回忆在情景记忆中。 实验心理学杂志. 综合卷, 2015.

[91] Inês Bramão, Jiefeng Jiang, A. Wagner, 和 M. Johansson. 在竞争性检索过程中，编码上下文被偶然重新激活，并追踪了记忆干扰的时序动态。*Cerebral Cortex*, 2022.

[92] 安德烈斯·M·布兰，山姆·考克斯，奥利弗·施利特，卡洛·巴尔达萨里，安德鲁·D·怀特和P·施瓦勒。使用化学工具增强大型语言模型。*Nat.Mac.Intell.*,2023.

[93] Maricela Claudia Bravo和Martha Coronel.对齐智能体通信协议——一种实用方法。软件与数据技术国际会议, 2008。

[94] F. Brazier, B. Dunin-Keplicz, N. Jennings和Jan Treur. Desire：在组合形式框架中对多智能体系统进行建模。协同信息系统国际杂志, 1997。

[95] Lorenz Brehme, Thomas Ströhle和Ruth Breu. 大型语言模型是否可以信任来评估检索增强生成系统？方法和数据集的调查，arXiv预印本arXiv:2504.20119，2025。URLhttps://arxiv.org/abs/2504.20119v2。

[96] R. Breil, D. Delahaye, Laurent Lapasset和E. Feron. 用于帮助管理空中交通结构的智能体系统。arXiv预印本, 2017。

[97] Alexander Brinkmann 和 Christian Bizer. 基于llm的产品属性值提取的自完善策略。 *Datenbanksysteme für Business,Technologieund Web*, 2025。

[98] D. Britz, M. Guan, 和 Minh-Thang Luong. 使用固定大小内存表示的高效注意力机制。 *Conferenceon Empirical Methods in Natural Language Processing*, 2017。

[99] Adam W. Broitman 和 M. J. Kahana. 循环事件的神经上下文恢复。 *bioRxiv*, 2024。

[100] Ethan A. Brooks, Logan Walls, Richard L. Lewis, 和 Satinder Singh. 大型语言模型可以实现策略迭代。*NeuralInformationProcessingSystems*, 2022。

[101] Rodney A. Brooks. 移动机器的鲁棒分层控制系统。 *IEEEJ.Robotics Autom.*, 1986。

[102] Tim Brooks, Aleksander Holynski, 和 Alexei A. Efros. Instructpix2pix：学习遵循图像编辑指令。 *Computer Visionand PatternRecognition*, 2022。

[103] 汤姆·B·布朗，本杰明·曼，尼克·莱德，梅兰妮·苏比亚，J·卡普兰，普拉富拉·达里瓦尔，阿维恩德·尼尔卡坦，普拉纳夫·夏姆，吉里什·萨斯特里，阿曼达·阿斯凯尔，桑迪尼·阿加瓦尔，阿丽尔·赫伯特-沃斯，格雷琴·克鲁格，T·亨尼汉，R·查尔德，A·拉梅什，丹尼尔·M·齐格勒，杰夫·吴，克莱门斯·温特，克里斯托弗·赫塞，马克·陈，埃里克·西格尔，马特乌斯·利特温，斯科特·格雷，本杰明·切斯，杰克·克拉克，克里斯托弗·伯纳，山姆·麦卡迪尔，亚历克·拉德福德，I·苏茨凯弗，以及达里奥·阿莫迪。语言模型是少样本学习器。神经信息处理系统, 2020。

[104] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, arXiv preprint arXiv:2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

[105] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, L. Baraldi, and R. Cucchiara. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

[106] Joyce Cahoon, Prerna Singh, Nick Litombe, Jonathan Larson, Ha Trinh, Yiwen Zhu, Andreas Mueller, Fotis Psallidas, and Carlo Curino. Optimizing open-domain question answering with graph-based retrieval augmented generation, arXiv preprint arXiv:2503.02922, 2025. URL https://arxiv.org/abs/2503.02922v1.

[107] Hongru Cai, Yongqi Li, Wenjie Wang, Fengbin Zhu, Xiaoyu Shen, Wenjie Li, and Tat-Seng Chua. Large language models empowered personalized web agents. *The Web Conference*, 2024.

[108] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 666–682, 2018.

[109] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2272–2281, 2019.

[110] Yujun Cai, Liuhao Ge, Jianfei Cai, Nadia Magnenat Thalmann, and Junsong Yuan. 3d hand pose estimation using synthetic data and weakly labeled rgb images. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3739–3753, 2020.

[111] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 226–242. Springer International Publishing, 2020.

[112] Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, et al. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11645–11655, 2021.

[113] V. Camos and P. Barrouillet. Attentional and non-attentional systems in the maintenance of verbal information in working memory: the executive and phonological loops. *Frontiers in Human Neuroscience*, 2014.

[114] Boxi Cao, Qiaoyu Tang, Hongyu Lin, Xianpei Han, Jiawei Chen, Tianshu Wang, and Le Sun. Retentive or forgetful? diving into the knowledge memorizing mechanism of language models. *International Conference on Language Resources and Evaluation*, 2023.

[104] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert- Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 语言模型是少样本学习器, arXiv 预印本 arXiv:2005.14165, 2020. URLhttps://arxiv.org/abs/2005.14165.

[105] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, L. Baraldi, and R. Cucchiara. Wiki-llava: 多模态语言模型的层次化检索增强生成. *2024 IEEE/CVF 计算机视觉与模式识别研讨会 (CVPRW)*, 2024.

[106] Joyce Cahoon, Prerna Singh, Nick Litombe, Jonathan Larson, Ha Trinh, Yiwen Zhu, Andreas Mueller, Fotis Psallidas, and Carlo Curino. 优化开放域问答的图检索增强生成, arXiv 预印本 arXiv:2503.02922, 2025. URLhttps://arxiv.org/abs/2503.02922v1.

[107] Hongru Cai, Yongqi Li, Wenjie Wang, Fengbin Zhu, Xiaoyu Shen, Wenjie Li, and Tat-Seng Chua. 大语言模型赋能个性化网页代理。 *The WebConference*,2024。

[108] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. 从单目 RGB 图像中进行弱监督 3d 手部姿态估计。在 欧洲计算机视觉会议论文集 *(ECCV)*, 第 666–682 页, 2018。

[109] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. 利用图卷积网络通过时空关系进行 3d 姿态估计。在 *IEEE/CVF* 国际计算机视觉会议论文集, 第 2272– 2281 页, 2019。

[110] 蔡玉军, 葛柳浩, 蔡建飞, Nadia Magnenat Thalmann, 和 袁俊松. 使用合成数据和弱标签RGB图像进行3D手部姿态估计. *IEEE transactions on patternanalysisand machine intelligence*, 43(11):3739–3753, 2020.

[111] 蔡玉军, 黄琳, 王怡伟, Tat-Jen Cham, 蔡建飞, 袁俊松, 刘军, 杨旭, 朱毅恒, 沈晓辉, 等. 学习渐进式关节传播进行人体运动预测. In *Computer Vision–ECCV2020: 16th EuropeanConference, Glasgow, UK, August 23–28, 2020,Proceedings, Part VII 16*, 页面 226–242. Springer International Publishing, 2020.

[112] 蔡玉军, 王怡伟, 朱毅恒, Tat-Jen Cham, 蔡建飞, 袁俊松, 刘军, 郑传霞, 闫思杰, 丁恒辉, 等. 通过条件变分自动编码器实现统一的3D人体运动合成模型. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 页面 11645–11655, 2021.

[113] V. Camos 和 P. Barrouillet. 注意力和非注意力系统在工作记忆中维护语言信息：执行回路和语音回路。 *Frontiersin HumanNeuroscience*, 2014。

[114] Boxi Cao, Qiaoyu Tang, Hongyu Lin, Xianpei Han, JiaweiChen, Tianshu Wang 和 Le Sun. 持久还是遗忘？深入语言模型的知识记忆机制。 *InternationalConferenceon Language Resources and Evaluation*, 2023。

[115] He Cao, Zhenwei An, Jiazhan Feng, Kun Xu, Liwei Chen, and Dongyan Zhao. A step closer to comprehensive answers: Constrained multi-stage question decomposition with large language models, arXiv preprint arXiv:2311.07491, 2023. URL https://arxiv.org/abs/2311.07491v1.

[116] Nicola De Cao, Wilker Aziz, and Ivan Titov. Question answering by reasoning across documents with graph convolutional networks. *North American Chapter of the Association for Computational Linguistics*, 2018.

[117] Pengfei Cao, Tianyi Men, Wencan Liu, Jingwen Zhang, Xuzhao Li, Xixun Lin, Dianbo Sui, Yanan Cao, Kang Liu, and Jun Zhao. Large language models for planning: A comprehensive and systematic survey, arXiv preprint arXiv:2505.19683, 2025. URL https://arxiv.org/abs/2505.19683v1.

[118] Yongcan Cao, Wenwu Yu, W. Ren, and Guanrong Chen. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial Informatics*, 2012.

[119] Yuji Cao, Huan Zhao, Yuheng Cheng, Ting Shu, Yue Chen, Guolong Liu, Gaoqi Liang, Junhua Zhao, Jinyue Yan, and Yunjie Li. Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[120] Yukun Cao, Zengyi Gao, Zhiyang Li, Xike Xie, and S. K. Zhou. Lego-graphrag: Modularizing graph-based retrieval-augmented generation for design space exploration. arXiv preprint, 2024.

[121] R. C. Cardoso and Angelo Ferrando. A review of agent-based programming for multi-agent systems. *De Computis*, 2021.

[122] Nicholas Carlini, Chang Liu, Ú. Erlingsson, Jernej Kos, and D. Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. *USENIX Security Symposium*, 2018.

[123] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, D. Song, Ú. Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. *USENIX Security Symposium*, 2020.

[124] Daniel Casanueva-Morato, A. Ayuso-Martinez, J. P. Dominguez-Morales, A. Jiménez-Fernandez, and G. Jiménez-Moreno. A bio-inspired implementation of a sparse-learning spike-based hippocampus memory model. *IEEE Transactions on Emerging Topics in Computing*, 2022.

[125] Daniel Casanueva-Morato, A. Ayuso-Martinez, J. P. Dominguez-Morales, A. Jiménez-Fernandez, and G. Jiménez-Moreno. Bio-inspired computational memory model of the hippocampus: an approach to a neuromorphic spike-based content-addressable memory. *Neural Networks*, 2023.

[126] Amartya Chakraborty, Paresh Dashore, Nadia Bathaee, Anmol Jain, Anirban Das, Shi-Xiong Zhang, Sambit Sahu, M. Naphade, and Genta Indra Winata. T1: A tool-oriented conversational dataset for multi-turn agentic planning, arXiv preprint arXiv:2505.16986, 2025. URL https://arxiv.org/abs/2505.16986v1.

[127] Kranti Chalamalasetti, Jana Gotze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. clembench: Using game play to evaluate chat-optimized language models as conversational agents. *Conference on Empirical Methods in Natural Language Processing*, 2023.

[128] Edward Y. Chang and Longling Geng. Sagallm: Context management, validation, and transaction guarantees for multi-agent llm planning, arXiv preprint arXiv:2503.11951, 2025. URL https://arxiv.org/abs/2503.11951v2.

[115] 何超, 安振伟, 冯家赞, 许坤, 陈立伟, 和赵东岩. 大语言模型驱动的约束多阶段问题分解：迈向全面答案的一步, arXiv preprint arXiv:2311.07491, 2023. URLhttps://arxiv.org/abs/2311.07491v1.

[116] Nicola De Cao, Wilker Aziz, 和 Ivan Titov. 基于图卷积网络的跨文档推理问答. 北美学术计算语言学协会, 2018.

[117] 曹鹏飞, 门天一, 刘文灿, 张景文, 李序昭, 林希勋, 隋电波, 曹亚楠, 刘康, 和赵军. 大语言模型在规划中的应用：全面而系统的综述, arXiv preprint arXiv:2505.19683, 2025. URLhttps://arxiv.org/abs/2505.19683v1.

[118] 曹永灿, 余文武, W. 任, 和 陈冠荣. 分布式多智能体协调研究近期进展概述. *IEEETransactions onIndustrialInformatics*, 2012.

[119] 曹宇飞, 赵欢, 程宇恒, 舒挺, 陈岳, 刘国龙, 梁高奇, 赵军华, 严金月, 和 李云杰. 大语言模型增强强化学习综述：概念、分类和方法. *IEEETransactionsonNeuralNetworks and Learning Systems*, 2024.

[120] 曹雨坤, 高增谊, 李志阳, 谢锡科, 和 S. K. 周子. Lego-graphrag：模块化基于图的检索增强生成用于设计空间探索. arXiv preprint, 2024.

[121] R. C. Cardoso 和 Angelo Ferrando. 基于智能体的多智能体系统编程综述.*De Computis*, 2021.

[122] Nicholas Carlini, Chang Liu, Ú. Erlingsson, Jernej Kos, and D. Song. 秘密共享者：评估和测试神经网络中的意外记忆。 *USENIXSecurity Symposium*, 2018。

[123] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, D. Song, Ú. Erlingsson, Alina Oprea, and Colin Raffel. 从大型语言模型中提取训练数据。 *USENIXSecurity Symposium*, 2020。

[124] Daniel Casanueva-Morato, A. Ayuso-Martinez, J. P. Dominguez-Morales, A. Jiménez-Fernandez, and G. Jiménez-Moreno. 一种受生物启发的稀疏学习脉冲式海马体记忆模型实现。 *IEEE Transactions onEmergingTopics inComputing*, 2022。

[125] Daniel Casanueva-Morato, A. Ayuso-Martinez, J. P. Dominguez-Morales, A. Jiménez-Fernandez, and G. Jiménez-Moreno. 海马体的生物启发计算记忆模型：一种基于脉冲式内容寻址内存的方法。 *NeuralNetworks*, 2023。

[126] 阿马蒂亚·查克拉博尔蒂，帕雷斯·达肖雷，纳迪亚·巴塔伊，安莫尔·贾因，安伊兰·达斯，张石雄，萨比特·萨胡，M·纳法德，以及Genta·因德拉·温塔。T1：一个面向工具的对话数据集，用于多轮代理规划，arXiv预印本 arXiv:2505.16986，2025。URL https://arxiv.org/abs/2505.16986v1.

[127] 克拉尼蒂·查拉马拉塞蒂，贾娜·戈特泽，谢尔佐德·哈基莫夫，布里伦·马杜雷拉，菲利普·萨德勒，以及大卫·施拉格。clembench：使用游戏玩法来评估聊天优化的语言模型作为对话代理。自然语言处理经验方法会议，2023。

[128] 爱德华·Y·张和龙玲·耿。Sagallm：多智能体llm规划中的上下文管理、验证和事务保证，arXiv预印本 arXiv:2503.11951，2025。URL https://arxiv.org/abs/2503.11951v2.

[129] Yu-Chu Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Weirong Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qian Yang, and Xingxu Xie. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.

[130] Subhajit Chaudhury, Payel Das, Sarathkrishna Swaminathan, Georgios Kollias, Elliot Nelson, Khushbu Pahwa, Tejaswini Pedapati, Igor Melnyk, and Matthew Riemer. Epman: Episodic memory attention for generalizing to longer contexts, arXiv preprint arXiv:2502.14280, 2025. URL https://arxiv.org/abs/2502.14280v1.

[131] Xueqi CHEGN, Shenghua Liu, and Ruqing ZHANG. Thinking on new system for big data technology. *Bulletin of Chinese Academy of Sciences (Chinese Version)*, 37(1):60–67, 2022.

[132] Viktoriia Chekalina, Anton Razzigaev, Elizaveta Goncharova, and Andrey Kuznetsov. Addressing hallucinations in language models with knowledge graph embeddings as an additional modality, arXiv preprint arXiv:2411.11531, 2024. URL https://arxiv.org/abs/2411.11531v2.

[133] Bo Chen, Yingyu Liang, Zhizhou Sha, Zhenmei Shi, and Zhao Song. Hsr-enhanced sparse attention acceleration, arXiv preprint arXiv:2410.10165, 2024. URL https://arxiv.org/abs/2410.10165v2.

[134] Boyu Chen, Zirui Guo, Zidan Yang, Yuluo Chen, Junze Chen, Zhenghao Liu, Chuan Shi, and Cheng Yang. Pathrag: Pruning graph-based retrieval augmented generation with relational paths, arXiv preprint arXiv:2502.14902, 2025. URL https://arxiv.org/abs/2502.14902v1.

[135] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023.

[136] Feiyang Chen, Yu Cheng, Lei Wang, Yuqing Xia, Ziming Miao, Lingxiao Ma, Fan Yang, Jilong Xue, Zhi Yang, Mao Yang, and Haibo Chen. Attentionengine: A versatile framework for efficient attention mechanisms on diverse hardware platforms, arXiv preprint arXiv:2502.15349, 2025. URL https://arxiv.org/abs/2502.15349v1.

[137] Huajun Chen. Large knowledge model: Perspectives and challenges. *Data Intelligence*, 2023.

[138] Jianing Chen, Zehao Li, Yujun Cai, Hao Jiang, Chengxuan Qian, Juyuan Kang, Shuqin Gao, Honglong Zhao, Tianlu Mao, and Yucheng Zhang. Haif-gs: Hierarchical and induced flow-guided gaussian splatting for dynamic scene. 2025.

[139] Jiaqi Chen, Xiaoye Zhu, Yue Wang, Tianyang Liu, Xinhui Chen, Ying Chen, Chak Tou Leong, Yifei Ke, Joseph Liu, Yiwen Yuan, Julian McAuley, and Li jia Li. Symbolic representation for any-to-any generative tasks, arXiv preprint arXiv:2504.17261v1, 2025. URL https://arxiv.org/abs/2504.17261v1.

[140] Jiayi Chen, J. Ye, and Guiling Wang. From standalone llms to integrated intelligence: A survey of compound al systems, arXiv preprint arXiv:2506.04565, 2025. URL https://arxiv.org/abs/2506.04565v1.

[141] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, Defu Lian, and Enhong Chen. When large language models meet personalization: Perspectives of challenges and opportunities. *World wide web (Bussum)*, 2023.

[129] 张玉初, 王旭, 王晋东, 吴元, 朱凯捷, 陈浩, 杨林怡, 易晓媛, 王存祥, 王奕东, 叶伟荣, 张越, 常毅, Philip S. Yu, 杨倩, 谢兴旭. 大型语言模型评估综述. *ACM Transactions on Intelligent Systemsand Technology*, 2023.

[130] Subhajit Chaudhury, Payel Das, Sarathkrishna Swaminathan, Georgios Kollias, Elliot Nelson, Khushbu Pahwa, Tejaswini Pedapati, Igor Melnyk, and Matthew Riemer. Epman: 用于泛化到更长上下文的情景记忆注意力机制, arXiv 预印本 arXiv:2502.14280, 2025. URL https://arxiv.org/abs/2502.14280v1.

[131] 谢启程, 刘胜华, 张睿晴. 大数据技术新系统思考.中国科学院院刊（中文版）, 37(1):60–67, 2022.

[132] Viktoriia Chekalina, Anton Razzigaev, Elizaveta Goncharova, and Andrey Kuznetsov. 使用知识图谱嵌入作为附加模态来解决语言模型中的幻觉，arXiv preprint arXiv:2411.11531, 2024. URLhttps://arxiv.org/abs/2411.11531v2.

[133] Bo Chen, Yingyu Liang, Zhizhou Sha, Zhenmei Shi, and Zhao Song. Hsr增强的稀疏注意力加速，arXiv preprint arXiv:2410.10165, 2024. URLhttps://arxiv.org/abs/2410.10165v2.

[134] Boyu Chen, Zirui Guo, Zidan Yang, Yuluo Chen, Junze Chen, Zhenghao Liu, Chuan Shi, and Cheng Yang. Pathrag：基于关系的路径剪枝图检索增强生成，arXiv preprint arXiv:2502.14902, 2025. URLhttps://arxiv.org/abs/2502.14902v1.

[135] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023.

[136] 陈飞阳，程宇，王雷，夏宇清，苗子明，马凌晓，杨帆，薛继龙，杨智，杨毛，陈海波. Attentionengine: 一个适用于不同硬件平台的通用高效注意力机制框架，arXiv预印本 arXiv:2502.15349，2025. URLhttps://arxiv.org/abs/2502.15349v1.

[137] 陈华军. 大知识模型：视角与挑战。《数据智能》，2023.

[138] 陈建宁, 李泽豪, 蔡宇君, 姜浩, 钱成玄, 康宇源, 高淑琴, 赵洪龙, 毛天路, 张宇成. Haif-gs: 分层和诱导流引导高斯splatting用于动态场景. 2025.

[139] 陈嘉琪, 朱晓晔, 王越, 刘天阳, 陈新辉, 陈莹, 拉姆·查克·托伦, 谢逸飞, 刘约瑟夫, 袁一闻, 朱利安·麦克阿莱, 李佳. 任意到任意的生成任务的符号表示, arXiv 预印本 arXiv:2504.17261v1, 2025. URLhttps://arxiv.org/abs/2504.17261v1.

[140] Jiayi Chen, J. Ye, and Guiling Wang. 从独立大语言模型到集成智能：复合AI系统的调查, arXiv预印本 arXiv:2506.04565, 2025. URL https://arxiv.org/abs/2506.04565v1.

[141] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, Defu Lian, and Enhong Chen. 当大语言模型遇到个性化：挑战与机遇的视角。 万维网 *(Bussum)*, 2023。

[142] Jiyu Chen, Shuang Peng, Daxiong Luo, Fan Yang, Renshou Wu, Fangyuan Li, and Xiaoxin Chen. Edgeinfinite: A memory-efficient infinite-context transformer for edge devices, arXiv preprint arXiv:2503.22196, 2025. URL https://arxiv.org/abs/2503.22196v1.

[143] Justin Chih-Yao Chen, Archiki Prasad, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. Magicore: Multi-agent, iterative, coarse-to-fine refinement for reasoning, arXiv preprint arXiv:2409.12147, 2024. URL https://arxiv.org/abs/2409.12147v1.

[144] Mingyang Chen, Haoze Sun, Tianpeng Li, Fan Yang, Hao Liang, Keer Lu, Bin Cui, Wentao Zhang, Zenan Zhou, and Weipeng Chen. Facilitating multi-turn function calling for llms via compositional instruction tuning. *International Conference on Learning Representations*, 2024.

[145] Nuo Chen, Yuhan Li, Jianheng Tang, and Jia Li. Graphwiz: An instruction-following language model for graph computational problems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 353–364, 2024.

[146] Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. Judgelrm: Large reasoning models as a judge, arXiv preprint arXiv:2504.00050, 2025. URL https://arxiv.org/abs/2504.00050v1.

[147] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wangxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models, arXiv preprint arXiv:2503.09567, 2025. URL https://arxiv.org/abs/2503.09567v3.

[148] Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, Yimeng Zhang, Yihao Liang, Yuhang Zhou, Jiaqi Wang, Zhi Chen, and Wanxiang Che. Ai4research: A survey of artificial intelligence for scientific research, arXiv preprint arXiv:2507.01903, 2025. URL https://arxiv.org/abs/2507.01903.

[149] S Chen, Y Wang, YF Wu, and Q Chen.... Advancing tool-augmented large language models: Integrating insights from errors in inference trees. 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c0f7ee1901fef1da4dae2b88dfd43195-Abstract-Conference.html.

[150] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, arXiv preprint arXiv:2306.15595, 2023. URL https://arxiv.org/abs/2306.15595v2.

[151] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *International Conference on Machine Learning*, 2020.

[152] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Trans. Mach. Learn. Res.*, 2022.

[153] Yanda Chen, Ruiqi Zhong, Sheng Zha, G. Karypis, and He He. Meta-learning via language model in-context tuning. *Annual Meeting of the Association for Computational Linguistics*, 2021.

[154] Yi Chen, JiaHao Zhao, and HaoHao Han. A survey on collaborative mechanisms between large and small language models, arXiv preprint arXiv:2505.07460, 2025. URL https://arxiv.org/abs/2505.07460v1.

[142] 陈继宇，彭双，罗大雄，杨帆，吴仁寿，李方圆，陈晓欣. Edgeinfinite：一种适用于边缘设备的内存高效无限上下文Transformer，arXiv预印本arXiv:2503.22196，2025. URLhttps://arxiv.org/abs/2503.22196v1.

[143] Justin Chih-Yao Chen, Archiki Prasad, Swarnadeep Saha, Elias Stengel-Eskin, 和 Mohit Bansal. Magi-core: 多智能体、迭代、由粗到细的推理优化, arXiv 预印本 arXiv:2409.12147, 2024. URLhttps://arxiv.org/abs/2409.12147v1.

陈明阳，孙昊泽，李天鹏，杨帆，梁浩，陆科儿，崔斌，张文涛，周泽南，陈伟鹏. 通过组合式指令微调促进大语言模型的多轮函数调用. 国际学习表征会议, 2024.

[145] Nuo Chen, Yuhan Li, Jianheng Tang, and Jia Li. Graphwiz: 一个用于图计算问题的指令跟随语言模型. 在 第*30*届*ACM SIGKDD*知识发现与数据挖掘会议论文集, 第353-364页, 2024.

[146] Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. Judgelrm: 作为裁判的大规模推理模型, arXiv预印本 arXiv:2504.00050, 2025. URL https://arxiv.org/abs/2504.00050v1.

[147] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wangxiang Che. 迈向推理时代: 推理大型语言模型的链式思维长链综述, arXiv预印本 arXiv:2503.09567, 2025. URLhttps://arxiv.org/abs/2503.09567v3.

[148] 陈启光，杨明达，秦立博，刘金浩，闫铮，关建南，彭登云，纪一岩，李汉静，胡梦康，张一梦，梁一浩，周宇航，王嘉琪，陈智，车万祥. Ai4research: 科学研究人工智能综述，arXiv预印本 arXiv:2507.01903，2025. URLhttps://arxiv.org/abs/2507.01903.

[149] 陈思，王毅，吴彦飞，和陈强。... 推进工具增强型大语言模型：整合推理树中错误的洞察。 2024。URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c0f7ee1901fef1da4dae2b88dfd43195-Abstract-Conference.html.

[150] 陈寿元, 黄世民, 陈良建, 和 田元东. 通过位置插值扩展大型语言模型的上下文窗口, arXiv 预印本 arXiv:2306.15595, 2023. URLhttps://arxiv.org/abs/2306.15595v2.

[151] 陈婷, Simon Kornblith, Mohammad Norouzi, 和 Geoffrey E. Hinton. 视觉表示对比学习的简单框架. 机器学习国际会议, 2020.

[152] 陈文虎, 马学广, 王新怡, 和 William W. Cohen. 思维提示程序: 为数值推理任务将计算与推理分离. 机器学习研究转述, 2022.

[153] 陈岩达, 钟瑞琪, 沈铮, G. Karypis, 和 何浩. 通过语言模型上下文调优进行元学习. 计算语言学协会年度会议, 2021.

[154] Yi Chen, JiaHao Zhao, and HaoHao Han. 大型和小型语言模型之间的协作机制综述, arXiv 预印本 arXiv:2505.07460, 2025. URL https://arxiv.org/abs/2505.07460v1.

[155] Yixin Chen, Shuai Zhang, Boran Han, Tong He, and Bo Li. Camml: Context-aware multimodal learner for large models. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[156] Z Chen, K Zhou, B Zhang, Z Gong, and WX Zhao.... Chatcot: Tool-augmented chain-of-thought reasoning on chat-based large language models. 2023. URL https://arxiv.org/abs/2305.14323.

[157] Zehui Chen, Weihua Du, Wenwei Zhang, Kuikun Liu, Jiangning Liu, Miao Zheng, Jingming Zhuo, Songyang Zhang, Dahua Lin, Kai Chen, et al. T-eval: Evaluating the tool utilization capability step by step. *arXiv preprint arXiv:2312.14033*, 2023.

[158] Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. Mindsearch: Mimicking human minds elicits deep ai searcher, arXiv preprint arXiv:2407.20183, 2024. URL https://arxiv.org/abs/2407.20183v1.

[159] Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Haifang Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, and Jiliang Tang. Exploring the potential of large language models (llms)in learning on graphs. *SIGKDD Explorations*, 2023.

[160] Zihan Chen, Song Wang, Zhen Tan, Xingbo Fu, Zhenyu Lei, Peng Wang, Huan Liu, Cong Shen, and Jundong Li. A survey of scaling in large language model reasoning, arXiv preprint arXiv:2504.02181, 2025. URL https://arxiv.org/abs/2504.02181v1.

[161] ZY Chen, S Shen, G Shen, and G Zhi.... Towards tool use alignment of large language models. 2024. URL https://aclanthology.org/2024.emnlp-main.82/.

[162] Mingyue Cheng, Yucong Luo, Ouyang Jie, Qi Liu, Huijie Liu, Li Li, Shuo Yu, Bohou Zhang, Jiawei Cao, Jie Ma, Daoyu Wang, and Enhong Chen. A survey on knowledge-oriented retrieval-augmented generation, arXiv preprint arXiv:2503.10677, 2025. URL https://arxiv.org/abs/2503.10677v2.

[163] Ning Cheng, Zhaohui Yan, Ziming Wang, Zhijie Li, Jiaming Yu, Zilong Zheng, Kewei Tu, Jinan Xu, and Wenjuan Han. Potential and limitations of llms in capturing structured semantics: A case study on srl. *International Conference on Intelligent Computing*, 2024.

[164] Sitao Cheng, Ziyuan Zhuang, Yong Xu, Fangkai Yang, Chaoyun Zhang, Xiaoting Qin, Xiang Huang, Ling Chen, Qingwei Lin, Dongmei Zhang, et al. Call me when necessary: Llms can efficiently and faithfully reason over structured environments. In *Association for Computational Linguistics 2024*, pages 4275–4295, 2024.

[165] Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. Lift yourself up: Retrieval-augmented text generation with self memory. *Neural Information Processing Systems*, 2023.

[166] Yao Cheng, Yibo Zhao, Jiapeng Zhu, Yao Liu, Xing Sun, and Xiang Li. Human cognition inspired rag with knowledge graph for complex problem solving, arXiv preprint arXiv:2503.06567, 2025. URL https://arxiv.org/abs/2503.06567v1.

[167] Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, and Xiuqiang He. Exploring large language model based intelligent agents: Definitions, methods, and prospects, arXiv preprint arXiv:2401.03428, 2024. URL https://arxiv.org/abs/2401.03428v1.

[155] Yixin Chen, Shuai Zhang, Boran Han, Tong He, 和 Bo Li. Camml: 针对大型模型的上下文感知多模态学习器. 计算语言学协会年度会议, 2024.

[156] Z Chen, K Zhou, B Zhang, Z Gong, 和 WX Zhao. . . . 基于聊天的大型语言模型的工具增强思维链推理. 2023. URLhttps://arxiv.org/abs/2305.14323.

[157] Zehui Chen, Weihua Du, Wenwei Zhang, Kuikun Liu, Jiangning Liu, Miao Zheng, Jingming Zhuo, Songyang Zhang, Dahua Lin, Kai Chen, 等. T-eval: 逐步评估工具使用能力. *arXiv* 预印本 *arXiv:2312.14033*, 2023.

[158] Zehui Chen,Kuikun Liu,Qiuchen Wang,JiangningLiu, Wenwei Zhang, Kai Chen, 和FengZhao. Mindsearch: 模仿人类思维激发深度 AI 搜索器, arXiv 预印本 arXiv:2407.20183, 2024. URLhttps://arxiv.org/abs/2407.20183v1.

[159] 陈志凯, 毛海涛, 李航, 金伟, 温海方, 魏晓驰, 王帅强, 尹大伟, 范文奇, 刘辉, 唐继良. 探索大型语言模型（llms）在图学习中的潜力. *SIGKDD Explorations*, 2023.

[160] 陈子涵, 王松, 谭振, 傅兴波, 雷振宇, 王鹏, 刘欢, 沈从, 李军东. 大型语言模型推理中的扩展性调查, arXiv preprint arXiv:2504.02181, 2025. URLhttps://arxiv.org/abs/2504.02181v1.

[161] 陈ZY, 沈S, 沈G, 和 赵G.. .. 大型语言模型工具使用对齐. 2024. URLhttps://aclanthology.org/2024.emnlp-main.82/.

[162] Mingyue Cheng, Yucong Luo, Ouyang Jie, Qi Liu, Huijie Liu, Li Li, Shuo Yu, Bohou Zhang, Jiawei Cao, Jie Ma, Daoyu Wang, and Enhong Chen. 知识导向检索增强生成综述，arXiv 预印本 arXiv:2503.10677，2025。URL https://arxiv.org/abs/2503.10677v2.

[163] Ning Cheng, Zhaohui Yan, Ziming Wang, Zhijie Li, Jiaming Yu, Zilong Zheng, Kewei Tu, Jinan Xu, and Wenjuan Han. 大型语言模型在捕获结构化语义中的潜力和局限性：基于句法角色标注的案例研究。 国际智能计算会议, 2024。

[164] Sitao Cheng, Ziyuan Zhuang, Yong Xu, Fangkai Yang, Chaoyun Zhang, Xiaoting Qin, Xiang Huang, Ling Chen, Qingwei Lin, Dongmei Zhang, et al. 必要时给我打电话：大型语言模型可以高效且忠实地对结构化环境进行推理。在 计算语言学协会 *2024*，第 4275–4295 页，2024。

[165] 辛成, 罗迪, 陈秀英, 刘乐毛, 赵冬岩, 和严瑞岩. 振作起来：具有自记忆的检索增强文本生成. 神经信息处理系统, 2023.

[166] 姚成, 赵一博, 朱嘉鹏, 刘瑶, 孙行, 和李翔. 受人类认知启发的rag与知识图谱用于复杂问题求解, arXiv预印本 arXiv:2503.06567, 2025. URLhttps://arxiv.org/abs/2503.06567v1.

[167] 程宇恒, 张晨瑶, 张正文, 孟祥瑞, 洪思睿, 李文豪, 王子豪, 王泽凯, 殷峰, 赵军华, 和何秀强. 探索基于大型语言模型的智能体: 定义、方法与展望, arXiv预印本 arXiv:2401.03428, 2024. URLhttps://arxiv.org/abs/2401.03428v1.

[168] Egor Cherepanov, Nikita Kachaev, A. Kovalev, and Aleksandr I. Panov. Memory, benchmark & robots: A benchmark for solving complex tasks with reinforcement learning, arXiv preprint arXiv:2502.10550, 2025. URL https://arxiv.org/abs/2502.10550v2.

[169] Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory, arXiv preprint arXiv:2504.19413, 2025. URL https://arxiv.org/abs/2504.19413.

[170] Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. *Findings*, 2022.

[171] Jihye Choi, Nils Palumbo, P. Chalasani, Matthew M. Engelhard, Somesh Jha, Anivarya Kumar, and David Page. Malade: Orchestration of llm-powered agents with retrieval augmented generation for pharmacovigilance, arXiv preprint arXiv:2408.01869, 2024. URL https://arxiv.org/abs/2408.01869v1.

[172] K. Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy J. Colwell, and Adrian Weller. Rethinking attention with performers. *International Conference on Learning Representations*, 2020.

[173] Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S. Yu, and Qingsong Wen. Llm agents for education: Advances and applications, arXiv preprint arXiv:2503.11733, 2025. URL https://arxiv.org/abs/2503.11733v1.

[174] Zhixuan Chu, Huaiyu Guo, Xinyuan Zhou, Yijia Wang, Fei Yu, Hong Chen, Wanqing Xu, Xin Lu, Qing Cui, Longfei Li, Junqing Zhou, and Sheng Li. Data-centric financial large language models, arXiv preprint arXiv:2310.17784, 2023. URL https://arxiv.org/abs/2310.17784v2.

[175] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. Simulating opinion dynamics with networks of llm-based agents, arXiv preprint arXiv:2311.09618, 2024. URL https://arxiv.org/abs/2311.09618.

[176] Yung-Sung Chuang, Benjamin Cohen-Wang, Shannon Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James Glass, Shang-Wen Li, and Wen tau Yih. Selfcite: Self-supervised alignment for context attribution in large language models, arXiv preprint arXiv:2502.09604, 2025. URL https://arxiv.org/abs/2502.09604v3.

[177] Julian Coda-Forno, Marcel Binz, Zeynep Akata, M. Botvinick, Jane X. Wang, and Eric Schulz. Meta-in-context learning in large language models. *Neural Information Processing Systems*, 2023.

[178] Joao Coelho, Jingjie Ning, Jingyuan He, Kangrui Mao, A. Paladugu, Pranav Setlur, Jiahe Jin, James P. Callan, João Magalhães, Bruno Martins, and Chenyan Xiong. Deepresearchgym: A free, transparent, and reproducible evaluation sandbox for deep research, arXiv preprint arXiv:2505.19253, 2025. URL https://arxiv.org/abs/2505.19253v2.

[179] Emile Contal and Garrin McGoldrick. Ragsys: Item-cold-start recommender as rag system. *IR-RAG@SIGIR*, 2024.

[180] Erica Coppolillo. Injecting knowledge graphs into large language models, arXiv preprint arXiv:2505.07554, 2025. URL https://arxiv.org/abs/2505.07554v1.

[168] Egor Cherepanov, Nikita Kachaev, A. Kovalev, and Aleksandr I. Panov. Memory, benchmark & robots: A benchmark for solving complex tasks with reinforcement learning, arXiv preprint arXiv:2502.10550, 2025. URL https://arxiv.org/abs/2502.10550v2.

[169] Prateek Chhikara、Dev Khant、Saket Aryan、Taranjeet Singh和Deshraj Yadav。Mem0：构建具有可扩展长期记忆的生产就绪式AI代理，arXiv预印本arXiv:2504.19413，2025年。URLhttps://arxiv.org/abs/2504.19413。

[170] Yew Ken Chia, Lidong Bing, Soujanya Poria 和 Luo Si. Relationprompt：利用提示生成合成数据以进行零样本关系三元组抽取。发现, 2022.

[171] Jihye Choi, Nils Palumbo, P. Chalasani, Matthew M. Engelhard, Somesh Jha, Anivarya Kumar, and David Page. Malade: 基于检索增强生成的大型语言模型驱动的药物警戒智能体编排, arXiv 预印本 arXiv:2408.01869, 2024. URLhttps://arxiv.org/abs/2408.01869v1.

[172] K. Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy J. Colwell, and Adrian Weller. 以表演者重新思考注意力。国际学习表征会议, 2020.

[173] 甄东楚, 王沈, 谢建, 朱停辉, 严一博, 叶金恒, 钟奥晓, 胡旭明, 梁静, 余思培, 文清松. 教育领域的大型语言模型代理：进展与应用, arXiv 预印本 arXiv:2503.11733, 2025. URLhttps://arxiv.org/abs/2503.11733v1.

[174] 朱志轩, 郭怀宇, 周新元, 王一嘉, 余飞, 陈红, 许万庆, 陆欣, 崔清, 李龙飞, 周俊庆, 和李胜. 数据驱动的金融大语言模型, arXiv 预印本 arXiv:2310.17784, 2023. URLhttps://arxiv.org/abs/2310.17784v2.

[175] 庄云水, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, 杨思嘉, Shah Dhavan, 胡俊杰, 和 Timothy T. Rogers. 基于llm代理的网络模拟意见动态, arXiv 预印本 arXiv:2311.09618, 2024. URL https://arxiv.org/abs/2311.09618.

[176] 庄永松, Benjamin Cohen-Wang, 沈泽江, 吴兆丰, 许虎, 林西维多利亚, Glass James, 李尚文, 和 Yih Wen tau. Selfcite: 大语言模型中上下文归因的自监督对齐, arXiv 预印本 arXiv:2502.09604, 2025. URL https://arxiv.org/abs/2502.09604v3.

[177] Julian Coda-Forno, Marcel Binz, Zeynep Akata, M. Botvinick, Jane X. Wang, and Eric Schulz. 元上下文学习在大语言模型中. *NeuralInformationProcessingSystems*, 2023.

[178] Joao Coelho, Jingjie Ning, Jingyuan He, Kangrui Mao, A. Paladugu, Pranav Setlur, Jiahe Jin, James P. Callan, João Magalhães, Bruno Martins, and Chenyan Xiong. Deepresearchgym: 一个免费、透明且可复现的深度研究评估沙盒, arXiv preprint arXiv:2505.19253, 2025. URL https://arxiv.org/abs/2505.19253v2.

[179] Emile Contal and Garrin McGoldrick. Ragsys: 作为RAG系统的项目冷启动推荐器. *IR-RAG@SIGIR*, 2024.

[180]Erica Coppolillo. 将知识图谱注入大型语言模型，arXiv 预印本 arXiv:2505.07554，2025。URLhttps://arxiv.org/abs/2505.07554v1.

[181] R. P. Costa, R. Froemke, P. J. Sjöström, and Mark C. W. van Rossum. Unified pre- and postsynaptic long-term plasticity enables reliable and flexible learning. *eLife*, 2015.

[182] Caia Costello, Simon Guo, Anna Goldie, and Azalia Mirhoseini. Think, prune, train, improve: Scaling reasoning without scaling models, arXiv preprint arXiv:2504.18116, 2025. URL https://arxiv.org/abs/2504.18116v1.

[183] Michael Craig, Karla Butterworth, Jonna Nilsson, Colin J Hamilton, P. Gallagher, and T. Smulders. How does intentionality of encoding affect memory for episodic information? *Learning & memory (Cold Spring Harbor, N.Y.)*, 2016.

[184] crewAI Inc. crewai: Framework for orchestrating role-playing, autonomous ai agents. https://github.com/crewAIInc/crewAI, 2024. [Online; accessed 17-July-2025].

[185] A. Cruz, André V. dos Santos, R. Santiago, and B. Bedregal. A fuzzy semantic for bdi logic. *Fuzzy Information and Engineering*, 2021.

[186] Florin Cuconasu, Giovanni Trappolini, F. Siciliano, Simone Filice, Cesare Campagnano, Y. Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems. *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

[187] Kai Cui, Anam Tahir, Gizem Ekinci, Ahmed Elshamanhory, Yannick Eich, Mengguang Li, and H. Koeppl. A survey on large-population systems and scalable multi-agent reinforcement learning, arXiv preprint arXiv:2209.03859, 2022. URL https://arxiv.org/abs/2209.03859v1.

[188] Yuanning Cui, Zequn Sun, and Wei Hu. A prompt-based knowledge graph foundation model for universal in-context reasoning. In *Advances in Neural Information Processing Systems*, 2024.

[189] Yue Cui, Liuyi Yao, Shuchang Tao, Weijie Shi, Yaliang Li, Bolin Ding, and Xiaofang Zhou. Enhancing tool learning in large language models with hierarchical error checklists, arXiv preprint arXiv:2506.00042, 2025. URL https://arxiv.org/abs/2506.00042v1.

[190] C. Curto, A. Degeratu, and V. Itskov. Flexible memory networks. *Bulletin of Mathematical Biology*, 2010.

[191] Ruiting Dai, Yuqiao Tan, Lisi Mo, Shuang Liang, Guohao Huo, Jiayi Luo, and Yao Cheng. G-sap: Graph-based structure-aware prompt learning over heterogeneous knowledge for commonsense reasoning. *International Conference on Multimedia Retrieval*, 2024.

[192] Xinnan Dai, Haohao Qu, Yifen Shen, Bohang Zhang, Qihao Wen, Wenqi Fan, Dongsheng Li, Jiliang Tang, and Caihua Shan. How do large language models understand graph patterns? a benchmark for graph pattern comprehension, arXiv preprint arXiv:2410.05298v2, 2024. URL https://arxiv.org/abs/2410.05298v2.

[193] Fatemeh Daneshfar and H. Bevrani. Multi-agent systems in control engineering: a survey. arXiv preprint, 2009.

[194] Yufan Dang, Cheng Qian, Xueheng Luo, Jingru Fan, Zihao Xie, Ruijie Shi, Weize Chen, Cheng Yang, Xiaoyin Che, Ye Tian, Xuantang Xiong, Lei Han, Zhiyuan Liu, and Maosong Sun. Multi-agent collaboration via evolving orchestration, arXiv preprint arXiv:2505.19591, 2025. URL https://arxiv.org/abs/2505.19591v1.

[181] R. P. Costa, R. Froemke, P. J. Sjöström, and Mark C. W. van Rossum. 统一的前后突触长时程可塑性使可靠和灵活的学习成为可能。 *eLife*, 2015。

[182] Caia Costello, Simon Guo, Anna Goldie, and Azalia Mirhoseini. 思考、剪枝、训练、改进：在不扩展模型的情况下扩展推理，arXiv 预印本 arXiv:2504.18116，2025。URL https://arxiv.org/abs/2504.18116v1.

[183] Michael Craig,Karla Butterworth, Jonna Nilsson, Colin J Hamilton, P. Gallagher, andT. Smulders. 编码的意向性如何影响情景信息的记忆？ 学习与记忆（冷泉港，纽约），2016。

[184] crewAI Inc. crewai：用于编排角色扮演、自主 AI 代理的框架。
https://github.com/crewAIInc/crewAI,2024.［在线；访问于 17-July-2025].

[185] A. Cruz, André V. dos Santos, R. Santiago, and B. Bedregal. 基于模糊语义的bdi逻辑。 *FuzzyInformationand Engineering*, 2021。

[186] Florin Cuconasu, Giovanni Trappolini, F. Siciliano, Simone Filice, Cesare Campagnano, Y. Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 噪声的力量：重新定义rag系统的检索。*Annual International ACM SIGIR Conference on Research and Development inInformation Retrieval*, 2024。

[187] Kai Cui, Anam Tahir, Gizem Ekinci, Ahmed Elshamanhory, Yannick Eich, Mengguang Li, and H. Koeppl. 大规模系统与可扩展多智能体强化学习的综述，arXiv预印本arXiv:2209.03859，2022。
URLhttps://arxiv.org/abs/2209.03859v1。

[188] 崔袁宁, 孙泽群, 和胡伟. 基于提示的知识图谱基础模型用于通用情境推理. 在神经信息处理系统进展, 2024.

[189] 崔越, 姚柳艺, 陶树昌, 石伟杰, 李亚良, 丁波林, 和周晓方. 基于分层错误检查表的工具学习增强大型语言模型, arXiv 预印本 arXiv:2506.00042, 2025. URLhttps://arxiv.org/abs/2506.00042v1.

[190] C. Curto, A. Degeratu, 和 V. Itskov. 灵活记忆网络. 数学生物学通报, 2010.

[191] 戴瑞婷, 谭宇桥, 摩立思, 梁双, 胡国豪, 罗佳怡, 和程瑶. G-sap: 基于图的异构知识结构感知提示学习用于常识推理. 多媒体检索国际会议, 2024.

[192] Xinnan Dai, Haohao Qu, Yifen Shen, Bohang Zhang, Qihao Wen, Wenqi Fan, Dongsheng Li, Jiliang Tang, and Caihua Shan. 大语言模型如何理解图模式？一个图模式理解的基准, arXiv preprint arXiv:2410.05298v2, 2024. URL https://arxiv.org/abs/2410.05298v2.

[193] Fatemeh Daneshfar and H. Bevrani. 控制工程中的多智能体系统：综述。arXiv preprint, 2009.

[194] Yufan Dang, Cheng Qian, Xueheng Luo, Jingru Fan, Zihao Xie, Ruijie Shi, Weize Chen, Cheng Yang, Xiaoyin Che, Ye Tian, Xuantang Xiong, Lei Han, Zhiyuan Liu, and Maosong Sun. 通过演化编排进行多智能体协作，arXiv preprint arXiv:2505.19591, 2025. URL https://arxiv.org/abs/2505.19591v1.

[195] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *International Conference on Learning Representations*, 2023.

[196] Tri Dao, Daniel Y. Fu, Stefano Ermon, A. Rudra, and Christopher R'e. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Neural Information Processing Systems*, 2022.

[197] D Das, D Banerjee, S Aditya, and A Kulkarni. Mathsensei: a tool-augmented large language model for mathematical reasoning. 2024. URL https://arxiv.org/abs/2402.17231.

[198] Payel Das, Subhajit Chaudhury, Elliot Nelson, Igor Melnyk, Sarath Swaminathan, Sihui Dai, Aurélie Lozano, Georgios Kollias, Vijil Chenthamarakshan, Jiří, Navrátil, Soham Dan, and Pin-Yu Chen. Larimar: Large language models with episodic memory control, arXiv preprint arXiv:2403.11901, 2024. URL https://arxiv.org/abs/2403.11901.

[199] Adrian de Wynter, Xun Wang, Qilong Gu, and Si-Qing Chen. On meta-prompting, arXiv preprint arXiv:2312.06562, 2023. URL https://arxiv.org/abs/2312.06562v3.

[200] Ramandeep Singh Dehal, Mehak Sharma, and Enayat Rajabi. Knowledge graphs and their reciprocal relationship with large language models. *Machine Learning and Knowledge Extraction*, 2025.

[201] Mauricio R. Delgado, V. Stenger, and J. Fiez. Motivation-dependent responses in the human caudate nucleus. *Cerebral Cortex*, 2004.

[202] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Neural Information Processing Systems*, 2023.

[203] Yang Deng, Wenqiang Lei, Hongru Wang, and Tat seng Chua. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. *Conference on Empirical Methods in Natural Language Processing*, 2023.

[204] Yang Deng, An Zhang, Yankai Lin, Xu Chen, Ji-Rong Wen, and Tat-Seng Chua. Large language model powered agents in the web. *The Web Conference*, 2024.

[205] Yang Deng, Xuan Zhang, Wenxuan Zhang, Yifei Yuan, See-Kiong Ng, and Tat-Seng Chua. On the multi-turn instruction following for conversational web agents. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[206] Brouillet Denis and Versace Rémy. The nature of the traces and the dynamics of memory. *Psychology and Behavioral Sciences*, 2019.

[207] Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees G. M. Snoek, M. Worring, and Yuki Asano. Self-supervised open-ended classification with small visual language models, arXiv preprint arXiv:2310.00500, 2023. URL https://arxiv.org/abs/2310.00500v2.

[208] Stefan Dernbach, Khushbu Agarwal, Alejandro Zuniga, Michael Henry, and Sutanay Choudhury. Glam: Fine-tuning large language models for domain knowledge graph alignment via neighborhood partitioning and generative subgraph encoding. *AAAI Spring Symposia*, 2024.

[209] Rushali Deshmukh, Rutuj Raut, Mayur Bhavsar, Sanika Gurav, and Y. Patil. Optimizing human-ai interaction: Innovations in prompt engineering. *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, 2025.

[195] Tri Dao. Flashattention-2: 更好的并行性和工作分区的高速注意力机制。 国际学习表征会议, 2023。

[196] Tri Dao, Daniel Y. Fu, Stefano Ermon, A. Rudra, 和 Christopher R'e. Flashattention: 快速且内存高效的精确注意力机制，具有io感知能力。 神经信息处理系统会议, 2022。

[197] D Das, D Banerjee, S Aditya, 和 A Kulkarni. Mathsensei: 一种用于数学推理的工具增强型大型语言模型。2024. URLhttps://arxiv.org/abs/2402.17231.

[198] PayelDas,SubhajitChaudhury,Elliot Nelson, IgorMelnyk, SarathSwaminathan, Sihui Dai,Aurélie Lozano, Georgios Kollias, Vijil Chenthamarakshan, Jiří, Navrátil, Soham Dan, 和 Pin-Yu Chen. Larimar: 具有情景记忆控制的大型语言模型，arXiv预印本 arXiv:2403.11901, 2024. URLhttps://arxiv.org/abs/2403.11901.

[199] 阿德里安·德·温特，王勋，顾启龙，陈思清。关于元提示，arXiv 预印本 arXiv:2312.06562，2023。URLhttps://arxiv.org/abs/2312.06562v3。

[200] Ramandeep Singh Dehal、Mehak Sharma和Enayat Rajabi。知识图谱及其与大型语言模型的互惠关系。机器学习与知识提取, 2025.

[201] Mauricio R. Delgado、V. Stenger 和 J. Fiez。动机依赖性的人类尾状核反应。*Cerebral Cortex*，2004。

[202] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Neural Information Processing Systems*, 2023.

杨登，雷文强，王红如，和蔡达生。针对主动对话提示和评估大型语言模型：澄清、目标导向和非协作。自然语言处理经验方法会议，2023。

[204] 杨登, 张安, 林岩凯, 陈旭, 温继荣, 和蔡天锡. 基于大型语言模型的网络智能体. 网络会议, 2024.

[205] 杨登, 张玄, 张文轩, 袁一飞, 郭志强, 和蔡天锡. 对话式网络智能体的多轮指令遵循. 计算语言学协会年会, 2024.

[206] 布罗伊莱特 丹尼斯 和 维尔萨克 雷米. 轨迹的本质和记忆的动态. 心理学与行为科学, 2019.

[207] 穆罕默德·马赫迪·德拉克什阿尼, 伊沃娜·纳热德科斯卡, Cees G. M. 斯诺克, M. 沃林, 和黑木由纪. 基于小型视觉语言模型的自监督开放式分类, arXiv 预印本 arXiv:2310.00500, 2023. URLhttps://arxiv.org/abs/2310.00500v2.

[208] Stefan Dernbach, Khushbu Agarwal, Alejandro Zuniga, Michael Henry, and Sutanay Choudhury. Glam: 基于邻域划分和生成子图编码的领域知识图谱对齐的大型语言模型微调. *AAAI SpringSymposia*, 2024.

[209] Rushali Deshmukh, Rutuj Raut, Mayur Bhavsar, Sanika Gurav, and Y. Patil. 优化人机交互：提示工程的创新. *20253rd International Conference on IntelligentData Communication Technologies and Internetof Things (IDCIoT)*, 2025.

[210] Darshan Deshpande, Varun Gangal, Hersh Mehta, Jitin Krishnan, Anand Kannappan, and Rebecca Qian. Trail: Trace reasoning and agentic issue localization, arXiv preprint arXiv:2505.08638, 2025. URL https://arxiv.org/abs/2505.08638v3.

[211] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics*, 2019.

[212] Dhruv Dhamani and Mary Lou Maher. The tyranny of possibilities in the design of task-oriented llm systems: A scoping survey, arXiv preprint arXiv:2312.17601, 2023. URL https://arxiv.org/abs/2312.17601v1.

[213] Frederick Dillon, Gregor Halvorsen, Simon Tattershall, Magnus Rowntree, and Gareth Vanderpool. Contextual memory reweaving in large language models using layered latent state reconstruction, arXiv preprint arXiv:2502.02046, 2025. URL https://arxiv.org/abs/2502.02046v2.

[214] Hanxing Ding, Shuchang Tao, Liang Pang, Zihao Wei, Jinyang Gao, Bolin Ding, Huawei Shen, and Xueqi Chen. Toolcoder: A systematic code-empowered tool learning framework for large language models, arXiv preprint arXiv:2502.11404, 2025. URL https://arxiv.org/abs/2502.11404v2.

[215] Hongxin Ding, Yue Fang, Runchuan Zhu, Xinke Jiang, Jinyang Zhang, Yongxin Xu, Xu Chu, Junfeng Zhao, and Yasha Wang. 3ds: Decomposed difficulty data selection's case study on llm medical domain adaptation. 2024.

[216] Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, and Furu Wei. Longnet: Scaling transformers to 1, 000, 000, 000 tokens. arXiv preprint, 2023.

[217] Tianyu Ding, Tianyi Chen, Haidong Zhu, Jiachen Jiang, Yiqi Zhong, Jinxin Zhou, Guangzhi Wang, Zhihui Zhu, Ilya Zharkov, and Luming Liang. The efficiency spectrum of large language models: An algorithmic survey, arXiv preprint arXiv:2312.00678, 2023. URL https://arxiv.org/abs/2312.00678v2.

[218] Yiran Ding, L. Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. *International Conference on Machine Learning*, 2024.

[219] Yiwen Ding, Zhiheng Xi, Wei He, Zhuoyuan Li, Yitao Zhai, Xiaowei Shi, Xunliang Cai, Tao Gui, Qi Zhang, and Xuanjing Huang. Mitigating tail narrowing in llm self-improvement via socratic-guided sampling. *North American Chapter of the Association for Computational Linguistics*, 2024.

[220] Christian Djeffal. Reflexive prompt engineering: A framework for responsible prompt engineering and ai interaction design. *Conference on Fairness, Accountability and Transparency*, 2025.

[221] G Dong, Y Chen, X Li, J Jin, H Qian, and Y Zhu.... Tool-star: Empowering llm-brained multi-tool reasoner via reinforcement learning. 2025. URL https://arxiv.org/abs/2505.16410.

[222] Guanting Dong, Jinxu Zhao, Tingfeng Hui, Daichi Guo, Wenlong Wan, Boqi Feng, Yueyan Qiu, Zhuoma Gongque, Keqing He, Zechen Wang, and Weiran Xu. Revisit input perturbation problems for llms: A unified robustness evaluation framework for noisy slot filling task. *Natural Language Processing and Chinese Computing*, 2023.

[210] Darshan Deshpande、Varun Gangal、Hersh Mehta、Jitin Krishnan、Anand Kannappan和Rebecca Qian。Trail：追踪推理和代理问题定位，arXiv预印本arXiv:2505.08638，2025年。URLhttps://arxiv.org/abs/2505.08638v3。

[211] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：为语言理解而进行深度双向 Transformer 的预训练。北美学术计算语言学协会，2019。

[212] Dhruv Dhamani 和 Mary Lou Maher. 任务导向型大语言模型系统设计中的可能性暴政：一项范围界定调查，arXiv 预印本 arXiv:2312.17601，2023年。URL https://arxiv.org/abs/2312.17601v1.

[213] 弗雷德里克·迪尔隆，格雷戈尔·哈夫索森，西蒙·塔特萨尔，马格纳斯·罗恩特里和加雷斯·范德普尔。使用分层潜在状态重建在大语言模型中进行情境记忆重织，arXiv预印本arXiv:2502.02046，2025年。URLhttps://arxiv.org/abs/2502.02046v2.

[214] 韩兴丁, 陶树昌, 庞亮, 魏子豪, 高金阳, 丁博林, 沈华伟, 陈雪琪. Toolcoder: 一种面向大型语言模型的系统化代码赋能工具学习框架, arXiv 预印本 arXiv:2502.11404, 2025. URL https://arxiv.org/abs/2502.11404v2.

[215] 丁红心，方越，朱润川，蒋新科，张晋阳，徐永欣，褚旭，赵俊峰，王亚莎。3ds：基于分解难度数据选择的大语言模型医疗领域自适应案例研究。2024。

[216] Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, and Furu Wei. Longnet: Scaling transformers to 1, 000, 000, 000 tokens. arXiv preprint, 2023.

[217] 丁天宇, 陈天毅, 朱海东, 姜嘉晨, 钟一奇, 周金鑫, 王光智, 朱志辉, 伊利亚·扎尔科夫, 和梁鲁明. 大型语言模型的效率谱：算法综述, arXiv preprint arXiv:2312.00678, 2023. URLhttps://arxiv.org/abs/2312.00678v2.

[218] 丁一然, 张丽, 张成荣东, 许媛媛, 尚宁, 许家航, 杨帆, 和杨毛. Longrope：将llm上下文窗口扩展到2000万个token. 机器学习国际会议, 2024.

[219] 丁一文, 西志恒, 何伟, 李卓远, 翟一涛, 石晓伟, 蔡训良, 郭涛, 张奇, 和黄宣静. 通过苏格拉底式引导采样缓解llm自我改进中的尾部变窄问题. 计算语言学协会北美分会, 2024.

[220] Christian Djeffal. 反思式提示工程：负责任提示工程和人工智能交互设计的框架。 公平、问责和透明度会议, 2025。

[221] G Dong, Y Chen, X Li, J Jin, H Qian, and Y Zhu.... Tool-star: 通过强化学习赋能具有大型语言模型思维的多工具推理器。 2025. URLhttps://arxiv.org/abs/2505.16410.

[222] Guanting Dong, Jinxu Zhao, Tingfeng Hui, Daichi Guo, Wenlong Wan, Boqi Feng, Yueyan Qiu, Zhuoma Gongque, Keqing He, Zechen Wang, and Weiran Xu. 重新审视大型语言模型的输入扰动问题：用于噪声槽填充任务的统一鲁棒性评估框架。 自然语言处理和中国计算, 2023。

[223] Guanting Dong, Yifei Chen, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Yutao Zhu, Hangyu Mao, Guorui Zhou, Zhicheng Dou, and Ji-Rong Wen. Tool-star: Empowering llm-brained multi-tool reasoner via reinforcement learning. arXiv preprint, 2025.

[224] Kaiwen Dong. Large language model applied in multi-agent systema survey. *Applied and Computational Engineering*, 2024.

[225] Peijie Dong, Zhenheng Tang, Xiang-Hong Liu, Lujun Li, Xiaowen Chu, and Bo Li. Can compressed llms truly act? an empirical evaluation of agentic capabilities in llm compression, arXiv preprint arXiv:2505.19433, 2025. URL https://arxiv.org/abs/2505.19433v2.

[226] Vicky Dong, Hao Yu, and Yao Chen. Graph-augmented relation extraction model with llms-generated support document, arXiv preprint arXiv:2410.23452, 2024. URL https://arxiv.org/abs/2410.23452v1.

[227] Xiangjue Dong, Maria Teleki, and James Caverlee. A survey on llm inference-time self-improvement, arXiv preprint arXiv:2412.14352, 2024. URL https://arxiv.org/abs/2412.14352v1.

[228] Yuxin Dong, Shuo Wang, Hongye Zheng, Jiajing Chen, Zhenhong Zhang, and Chihang Wang. Advanced rag models with graph structures: Optimizing complex knowledge reasoning and text generation. *2024 5th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC)*, 2024.

[229] Zican Dong, Junyi Li, Xin Men, Wayne Xin Zhao, Bingbing Wang, Zhen Tian, Weipeng Chen, and Ji-Rong Wen. Exploring context window of large language models via decomposed positional vectors. *Neural Information Processing Systems*, 2024.

[230] Ehsan Doostmohammadi and Marco Kuhlmann. Studying the role of input-neighbor overlap in retrieval-augmented language models training efficiency, arXiv preprint arXiv:2505.14309, 2025. URL https://arxiv.org/abs/2505.14309v1.

[231] Mohammadreza Doostmohammadian, Alireza Aghasi, Mohammad Pirani, Ehsan Nekouei, H. Zarrabi, Reza Keypour, Apostolos I. Rikos, and K. H. Johansson. Survey of distributed algorithms for resource allocation over multi-agent systems, arXiv preprint arXiv:2401.15607, 2024. URL https://arxiv.org/abs/2401.15607v1.

[232] A. Dorri, S. Kanhere, and R. Jurdak. Multi-agent systems: A survey. *IEEE Access*, 2018.

[233] Mauro Dragone. Component & service-based agent systems: Self-osgi. *International Conference on Agents and Artificial Intelligence*, 2012.

[234] Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. WorkArena: How capable are web agents at solving common knowledge work tasks? In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11642–11662. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/drouin24a.html.

[235] Hung Du, Srikanth Thudumu, Rajesh Vasa, and K. Mouzakis. A survey on context-aware multi-agent systems: Techniques, challenges and future directions, arXiv preprint arXiv:2402.01968, 2024. URL https://arxiv.org/abs/2402.01968v2.

[223] 郭廷东, 陈一飞, 李晓曦, 金佳杰, 钱宏进, 朱宇涛, 毛杭宇, 周国瑞, 莎振程, 文继荣. Tool-star: 通过强化学习赋能具有llm大脑的多工具推理器. arXiv预印本, 2025.

[224] 董凯文. 大语言模型在多智能体系统中的应用调查. 应用与计算工程, 2024.

[225] 董培杰, 唐振恒, 刘祥红, 李路军, 崔晓文, 和李波. 压缩llm真的能行动吗? 对llm压缩中代理能力的实证评估, arXiv预印本 arXiv:2505.19433, 2025. URLhttps://arxiv.org/abs/2505.19433v2.

[226] 董薇, 余浩, 和陈瑶. 基于llm生成支持文档的图增强关系抽取模型, arXiv预印本 arXiv:2410.23452, 2024. URL https://arxiv.org/abs/2410.23452v1.

[227] 向觉东, Maria Teleki, 和 James Caverlee. 关于大语言模型推理时自我改进的调查, arXiv 预印本 arXiv:2412.14352, 2024. URLhttps://arxiv.org/abs/2412.14352v1.

[228] 董宇欣, 王硕, 郑红叶, 陈嘉静, 张振红, 和 王驰航. 具有图结构的先进 RAG 模型: 优化复杂知识推理和文本生成. *2024* 年第 *5* 届计算机工程与智能通信国际 *symposium (ISCEIC)*, 2024.

[229] 董子灿, 李俊毅, 闵欣, 赵文新, 王冰冰, 田震, 陈伟鹏, 和 温继荣. 通过分解位置向量探索大语言模型的上下文窗口.神经信息处理系统, 2024.

[230] Ehsan Doostmohammadi 和 Marco Kuhlmann. 研究输入邻居重叠在检索增强语言模型训练效率中的作用, arXiv 预印本 arXiv:2505.14309, 2025. URL https://arxiv.org/abs/2505.14309v1.

[231] Mohammadreza Doostmohammadian, Alireza Aghasi, Mohammad Pirani, Ehsan Nekouei, H. Zarrabi, Reza Keypour, Apostolos I. Rikos, and K. H. Johansson. 分布式算法综述: 多智能体系统中的资源分配, arXiv 预印本 arXiv:2401.15607, 2024. URL https://arxiv.org/abs/2401.15607v1.

[232] A. Dorri, S. Kanhere, and R. Jurdak. 多智能体系统: 综述. *IEEEAccess*, 2018.

[233] Mauro Dragone. 基于组件和服务的中继系统: Self-osgi. 国际智能体与人工智能会议, 2012.

[234] Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. WorkArena: 网络智能体在解决常见知识工作任务方面的能力如何? 收录于 Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, 和 Felix Berkenkamp 编的 第*41*届国际机器学习会议论文集, 第235卷的 机器学习研究论文集, 第11642–11662页。PMLR, 2024年7月21–27日. URL https://proceedings.mlr.press/v235/drouin24a.html.

[235] Hung Du, Srikanth Thudumu, Rajesh Vasa, and K. Mouzakis. 大语言模型的上下文工程综述: 技术、挑战和未来方向, arXiv 预印本 arXiv:2402.01968, 2024. URLhttps://arxiv.org/abs/2402.01968v2.

[236] Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Çelebi, Michael Auli, Ves Stoyanov, and Alexis Conneau. Self-training improves pre-training for natural language understanding. *North American Chapter of the Association for Computational Linguistics*, 2020.

[237] Jusen Du, Weigao Sun, Disen Lan, Jiaxi Hu, and Yu Cheng. Mom: Linear sequence modeling with mixture-of-memories, arXiv preprint arXiv:2502.13685, 2025. URL https://arxiv.org/abs/2502.13685v2.

[238] Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents, arXiv preprint arXiv:2506.11763, 2025. URL https://arxiv.org/abs/2506.11763v1.

[239] Shangheng Du, Jiabao Zhao, Jinxin Shi, Zhentao Xie, Xin Jiang, Yanhong Bai, and Liang He. A survey on the optimization of large language model-based agents, arXiv preprint arXiv:2503.12434, 2025. URL https://arxiv.org/abs/2503.12434v1.

[240] Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. Perltqa: A personal long-term memory dataset for memory classification, retrieval, and synthesis in question answering, arXiv preprint arXiv:2402.16288, 2024. URL https://arxiv.org/abs/2402.16288v1.

[241] Hanqi Duan, Yao Cheng, Jianxiang Yu, and Xiang Li. Can large language models act as ensembler for multi-gnns?, arXiv preprint arXiv:2410.16822, 2024. URL https://arxiv.org/abs/2410.16822v2.

[242] Peitong Duan, Chin yi Chen, Bjoern Hartmann, and Yang Li. Visual prompting with iterative refinement for design critique generation, arXiv preprint arXiv:2412.16829, 2024. URL https://arxiv.org/abs/2412.16829v2.

[243] Brown Ebouky, A. Bartezzaghi, and Mattia Rigotti. Eliciting reasoning in language models with cognitive tools, arXiv preprint arXiv:2506.12115, 2025. URL https://arxiv.org/abs/2506.12115v1.

[244] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

[245] Candace Edwards. Hybrid context retrieval augmented generation pipeline: Llm-augmented knowledge graphs and vector database for accreditation reporting assistance, arXiv preprint arXiv:2405.15436, 2024. URL https://arxiv.org/abs/2405.15436v1.

[246] Abul Ehtesham, Aditi Singh, Gaurav Kumar Gupta, and Saket Kumar. A survey of agent interoperability protocols: Model context protocol (mcp), agent communication protocol (acp), agent-to-agent protocol (a2a), and agent network protocol (anp), arXiv preprint arXiv:2505.02279, 2025. URL https://arxiv.org/abs/2505.02279v2.

[247] Will Epperson, Gagan Bansal, Victor Dibia, Adam Fourney, Jack Gerrits, Erkang Zhu, and Saleema Amershi. Interactive debugging and steering of multi-agent ai systems. *International Conference on Human Factors in Computing Systems*, 2025.

[236] 景斐·杜, 艾德华·格拉夫, 贝利兹·古内尔, 维什拉夫·乔杜里, 奥努尔·切莱比, 迈克尔·奥利, 韦斯·斯托扬诺夫, 和 亚历克西斯·孔诺。自训练改进自然语言理解的预训练。北美学术计算语言学协会分会, 2020。

[237] 朱森·杜, 孙伟高, 兰迪森·兰, 胡嘉熙, 和 成宇。Mom：基于记忆混合的线性序列建模，arXiv 预印本 arXiv:2502.13685，2025。URLhttps://arxiv.org/abs/2502.13685v2。

[238] 杜明轩, 许本峰, 朱志伟, 王晓瑞, 和 毛振东。Deepresearch 基准：深度研究智能体的综合基准，arXiv 预印本 arXiv:2506.11763，2025。URLhttps://arxiv.org/abs/2506.11763v1。

[239] 商恒杜, 赵家宝, 石金鑫, 谢振涛, 江欣, 白彦宏, 和梁鹤. 基于大型语言模型的智能体优化的调查，arXiv 预印本 arXiv:2503.12434, 2025. URLhttps://arxiv.org/abs/2503.12434v1.

[240] 杜一鸣, 王红如, 赵正谊, 梁斌, 王宝军, 钟万军, 王泽中, 和黄锦辉. Perltqa: 用于问答中记忆分类、检索和合成的个人长期记忆数据集, arXiv 预印本 arXiv:2402.16288, 2024. URL https://arxiv.org/abs/2402.16288v1.

[241] 段汉奇, 成瑶, 余建祥, 和李翔. 大型语言模型能否作为多GNN的集成器?, arXiv 预印本 arXiv:2410.16822, 2024. URL https://arxiv.org/abs/2410.16822v2.

[242] Peitong Duan, Chin yi Chen, Bjoern Hartmann, and Yang Li. 视觉提示与迭代优化用于设计评论生成，arXiv 预印本 arXiv:2412.16829，2024。URL https://arxiv.org/abs/2412.16829v2.

[243] Brown Ebouky, A. Bartezzaghi, and Mattia Rigotti. 使用认知工具激发语言模型的推理，arXiv 预印本 arXiv:2506.12115，2025。URLhttps://arxiv.org/abs/2506.12115v1.

[244] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 从局部到全局：一种用于查询导向摘要的图 RAG 方法。 *arXiv* 预印本 *arXiv:2404.16130*, 2024。

[245] Candace Edwards. 混合上下文检索增强生成管道：Llm增强知识图谱和向量数据库用于认证报告辅助，arXiv 预印本 arXiv:2405.15436，2024。URLhttps://arxiv.org/abs/2405.15436v1.

[246] Abul Ehtesham, Aditi Singh, Gaurav Kumar Gupta, 和 Saket Kumar. 智能体互操作性协议调查：模型上下文协议 (mcp)、智能体通信协议 (acp)、智能体间协议 (a2a) 和智能体网络协议 (anp)，arXiv 预印本 arXiv:2505.02279，2025。URLhttps://arxiv.org/abs/2505.02279v2.

[247] Will Epperson, Gagan Bansal, Victor Dibia, Adam Fourney, Jack Gerrits, Erkang Zhu, 和 Saleema Amershi。多智能体人工智能系统的交互式调试和引导。国际人机交互系统会议, 2025。

[248] Lutfi Eren Erdogan, Nicholas Lee, Siddharth Jha, Sehoon Kim, Ryan Tabrizi, Suhong Moon, Coleman Hooper, G. Anumanchipalli, Kurt Keutzer, and A. Gholami. Tinyagent: Function calling at the edge. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[249] Oluwole Fagbohun, Rachel M. Harrison, and Anton Dereventsov. An empirical categorization of prompting techniques for large language models: A practitioner's guide. arXiv preprint, 2024.

[250] Kazem Faghih, Wenxiao Wang, Yize Cheng, Siddhant Bharti, Gaurang Sriramanan, S. Balasubramanian, Parsa Hosseini, and S. Feizi. Gaming tool preferences in agentic llms, arXiv preprint arXiv:2505.18135, 2025. URL https://arxiv.org/abs/2505.18135v1.

[251] Linxi (Jim) Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Neural Information Processing Systems*, 2022.

[252] Siqi Fan, Xiusheng Huang, Yiqun Yao, Xuezhi Fang, Kang Liu, Peng Han, Shuo Shang, Aixin Sun, and Yequan Wang. If an llm were a character, would it know its own story? evaluating lifelong learning in llms, arXiv preprint arXiv:2503.23514, 2025. URL https://arxiv.org/abs/2503.23514v1.

[253] Wenqi Fan, Yujuan Ding, Liang bo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. *Knowledge Discovery and Data Mining*, 2024.

[254] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding, arXiv preprint arXiv:2403.11481, 2024. URL https://arxiv.org/abs/2403.11481v2.

[255] Hongchao Fang and Pengtao Xie. An end-to-end contrastive self-supervised learning framework for language understanding. *Transactions of the Association for Computational Linguistics*, 2022.

[256] Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. Cert: Contrastive self-supervised learning for language understanding, arXiv preprint arXiv:2005.12766, 2020. URL https://arxiv.org/abs/2005.12766v2.

[257] Junfeng Fang, Zijun Yao, Ruipeng Wang, Haokai Ma, Xiang Wang, and Tat-Seng Chua. We should identify and mitigate third-party safety risks in mcp-powered agent systems, arXiv preprint arXiv:2506.13666, 2025. URL https://arxiv.org/abs/2506.13666v1.

[258] Siyuan Fang, Kaijing Ma, Tianyu Zheng, Xinrun Du, Ningxuan Lu, Ge Zhang, and Qingkun Tang. Karpa: A training-free method of adapting knowledge graph as references for large language model's reasoning path aggregation. arXiv preprint, 2024.

[259] Wei-Wen Fang, Yang Zhang, Kaizhi Qian, James Glass, and Yada Zhu. Play2prompt: Zero-shot tool instruction optimization for llm agents via tool play, arXiv preprint arXiv:2503.14432, 2025. URL https://arxiv.org/abs/2503.14432v2.

[260] Yi Fang, Dongzhe Fan, D. Zha, and Qiaoyu Tan. Gaugllm: Improving graph contrastive learning for text-attributed graphs with large language models. *Knowledge Discovery and Data Mining*, 2024.

[261] Yi Fang, Bowen Jin, Jiacheng Shen, Sirui Ding, Qiaoyu Tan, and Jiawei Han. Graphgpt-o: Synergistic multimodal comprehension and generation on graphs. arXiv preprint, 2025.

Lutfi Eren Erdogan, Nicholas Lee, Siddharth Jha, Sehoon Kim, Ryan Tabrizi, Suhong Moon, Coleman Hooper, G. Anumanchipalli, Kurt Keutzer, 和 A. Gholami. Tinyagent: 边缘函数调用.自然语言处理经验方法会议, 2024.

[249] Oluwole Fagbohun、Rachel M. Harrison和Anton Dereventsov。大型语言模型的提示技术经验性分类：实践者的指南。arXiv预印本，2024年。

[250] Kazem Faghih, Wenxiao Wang, Yize Cheng, Siddhant Bharti, Gaurang Sriramanan, S. Balasubramanian, Parsa Hosseini, and S. Feizi. 自主式大语言模型的博弈工具偏好，arXiv 预印本 arXiv:2505.18135，2025。URLhttps://arxiv.org/abs/2505.18135v1.

[251] 林溪（Jim）范，官志王，江云帆，Ajay Mandlekar，杨云丛，朱皓毅，Andrew Tang，黄德安，朱玉，Anima Anandkumar。Minedojo：构建具有互联网规模知识的开放式具身智能体。神经信息处理系统, 2022。

[252] 范思琪，黄修生，姚一群，方学志，刘康，韩鹏，尚硕，孙爱鑫，和王业权。如果llm是一个角色，它会知道自己的故事吗？评估llm中的终身学习，arXiv预印本arXiv:2503.23514，2025。URLhttps://arxiv.org/abs/2503.23514v1。

[253] 范文琪，丁宇娟，宁亮波，王时杰，李恒云，尹大卫，Chua Tat-Seng，和李庆。关于rag与llms的调查：迈向检索增强型大型语言模型。知识发现与数据挖掘, 2024。

[254] 岳帆, 马晓健, 吴如杰, 杜云涛, 李嘉琪, 高志, 和 李清. Videoagent: 一种用于视频理解的记忆增强多模态代理, arXiv 预印本 arXiv:2403.11481, 2024. URLhttps://arxiv.org/abs/2403.11481v2.

[255] 方洪超和谢鹏涛. 一种用于语言理解的端到端对比自监督学习框架. 计算语言学协会汇刊, 2022.

[256] 方洪超, 王思诚, 周梦, 丁家源, 和 谢鹏涛. Cert: 用于语言理解的对比自监督学习, arXiv 预印本 arXiv:2005.12766, 2020. URLhttps://arxiv.org/abs/2005.12766v2.

[257] 方俊峰, 姚子军, 王瑞鹏, 马浩凯, 王翔, 和 Chua Tat-Seng. 我们应该在 mcp 驱动的代理系统中识别和缓解第三方安全风险, arXiv 预印本 arXiv:2506.13666, 2025. URLhttps://arxiv.org/abs/2506.13666v1.

[258] 方思源, 马凯敬, 郑天宇, 杜新润, 陆宁轩, 张格, 唐清坤. Karpa: 一种无需训练的知识图谱作为大型语言模型推理路径聚合参考的方法. arXiv 预印本, 2024.

[259] 方伟文, 张阳, 钱凯之, James Glass, 朱亚达. Play2prompt: 通过工具玩耍进行零样本工具指令优化, 用于 llm 代理, arXiv preprint arXiv:2503.14432, 2025. URLhttps://arxiv.org/abs/2503.14432v2.

[260] 方毅, 范东哲, Zha D., 和 谭巧宇. Gaugllm: 使用大型语言模型改进文本属性图的图对比学习. 知识发现与数据挖掘, 2024.

[261] 方毅, 金 Bowen, 沈家成, 丁思睿, 谭巧宇, 和 韩家伟. Graphgpt-o: 图上的协同多模态理解和生成. arXiv 预印本, 2025.

[262] Bahare Fatemi, Jonathan J. Halcrow, and Bryan Perozzi. Talk like a graph: Encoding graphs for large language models. *International Conference on Learning Representations*, 2023.

[263] George Fatouros, Georgios Makridis, George Kousiouris, John Soldatos, A. Tsadimas, and D. Kyriazis. Towards conversational ai for human-machine collaborative mlops, arXiv preprint arXiv:2504.12477, 2025. URL https://arxiv.org/abs/2504.12477v1.

[264] M. Fauth, F. Wörgötter, and Christian Tetzlaff. Formation and maintenance of robust long-term information storage in the presence of synaptic turnover. *bioRxiv*, 2015.

[265] Zahra Fayyaz, Aya Altamimi, Sen Cheng, and Laurenz Wiskott. A model of semantic completion in generative episodic memory. *Neural Computation*, 2021.

[266] Xiang Fei, Xiawu Zheng, and Hao Feng. Mcp-zero: Proactive toolchain construction for llm agents from scratch. arXiv preprint, 2025.

[267] Philip Feldman, James R. Foulds, and Shimei Pan. Ragged edges: The double-edged sword of retrieval-augmented chatbots, arXiv preprint arXiv:2403.01193, 2024. URL https://arxiv.org/abs/2403.01193v3.

[268] Aosong Feng, Rex Ying, and L. Tassiulas. Long sequence modeling with attention tensorization: From sequence to tensor learning. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[269] Erhu Feng, Wenbo Zhou, Zibin Liu, Le Chen, Yunpeng Dong, Cheng Zhang, Yisheng Zhao, Dong Du, Zhi-Hua Zhou, Yubin Xia, and Haibo Chen. Get experience from practice: Llm agents with record & replay, arXiv preprint arXiv:2505.17716, 2025. URL https://arxiv.org/abs/2505.17716v1.

[270] Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. Retool: Reinforcement learning for strategic tool use in llms, arXiv preprint arXiv:2504.11536, 2025. URL https://arxiv.org/abs/2504.11536v2.

[271] Kaituo Feng, Changsheng Li, Xiaolu Zhang, Jun Zhou, Ye Yuan, and Guoren Wang. Keypoint-based progressive chain-of-thought distillation for llms. *International Conference on Machine Learning*, 2024.

[272] Leo Feng, Frederick Tung, Hossein Hajimirsadeghi, Y. Bengio, and M. O. Ahmed. Constant memory attention block, arXiv preprint arXiv:2306.12599, 2023. URL https://arxiv.org/abs/2306.12599v1.

[273] Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. Scalable multi-hop relational reasoning for knowledge-aware question answering. *Conference on Empirical Methods in Natural Language Processing*, 2020.

[274] Yifan Feng, Shiquan Liu, Xiangmin Han, Shaoyi Du, Zongze Wu, Han Hu, and Yue Gao. Hypergraph foundation model, arXiv preprint arXiv:2503.01203v1, 2025. URL https://arxiv.org/abs/2503.01203v1.

[275] Chrisantha Fernando, Dylan Banarse, H. Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. *International Conference on Machine Learning*, 2023.

Bahare Fatemi、Jonathan J. Halcrow和Bryan Perozzi。像图形一样说话：为大型语言模型编码图形。国际学习表征会议, 2023.

[263] 乔治·法图罗斯，乔治奥斯·马克里迪斯，乔治·科斯尤里斯，约翰·索达托斯，A·茨达马斯，和D·基里亚西斯。迈向人机协作的对话式AI，arXiv预印本arXiv:2504.12477，2025年。URLhttps://arxiv.org/abs/2504.12477v1。

[264] M. Fauth, F. Wörgötter 和 Christian Tetzlaff. 在突触更替存在的情况下形成和维持稳健的长期信息存储。*bioRxiv*, 2015.

[265] Zahra Fayyaz, Aya Altamimi, Sen Cheng, and Laurenz Wiskott. A model of semantic completion in generative episodic memory. *Neural Computation*, 2021.

[266] 向飞、郑晓武和冯浩。Mcp-zero：从零开始为llm代理构建主动工具链。arXiv预印本，2025。

[267] Philip Feldman、James R. Foulds 和 Shimei Pan。参差不齐的边缘：检索增强型聊天机器人的双刃剑，arXiv 预印本 arXiv:2403.01193，2024年。URLhttps://arxiv.org/abs/2403.01193v3。

冯澳松、邢雷和L. Tassiulas。长序列建模与注意力张量化：从序列到张量学习。自然语言处理经验方法会议，2024。

[269] 艾虎峰，周文博，刘子斌，陈乐，董云鹏，张成，赵一盛，杜东，周志华，夏宇斌，陈海波. 从实践中获取经验：具有记录与回放功能的 LLM 代理，arXiv 预印本 arXiv:2505.17716，2025. URLhttps://arxiv.org/abs/2505.17716v1.

[270] Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. Retool: 强化学习用于大型语言模型的策略工具使用, arXiv preprint arXiv:2504.11536, 2025. URLhttps://arxiv.org/abs/2504.11536v2.

[271] Kaituo Feng, Changsheng Li, Xiaolu Zhang, Jun Zhou, Ye Yuan, and Guoren Wang. 基于关键点的渐进式思维链蒸馏用于大型语言模型. 机器学习国际会议, 2024.

[272] Leo Feng, Frederick Tung, Hossein Hajimirsadeghi, Y. Bengio, and M. O. Ahmed. 恒定内存注意力块, arXiv preprint arXiv:2306.12599, 2023. URL https://arxiv.org/abs/2306.12599v1.

[273] Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, PeifengWang, Jun Yan, and Xiang Ren. 可扩展的多跳关系推理用于知识感知问答. 自然语言处理经验方法会议, 2020.

[274] Yifan Feng, Shiquan Liu, Xiangmin Han, Shaoyi Du, Zongze Wu, Han Hu, and Yue Gao. 超图基础模型, arXiv 预印本 arXiv:2503.01203v1, 2025. URLhttps://arxiv.org/abs/2503.01203v1.

[275] Chrisantha Fernando, Dylan Banarse, H. Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. *International Conference on Machine Learning*, 2023.

[276] Tharindu Fernando, Simon Denman, A. Mcfadyen, S. Sridharan, and C. Fookes. Tree memory networks for modelling long-term temporal dependencies. *Neurocomputing*, 2017.

[277] M. Ferrag, Norbert Tihanyi, and M. Debbah. From llm reasoning to autonomous ai agents: A comprehensive review, arXiv preprint arXiv:2504.19678, 2025. URL https://arxiv.org/abs/2504.19678v1.

[278] M. Ferrag, Norbert Tihanyi, and M. Debbah. Reasoning beyond limits: Advances and open problems for llms, arXiv preprint arXiv:2503.22732, 2025. URL https://arxiv.org/abs/2503.22732v1.

[279] Christopher Fifty, Dennis Duan, Ronald G. Junkins, Ehsan Amid, Jurij Leskovec, Christopher R'e, and Sebastian Thrun. Context-aware meta-learning. *International Conference on Learning Representations*, 2023.

[280] Tim Finin, Richard Fritzson, Donald P McKay, Robin McEntire, et al. Kqml-a language and protocol for knowledge and information exchange. In *13th Int. Distributed Artificial Intelligence Workshop*, pages 93–103, 1994.

[281] Chelsea Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning*, 2017.

[282] Paolo Finotelli and Francis Eustache. Mathematical modeling of human memory. *Frontiers in Psychology*, 2023.

[283] Ferdinando Fioretto, Enrico Pontelli, and W. Yeoh. Distributed constraint optimization problems and applications: A survey. *Journal of Artificial Intelligence Research*, 2016.

[284] Meire Fortunato, Melissa Tan, Ryan Faulkner, S. Hansen, Adrià Puigdomènech Badia, Gavin Buttimore, Charlie Deck, Joel Z. Leibo, and C. Blundell. Generalization of reinforcement learners with working and episodic memory. *Neural Information Processing Systems*, 2019.

[285] Samy Foudil, Claire Pleche, and E. Macaluso. Memory for spatio-temporal contextual details during the retrieval of naturalistic episodes. *Scientific Reports*, 2021.

[286] Zafeirios Fountas, Martin A Benfeghoul, Adnan Oomerjee, Fenia Christopoulou, Gerasimos Lampouras, Haitham Bou-Ammar, and Jun Wang. Human-like episodic memory for infinite context llms, arXiv preprint arXiv:2407.09450, 2024. URL https://arxiv.org/abs/2407.09450.

[287] Quentin Fournier, G. Caron, and D. Aloise. A practical survey on faster and lighter transformers. *ACM Computing Surveys*, 2021.

[288] Luca Franceschi, P. Frasconi, Saverio Salzo, Riccardo Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. *International Conference on Machine Learning*, 2018.

[289] Eduard Frankford, Daniel Crazzolara, Clemens Sauerwein, Michael Vierhauser, and Ruth Breu. Requirements for an online integrated development environment for automated programming assessment systems. *International Conference on Computer Supported Education*, 2024.

[276] Tharindu Fernando, Simon Denman, A. Mcfadyen, S. Sridharan, and C. Fookes. 树形记忆网络用于建模长期时序依赖性. *Neurocomputing*,2017.

[277] M. Ferrag, Norbert Tihanyi, and M. Debbah. 从 llm 推理到自主 AI 代理：综合综述, arXiv preprint arXiv:2504.19678, 2025. URLhttps://arxiv.org/abs/2504.19678v1.

[278] M. Ferrag, Norbert Tihanyi, and M. Debbah. 超越极限的推理：llms 的进展和开放问题, arXiv preprint arXiv:2503.22732, 2025. URL https://arxiv.org/abs/2503.22732v1.

[279] Christopher Fifty, Dennis Duan, Ronald G. Junkins, Ehsan Amid, Jurij Leskovec, Christopher R' e, and SebastianThrun. 基于上下文的元学习. 国际学习表征会议, 2023.

[280] Tim Finin, Richard Fritzson, Donald P McKay, Robin McEntire, 等. Kqml-一种用于知识和信息交换的语言和协议. 在第*13*届国际分布式人工智能研讨会, 第93-103页, 1994年.

[281] Chelsea Finn, P. Abbeel, 和 S. Levine. 用于深度网络快速适应的模型无关元学习. 机器学习国际会议, 2017年.

[282] Paolo Finotelli 和 Francis Eustache. 人类记忆的数学建模. 心理学前沿, 2023年.

[283] Ferdinando Fioretto, Enrico Pontelli, 和 W. Yeoh. 分布式约束优化问题及其应用：一项调查. 人工智能研究杂志, 2016 年 .

[284] Meire Fortunato, Melissa Tan, Ryan Faulkner, S. Hansen, Adrià Puigdomènech Badia, Gavin Buttimore, Charlie Deck, Joel Z. Leibo, 和 C. Blundell. 具备工作记忆和情景记忆的强化学习者的泛化. 神经信息处理系统, 2019年.

[285] Samy Foudil、Claire Pleche 和 E. Macaluso。在检索自然主义事件期间对时空上下文细节的记忆。*ScientificReports*, 2021。

[286] Zafeirios Fountas、Martin A Benfeghoul、Adnan Oomerjee、Fenia Christopoulou、Gerasimos Lampouras、Haitham Bou-Ammar 和 Jun Wang。无限上下文 LLM 的人类式情景记忆，arXiv 预印本 arXiv:2407.09450，2024。URLhttps://arxiv.org/abs/2407.09450。

[287] Quentin Fournier、G. Caron 和 D. Aloise。关于更快、更轻的 Transformer 的实用综述。*ACM Computing Surveys*, 2021。

[288] Luca Franceschi、P. Frasconi、Saverio Salzo、Riccardo Grazzi 和 M. Pontil。用于超参数优化和元学习的双层规划。*International Conference on MachineLearning*, 2018。

[289] Eduard Frankford, Daniel Crazzolara, Clemens Sauerwein, Michael Vierhauser, and Ruth Breu. 对自动化编程评估系统在线集成开发环境的要求. 国际计算机支持教育会议,2024.

[290] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint, 2024.

[291] Honghao Fu, Yilang Shen, Yuxuan Liu, Jingzhong Li, and Xiang Zhang. Sgcn: a multi-order neighborhood feature fusion landform classification method based on superpixel and graph convolutional network. *International Journal of Applied Earth Observation and Geoinformation*, 122:103441, 2023.

[292] Honghao Fu, Yufei Wang, Wenhan Yang, Alex C Kot, and Bihan Wen. Dp-iqa: Utilizing diffusion prior for blind image quality assessment in the wild. 2024.

[293] Honghao Fu, Hao Wang, Jing Jih Chin, and Zhiqi Shen. Brainvis: Exploring the bridge between brain and visual signals via image reconstruction. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[294] Yuchuan Fu, Xiaohan Yuan, and Dongxia Wang. Ras-eval: A comprehensive benchmark for security evaluation of llm agents in real-world environments. arXiv preprint, 2025.

[295] Zichuan Fu, Wentao Song, Yejing Wang, Xian Wu, Yefeng Zheng, Yingying Zhang, Derong Xu, Xuetao Wei, Tong Xu, and Xiangyu Zhao. Sliding window attention training for efficient large language models, arXiv preprint arXiv:2502.18845, 2025. URL https://arxiv.org/abs/2502.18845v2.

[296] Stefano Fusi. Memory capacity of neural network models, arXiv preprint arXiv:2108.07839, 2021. URL https://arxiv.org/abs/2108.07839v2.

[297] Tiantian Gan and Qiyao Sun. Rag-mcp: Mitigating prompt bloat in llm tool selection via retrieval-augmented generation. arXiv preprint, 2025.

[298] Kanishk Gandhi, Gala Stojnic, B. Lake, and M. Dillon. Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others. *Neural Information Processing Systems*, 2021.

[299] Anish Ganguli, Prabal Deb, and Debleena Banerjee. Mark: Memory augmented refinement of knowledge, arXiv preprint arXiv:2505.05177, 2025. URL https://arxiv.org/abs/2505.05177v1.

[300] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 2023.

[301] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S^3: Social-network simulation system with large language model-empowered agents, arXiv preprint arXiv:2307.14984, 2025. URL https://arxiv.org/abs/2307.14984.

[302] Hang Gao and Yongfeng Zhang. Memory sharing for large language model based agents, arXiv preprint arXiv:2404.09982, 2024. URL https://arxiv.org/abs/2404.09982v2.

[303] L Gao, A Madaan, S Zhou, and U Alon.... Pal: Program-aided language models. 2023. URL https://proceedings.mlr.press/v202/gao23f.

[304] Luyu Gao, Xueguang Ma, Jimmy J. Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. *Annual Meeting of the Association for Computational Linguistics*, 2022.

[290] 傅超友, 戴宇涵, 罗永东, 李雷, 任树槐, 张仁睿, 王志涵, 周晨宇, 沈云航, 张梦丹, 陈沛先, 李延伟, 林少辉, 赵思睿, 李科, 许通, 郑晓武, 陈恩红, 贾荣荣, 和孙行. Video-mme: 视频分析中多模态llms的首次综合评估基准. arXiv预印本, 2024.

[291] 傅宏浩, 沈一朗, 刘宇轩, 李景中, 和张翔. Sgcn: 基于超像素和图卷积网络的多阶邻域特征融合地貌分类方法. 国际地球观测与地理信息杂志, 122:103441, 2023.

[292] 傅宏浩, 王宇飞, 杨文涵, Alex C Kot, 和温比安. Dp-iqa: 在自然场景中利用扩散先验进行盲图像质量评估. 2024.

[293] 付宏浩, 王浩, Chin Jing Jih, 和 Shen Zhiqi. Brainvis: 通过图像重建探索大脑与视觉信号之间的桥梁. 在 *ICASSP 2025-2025 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*, 页面 1–5. IEEE, 2025.

[294] 付宇川, 袁晓寒, 和 王东霞. Ras-eval: 用于现实世界环境中 LLM 代理安全评估的综合基准. arXiv 预印本, 2025.

[295] 付子川, 宋文涛, 王叶晶, 吴娴, 郑叶峰, 张莹莹, 许德荣, 魏雪涛, 许桐, 和 赵祥宇. 用于高效大型语言模型的滑动窗口注意力训练, arXiv 预印本arXiv:2502.18845, 2025. URL https://arxiv.org/abs/2502.18845v2.

[296] Stefano Fusi. 神经网络模型的记忆容量，arXiv 预印本 arXiv:2108.07839，2021。URL https://arxiv.org/abs/2108.07839v2.

[297] Tiantian Gan 和 Qiyao Sun. Rag-mcp: 通过检索增强生成来缓解 LLM 工具选择中的提示膨胀。arXiv 预印本, 2025。

[298] Kanishk Gandhi, Gala Stojnic, B. Lake, 和 M. Dillon. Baby intuitions benchmark (bib): 区分他人的目标、偏好和行动。神经信息处理系统，2021。

[299] Anish Ganguli, Prabal Deb, 和 Debleena Banerjee. Mark: 基于记忆的知识增强细化，arXiv 预印本 arXiv:2505.05177，2025。URL https://arxiv.org/abs/2505.05177v1.

[300] 陈高, 兰晓冲, 李念, 袁媛, 丁景涛, 周志伦, 许峰立, 和 李勇. 大语言模型赋能的基于代理的建模与仿真: 调查与研究展望.人文社会科学传播, 2023.

[301] 陈高, 兰晓冲, 陆志宏, 毛金珠, 皮景华, 王欢东, 金德鹏, 和 李勇. S_3: 大语言模型赋能的社交网络仿真系统, arXiv 预印本 arXiv:2307.14984, 2025. URLhttps://arxiv.org/abs/2307.14984.

[302] 高航和张永峰. 基于大语言模型的代理的记忆共享, arXiv 预印本 arXiv:2404.09982, 2024. URLhttps://arxiv.org/abs/2404.09982v2.

[303] L Gao, A Madaan, S Zhou, and U Alon.... Pal: Program-aided language models. 2023. URL https://proceedings.mlr.press/v202/gao23f.

[304] 刘宇高，马学广，Jimmy J. Lin，和 Jamie Callan。无需相关性标签的精确零样本密集检索。计算语言学协会年会，2022。

[305] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. *International Conference on Machine Learning*, 2022.

[306] Shuzheng Gao, Xinjie Wen, Cuiyun Gao, Wenxuan Wang, and Michael R. Lyu. What makes good in-context demonstrations for code intelligence tasks with llms? *International Conference on Automated Software Engineering*, 2023.

[307] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *Annual Meeting of the Association for Computational Linguistics*, 2021.

[308] Weiguo Gao. Mep: Multiple kernel learning enhancing relative positional encoding length extrapolation, arXiv preprint arXiv:2403.17698, 2024. URL https://arxiv.org/abs/2403.17698v1.

[309] Xian Gao, Zongyun Zhang, Mingye Xie, Ting Liu, and Yuzhuo Fu. Graph of ai ideas: Leveraging knowledge graphs and llms for ai research idea generation, arXiv preprint arXiv:2503.08549, 2025. URL https://arxiv.org/abs/2503.08549v1.

[310] Xuanqi Gao, Siyi Xie, Juan Zhai, Shqing Ma, and Chao Shen. Mcp-radar: A multi-dimensional benchmark for evaluating tool use capabilities in large language models. arXiv preprint, 2025.

[311] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997, 2023. URL https://arxiv.org/abs/2312.10997v5.

[312] Yunfan Gao, Yun Xiong, Meng Wang, and Haofen Wang. Modular rag: Transforming rag systems into lego-like reconfigurable frameworks, arXiv preprint arXiv:2407.21059, 2024. URL https://arxiv.org/abs/2407.21059v1.

[313] Yunfan Gao, Yun Xiong, Yijie Zhong, Yuxi Bi, Ming Xue, and Haofen Wang. Synergizing rag and reasoning: A systematic review, arXiv preprint arXiv:2504.15909, 2025. URL https://arxiv.org/abs/2504.15909v2.

[314] Zhangyang Gao, Daize Dong, Cheng Tan, Jun Xia, Bozhen Hu, and Stan Z. Li. A graph is worth k words: Euclideanizing graph using pure transformer. *International Conference on Machine Learning*, 2024.

[315] Itai Gat, Idan Schwartz, and A. Schwing. Perceptual score: What data modalities does your model perceive? *Neural Information Processing Systems*, 2021.

[316] Itai Gat, Felix Kreuk, Tu Nguyen, Ann Lee, Jade Copet, Gabriel Synnaeve, Emmanuel Dupoux, and Yossi Adi. Augmentation invariant discrete representation for generative spoken language modeling. *International Workshop on Spoken Language Translation*, 2022.

[317] Tao Ge, Jing Hu, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. *International Conference on Learning Representations*, 2023.

[318] Yuyao Ge, Zhongguo Yang, Lizhe Chen, Yiming Wang, and Chengyang Li. Attack based on data: a novel perspective to attack sensitive points directly. *Cybersecurity*, 6(1):43, 2023.

[319] Yuyao Ge, Shenghua Liu, Baolong Bi, Yiwei Wang, Lingrui Mei, Wenjie Feng, Lizhe Chen, and Xueqi Cheng. Can graph descriptive order affect solving graph problems with llms? *ACL 2025*, 2024.

[305] 刘吕，阿曼·马达安，周舒言，Uri·阿隆，刘鹏飞，杨一鸣，Jamie·卡兰，以及Graham·纽比格。Pal：程序辅助语言模型。机器学习国际会议，2022。

[306] 高树正，温新杰，高翠云，王文轩，以及Michael·R·李。什么是使基于LLM的代码智能任务的良好上下文演示？自动化软件工程国际会议，2023。

[307] 高天宇，Adam·费施，以及陈丹琪。使预训练语言模型成为更好的少样本学习者。计算语言学协会年度会议，2021。

[308] 高伟国。Mep：多核学习增强相对位置编码长度外推，arXiv预印本arXiv:2403.17698，2024。URLhttps://arxiv.org/abs/2403.17698v1。

[309] Xian Gao, Zongyun Zhang, Mingye Xie, Ting Liu, and Yuzhuo Fu. 图形化ai思想：利用知识图谱和llms进行ai研究思想生成，arXiv预印本 arXiv:2503.08549，2025。URLhttps://arxiv.org/abs/2503.08549v1.

[310] Xuanqi Gao, Siyi Xie, Juan Zhai, Shqing Ma, and Chao Shen. Mcp-radar：一个用于评估大型语言模型工具使用能力的多维基准。arXiv预印本，2025。

[311] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 大型语言模型的检索增强生成：一项调查，arXiv预印本 arXiv:2312.10997，2023。URLhttps://arxiv.org/abs/2312.10997v5.

[312] 高云帆, 邢云, 王梦, 王浩帆. 模块化rag：将rag系统转换为积木般的可重构框架，arXiv预印本arXiv:2407.21059，2024年。URLhttps://arxiv.org/abs/2407.21059v1.

[313] 高云帆, 邢云, 中一杰, 毕宇曦, 薛明, 王浩帆. 融合检索与推理: 系统性综述, arXiv 预印本 arXiv:2504.15909, 2025. URLhttps://arxiv.org/abs/2504.15909v2.

张阳高，刀则东，陈谭，夏军，胡伯珍，和斯坦·Z·李。一张图值k个词：使用纯Transformer对图进行欧几里得化。机器学习国际会议，2024.

[315] Itai Gat, Idan Schwartz, 和 A. Schwing. 感知分数：您的模型感知哪些数据模态？ *Neural InformationProcessingSystems*, 2021.

[316] Itai Gat, Felix Kreuk, Tu Nguyen, Ann Lee, Jade Copet, Gabriel Synnaeve, Emmanuel Dupoux, 和 Yossi Adi. 增强不变离散表示用于生成式口语语言建模。*International Workshop on Spoken Language Translation*, 2022.

[317] Tao Ge, Jing Hu, Xun Wang, Si-Qing Chen, 和 Furu Wei. 上下文自动编码器用于大型语言模型中的上下文压缩。*International Conferenceon Learning Representations*, 2023.

[318] Yuyao Ge, Zhongguo Yang, Lizhe Chen, Yiming Wang, 和 Chengyang Li. 基于数据攻击：一种直接攻击敏感点的新视角。*Cybersecurity*, 6(1):43, 2023.

[319] Yuyao Ge, Shenghua Liu, Baolong Bi, Yiwei Wang, Lingrui Mei, Wenjie Feng, Lizhe Chen, 和 Xueqi Cheng. 图形描述顺序会影响使用 LLM 解决图形问题吗？ *ACL 2025*, 2024.

[320] Yuyao Ge, Shenghua Liu, Yiwei Wang, Lingrui Mei, Lizhe Chen, Baolong Bi, and Xueqi Cheng. Innate reasoning is not enough: In-context learning enhances reasoning large language models with less overthinking. 2025.

[321] Binzong Geng, Zhaoxin Huan, Xiaolu Zhang, Yong He, Liang Zhang, Fajie Yuan, Jun Zhou, and Linjian Mo. Breaking the length barrier: Llm-enhanced ctr prediction in long textual user behaviors. *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

[322] Hejia Geng, Boxun Xu, and Peng Li. Upar: A kantian-inspired prompting framework for enhancing large language model capabilities, arXiv preprint arXiv:2310.01441, 2023. URL https://arxiv.org/abs/2310.01441v2.

[323] Antonios Georgiou, M. Katkov, and M. Tsodyks. Retroactive interference model of forgetting. *Journal of Mathematical Neuroscience*, 2021.

[324] S. Gershman, A. Schapiro, A. Hupbach, and K. Norman. Neural context reinstatement predicts memory misattribution. *Journal of Neuroscience*, 2013.

[325] Mor Geva, R. Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *Conference on Empirical Methods in Natural Language Processing*, 2020.

[326] Mor Geva, Avi Caciularu, Ke Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *Conference on Empirical Methods in Natural Language Processing*, 2022.

[327] Arda Gezdur and J. Bhattacharjya. Innovators and transformers: enhancing supply chain employee training with an innovative application of a large language model. *International Journal of Physical Distribution & Logistics Management*, 2025.

[328] Alireza Ghafarollahi and Markus J. Buehler. Protagents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, 2024.

[329] Abdellah Ghassel, Ian Robinson, Gabriel Tanase, Hal Cooper, Bryan Thompson, Zhen Han, V. Ioannidis, Soji Adeshina, and H. Rangwala. Hierarchical lexical graph for enhanced multi-hop retrieval, arXiv preprint arXiv:2506.08074, 2025. URL https://arxiv.org/abs/2506.08074v1.

[330] S. Ghetti and S. Bunge. Neural changes underlying the development of episodic memory during middle childhood. *Developmental Cognitive Neuroscience*, 2012.

[331] D. Ghica. Function interface models for hardware compilation: Types, signatures, protocols, arXiv preprint arXiv:0907.0749, 2009. URL https://arxiv.org/abs/0907.0749v1.

[332] Tyler Giallanza, Declan Campbell, and Jonathan D. Cohen. Toward the emergence of intelligent control: Episodic generalization and optimization. *Open Mind*, 2024.

[333] In Gim, Seung seob Lee, and Lin Zhong. Asynchronous llm function calling, arXiv preprint arXiv:2412.07017, 2024. URL https://arxiv.org/abs/2412.07017v1.

[334] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory

[320] 余瑶, 刘胜华, 王一伟, 梅凌瑞, 陈立哲, 毕宝龙, 和程学祺. 天生推理不够：情境学习增强大型语言模型的推理能力，减少过度思考. 2025.

[321] 耿斌宗, 黄兆鑫, 张晓露, 何勇, 张亮, 袁发杰, 周军, 和莫林健. 突破长度障碍：长文本用户行为中的Llm增强Ctr预测.年度国际*ACM SIGIR*信息检索研究与发展会议, 2024.

[322] 耿鹤佳, 许博轩, 和李鹏. Upar：一个受康德启发提示框架，用于增强大型语言模型能力，arXiv预印本arXiv:2310.01441, 2023. URL https://arxiv.org/abs/2310.01441v2.

[323] 安东尼奥斯·乔治尤, M. Katkov, 和 M. Tsodyks. 遗忘的逆向干扰模型.数学神经科学杂志, 2021.

[324] S. Gershman, A. Schapiro, A. Hupbach, and K. Norman. 神经上下文恢复预测记忆错误归因. *Journal ofNeuroscience*, 2013.

[325] Mor Geva, R. Schuster, Jonathan Berant, and Omer Levy. Transformer 前馈层是键值记忆. *Conference onEmpiricalMethods inNatural LanguageProcessing*, 2020.

[326] Mor Geva, Avi Caciularu, Ke Wang, and Yoav Goldberg. Transformer 前馈层通过在词汇空间中促进概念来构建预测. *Conference onEmpirical Methods inNatural LanguageProcessing*, 2022.

[327] Arda Gezdur and J. Bhattacharjya. 创新者与 Transformer：通过大型语言模型创新应用增强供应链员工培训. *International Journalof PhysicalDistribution & Logistics Management*, 2025.

[328] Alireza Ghafarollahi and Markus J. Buehler. Protagents：通过大型语言模型多智能体协作结合物理和机器学习的蛋白质发现. *Digital Discovery*, 2024.

[329] Abdellah Ghassel, Ian Robinson, Gabriel Tanase, Hal Cooper, Bryan Thompson, Zhen Han, V. Ioannidis, Soji Adeshina, and H. Rangwala. 层级词汇图用于增强多跳检索，arXiv 预印本 arXiv:2506.08074，2025。URLhttps://arxiv.org/abs/2506.08074v1.

[330] S. Ghetti 和 S. Bunge. 中学阶段情景记忆发展的神经基础. 发展认知神经科学, 2012.

[331] D. Ghica. 硬件编译的功能接口模型：类型、签名、协议，arXiv 预印本 arXiv:0907.0749，2009年。URLhttps://arxiv.org/abs/0907.0749v1.

[332] Tyler Giallanza、Declan Campbell和Jonathan D. Cohen。迈向智能控制的涌现：情景泛化和优化。*Open Mind*, 2024.

[333] 李胜燮和李钟. 异步LLM函数调用, arXiv预印本 arXiv:2412.07017, 2024. URLhttps://arxiv.org/abs/2412.07017v1.

[334] 阿米莉亚·格莱泽, 内特·麦卡利斯, 玛雅·特雷巴奇, 约翰·阿斯拉尼德斯, 弗拉德·菲鲁伊, 蒂莫·埃瓦尔德斯, 马里贝斯·劳, 劳拉 · 韦丁格 , 马丁 · 查德威克 , 菲比 · 萨克 , 露西 · 坎贝尔 - 吉林厄姆 , 乔纳森 · 乌埃萨托 , 黄伯森 , 拉莫娜 · 科曼内斯库 , 杨帆, 艾比盖尔·西, 苏曼斯·达塔里, 罗里

Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements, arXiv preprint arXiv:2209.14375, 2022. URL https://arxiv.org/abs/2209.14375.

[335] D. Godden and A. Baddeley. Context-dependent memory in two natural environments: on land and underwater. arXiv preprint, 1975.

[336] Arda Goknil, Femke B. Gelderblom, Simeon Tverdal, Shukun Tokas, and Hui Song. Privacy policy analysis through prompt engineering for llms, arXiv preprint arXiv:2409.14879, 2024. URL https://arxiv.org/abs/2409.14879v1.

[337] Yaroslav Golubev, Zarina Kurbatova, E. Alomar, T. Bryksin, and Mohamed Wiem Mkaouer. One thousand and one stories: a large-scale survey of software refactoring. *ESEC/SIGSOFT FSE*, 2021.

[338] Alan M Gordon, Jesse Rissman, Roozbeh Kiani, and Anthony D Wagner. Cortical reinstatement mediates the relationship between content-specific encoding activity and subsequent recollection decisions. *Cerebral Cortex*, 2014.

[339] E. Gordon and B. Logan. Managing goals and resources in dynamic environments. arXiv preprint, 2005.

[340] Z Gou, Z Shao, Y Gong, Y Shen, and Y Yang. . . . Critic: Large language models can self-correct with tool-interactive critiquing. 2023. URL https://arxiv.org/abs/2305.11738.

[341] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. Tora: A tool-integrated reasoning agent for mathematical problem solving. *International Conference on Learning Representations*, 2023.

[342] Alex Graves, Abdel rahman Mohamed, and Geoffrey E. Hinton. Speech recognition with deep recurrent neural networks. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[343] Ekaterina Grishina, Mikhail Gorbunov, and Maxim Rakhuba. Procrustesgpt: Compressing llms with structured matrices and orthogonal transformations, arXiv preprint arXiv:2506.02818, 2025. URL https://arxiv.org/abs/2506.02818v1.

[344] Sven Gronauer and K. Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, 2021.

[345] C. Gros. Complex and adaptive dynamical systems, arXiv preprint arXiv:0807.4838, 2008. URL https://arxiv.org/abs/0807.4838v3.

[346] Albert Gu, Karan Goel, and Christopher R'e. Efficiently modeling long sequences with structured state spaces. *International Conference on Learning Representations*, 2021.

[347] Albert Gu, Ankit Gupta, Karan Goel, and Christopher Ré. On the parameterization and initialization of diagonal state space models. *Neural Information Processing Systems*, 2022.

[348] Jian Gu, Chunyang Chen, and A. Aleti. Vocabulary-defined semantics: Latent space clustering for improving in-context learning, arXiv preprint arXiv:2401.16184, 2024. URL https://arxiv.org/abs/2401.16184v6.

Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 通过目标人类判断改进对话代理的对齐，arXiv preprint arXiv:2209.14375, 2022. URLhttps://arxiv.org/abs/2209.14375.

[335] D. Godden and A. Baddeley. 在两种自然环境中依赖上下文的记忆：陆地和水下。arXiv preprint, 1975.

[336] Arda Goknil, Femke B. Gelderblom, Simeon Tverdal, Shukun Tokas, and Hui Song. 通过提示工程对大型语言模型的隐私政策分析，arXiv preprint arXiv:2409.14879, 2024. URLhttps://arxiv.org/abs/2409.14879v1.

[337] Yaroslav Golubev, Zarina Kurbatova, E. Alomar, T. Bryksin, and Mohamed Wiem Mkaouer. 一千零一个故事：一项大规模的软件重构调查。 *ESEC/SIGSOFT FSE*, 2021.

[338] Alan M Gordon, Jesse Rissman, Roozbeh Kiani, and Anthony D Wagner. 皮层重建介导了内容特定编码活动与后续回忆决策之间的关系。 *Cerebral Cortex*, 2014。

[339] E. Gordon and B. Logan. 在动态环境中管理目标和资源。 arXiv preprint, 2005。

[340] Z Gou, Z Shao, Y Gong, Y Shen, and Y Yang. . . . Critic: 大型语言模型可以通过工具交互式评论进行自我纠正。2023. URLhttps://arxiv.org/abs/2305.11738.

[341] 郭志斌, 邵志宏, 龚叶云, 沈延龙, 杨宇宇, 黄敏烈, 段南, 和陈伟珠. Tora: 一个用于数学问题求解的工具集成推理代理。国际学习表征会议, 2023.

[342] Alex Graves, Abdel rahman Mohamed, 和 Geoffrey E. Hinton. 深度循环神经网络中的语音识别。*IEEE*国际声学、语音与信号处理会议, 2013.

[343] Ekaterina Grishina, Mikhail Gorbunov, 和 Maxim Rakhuba. Procrustesgpt: 使用结构化矩阵和正交变换压缩大型语言模型，arXiv 预印本 arXiv:2506.02818，2025。URLhttps://arxiv.org/abs/2506.02818v1.

[344] Sven Gronauer 和 K. Diepold. 多智能体深度强化学习：一项调查。人工智能评论, 2021.

[345] C. Gros. 复杂和自适应动力系统，arXiv 预印本 arXiv:0807.4838，2008。URLhttps://arxiv.org/abs/0807.4838v3.

[346] Albert Gu, Karan Goel, 和 Christopher R'e. 使用结构化状态空间高效建模长序列。 学习表示国际会议, 2021。

[347] Albert Gu, Ankit Gupta, Karan Goel, 和 Christopher Ré. 对对角状态空间模型的参数化和初始化。 神经信息处理系统, 2022。

[348] Jian Gu, Chunyang Chen, 和 A. Aleti. 词汇定义语义：用于改进情境学习的潜在空间聚类，arXiv 预印本 arXiv:2401.16184，2024。URLhttps://arxiv.org/abs/2401.16184v6.

[349] Yongli Gu, Xiang Yan, Hanlin Qin, Naveed Akhtar, Shuai Yuan, Honghao Fu, Shuowen Yang, and Ajmal Mian. Hdtcnet: A hybrid-dimensional convolutional network for multivariate time series classification. *Pattern Recognition*, page 111837, 2025.

[350] Zhuohan Gu, Jiayi Yao, Kuntai Du, and Junchen Jiang. Llmsteer: Improving long-context llm inference by steering attention on reused contexts, arXiv preprint arXiv:2411.13009, 2024. URL https://arxiv.org/abs/2411.13009v2.

[351] Shengyue Guan, Haoyi Xiong, Jindong Wang, Jiang Bian, Bin Zhu, and Jian guang Lou. Evaluating llm-based agents for multi-turn conversations: A survey, arXiv preprint arXiv:2503.22458, 2025. URL https://arxiv.org/abs/2503.22458v1.

[352] Zhong Guan, Hongke Zhao, Likang Wu, Ming He, and Jianpin Fan. Langtopo: Aligning language descriptions of graphs with tokenized topological modeling, arXiv preprint arXiv:2406.13250, 2024. URL https://arxiv.org/abs/2406.13250v1.

[353] Tom Gunter, Zirui Wang, Chong Wang, Ruoming Pang, Andy Narayanan, Aonan Zhang, Bowen Zhang, Chen Chen, Chung-Cheng Chiu, David Qiu, Deepak Gopinath, Dian Ang Yap, Dong Yin, Feng Nan, Floris Weers, Guoli Yin, Haoshuo Huang, Jianyu Wang, Jiarui Lu, John Peebles, Kewei Ye, Mark Lee, Nan Du, Qibin Chen, Quentin Keunebroek, Sam Wiseman, Syd Evans, Tao Lei, Vivek Rathod, Xiang Kong, Xianzhi Du, Yanghao Li, Yongqiang Wang, Yuan Gao, Zaid Ahmed, Zhaoyang Xu, Zhiyun Lu, Al Rashid, Albin Madappally Jose, Alec Doane, Alfredo Bencomo, Allison Vanderby, Andrew Hansen, Ankur Jain, A. Anupama, Areeba Kamal, Bugu Wu, Carolina Brum, Charlie Maalouf, Chinguun Erdenebileg, Chris Dulhanty, Dominik Moritz, Doug Kang, Eduardo Jimenez, Evan Ladd, Fang Shi, Felix Bai, Frank Chu, Fred Hohman, Hadas Kotek, Hannah Gillis Coleman, Jane Li, Jeffrey P. Bigham, Jeffery Cao, Jeff Lai, Jessica Cheung, Jiulong Shan, Joe Zhou, John Li, Jun Qin, Karanjeet Singh, Karla Vega, Kelvin Zou, Laura Heckman, Lauren Gardiner, Margit Bowler, Maria Cordell, Meng Cao, Nicole Hay, Nilesh Shahdadpuri, Otto Godwin, Pranay Dighe, Pushyami Rachapudi, Ramsey Tantawi, Roman Frigg, Sam Davarnia, Sanskruti Shah, Saptarshi Guha, Sasha Sirovica, Shen Ma, Shuang Ma, Simon Wang, Sulgi Kim, Suma Jayaram, Vaishaal Shankar, Varsha Paidi, Vivek Kumar, Xin Wang, Xin Zheng, Walker Cheng, Y. Shrager, Yang Ye, Yasu Tanaka, Yihao Guo, Yun Meng, Zhaoping Luo, Ouyang Zhi, Alp Aygar, Alvin Wan, Andrew D. Walkingshaw, Tzu-Hsiang Lin, Arsalan Farooq, Brent Ramerth, Colorado Reed, Chris Bartels, Chris Chaney, David Riazati, Eric Liang Yang, Erin Feldman, Gabriel Hochstrasser, Guillaume Seguin, Irina Belousova, J. Pelemans, Karen Yang, Keivan A. Vahid, Liangliang Cao, Mahyar Najibi, Marco Zuliani, Max Horton, Minsik Cho, Nikhil Bhendawade, Patrick Dong, Piotr Maj, Pulkit Agrawal, Qi Shan, Qichen Fu, R. Poston, Sam Xu, Shuangning Liu, Sushma Rao, Tashweena Heeramun, Thomas Merth, Uday Rayala, Victor Cui, Vivek Rangarajan Sridhar, Wencong Zhang, Wenqi Zhang, Wentao Wu, Xingyu Zhou, Xinwen Liu, Yang Zhao, Yin Xia, Zhile Ren, and Zhongzheng Ren. Apple intelligence foundation language models, arXiv preprint arXiv:2407.21075, 2024. URL https://arxiv.org/abs/2407.21075v1.

[354] Jiayan Guo, Lun Du, and Hengyu Liu. Gpt4graph: Can large language models understand graph structured data ? an empirical evaluation and benchmarking, arXiv preprint arXiv:2305.15066, 2023. URL https://arxiv.org/abs/2305.15066v2.

[355] Jing Guo, Nan Li, Jianchuan Qi, Hang Yang, Ruiqiao Li, Yuzhen Feng, Si Zhang, and Ming Xu. Empowering working memory for large language model agents, arXiv preprint arXiv:2312.17259, 2024. URL https://arxiv.org/abs/2312.17259.

[349] Yongli Gu, Xiang Yan, Hanlin Qin, Naveed Akhtar, Shuai Yuan, Honghao Fu, Shuowen Yang, and Ajmal Mian. Hdtcnet: 一种用于多变量时间序列分类的混合维度卷积网络. 模式识别, 第111837页, 2025.

[350] 朱浩然，姚佳怡，杜坤泰，蒋峻辰. Llmsteer: 通过在重复上下文中引导注意力来改进长上下文llm推理，arXiv预印本arXiv:2411.13009，2024. URLhttps://arxiv.org/abs/2411.13009v2.

[351] 盛月关，黄浩毅，王金东，边江，朱斌，以及刘建光。评估基于LLM的多轮对话代理：一项调查，arXiv预印本arXiv:2503.22458，2025年。URLhttps://arxiv.org/abs/2503.22458v1。

[352] 钟冠，赵洪科，吴立康，何明，范建平. Langtopo：将图的语言描述与分词拓扑建模对齐，arXiv预印本arXiv:2406.13250，2024年. URLhttps://arxiv.org/abs/2406.13250v1.

[353] Tom Gunter, Zirui Wang, Chong Wang, Ruoming Pang, Andy Narayanan, Aonan Zhang, Bowen Zhang, Chen Chen, Chung-Cheng Chiu, David Qiu, Deepak Gopinath, Dian Ang Yap, Dong Yin, Feng Nan, Floris Weers, Guoli Yin, Haoshuo Huang, Jianyu Wang, Jiarui Lu, John Peebles, Kewei Ye, Mark Lee, Nan Du, Qibin Chen, Quentin Keunebroek, Sam Wiseman, Syd Evans, Tao Lei, Vivek Rathod, Xiang Kong, Xianzhi Du, Yanghao Li, Yongqiang Wang, Yuan Gao, Zaid Ahmed, Zhaoyang Xu, Zhiyun Lu, Al Rashid, Albin Madappally Jose, Alec Doane, Alfredo Bencomo, Allison Vanderby, Andrew Hansen, Ankur Jain, A. Anupama, Areeba Kamal, Bugu Wu, Carolina Brum, Charlie Maalouf, Chinguun Erdenebileg, Chris Dulhanty, Dominik Moritz, Doug Kang, Eduardo Jimenez, Evan Ladd, Fang Shi, Felix Bai, Frank Chu, Fred Hohman, Hadas Kotek, Hannah Gillis Coleman, Jane Li, Jeffrey P. Bigham, Jeffery Cao, Jeff Lai, Jessica Cheung, Jiulong Shan, Joe Zhou, John Li, Jun Qin, Karanjeet Singh, Karla Vega, Kelvin Zou, Laura Heckman, Lauren Gardiner, Margit Bowler, Maria Cordell, Meng Cao, Nicole Hay, Nilesh Shahdadpuri, Otto Godwin, Pranay Dighe, Pushyami Rachapudi, Ramsey Tantawi, Roman Frigg, Sam Davarnia, Sanskruti Shah, Saptarshi Guha, Sasha Sirovica, Shen Ma, Shuang Ma, Simon Wang, Sulgi Kim, Suma Jayaram, Vaishaal Shankar, Varsha Paidi, Vivek Kumar, Xin Wang, Xin Zheng, Walker Cheng, Y. Shrager, Yang Ye, Yasu Tanaka, Yihao Guo, Yun Meng, Zhaoping Luo, Ouyang Zhi, Alp Aygar, Alvin Wan, Andrew D. Walkingshaw, Tzu-Hsiang Lin, Arsalan Farooq, Brent Ramerth, Colorado Reed, Chris Bartels, Chris Chaney, David Riazati, Eric Liang Yang, Erin Feldman, Gabriel Hochstrasser, Guillaume Seguin, Irina Belousova, J. Pelemans, Karen Yang, Keivan A. Vahid, Liangliang Cao, Mahyar Najibi, Marco Zuliani, Max Horton, Minsik Cho, Nikhil Bhendawade, Patrick Dong, Piotr Maj, Pulkit Agrawal, Qi Shan, Qichen Fu, R. Poston, Sam Xu, Shuangning Liu, Sushma Rao, Tashweena Heeramun, Thomas Merth, Uday Rayala, Victor Cui, Vivek Rangarajan Sridhar, Wencong Zhang, Wenqi Zhang, Wentao Wu, Xingyu Zhou, Xinwen Liu, Yang Zhao, Yin Xia, Zhile Ren, and Zhongzheng Ren. Apple intelligence foundation language models, arXiv preprint arXiv:2407.21075, 2024. URL https://arxiv.org/abs/2407.21075v1.

[354] 郭嘉言, 杜伦, 和 刘恒宇. Gpt4graph: 大型语言模型能否理解图结构数据？实证评估与基准测试, arXiv preprint arXiv:2305.15066, 2023. URLhttps://arxiv.org/abs/2305.15066v2.

[355] 郭静, 李娜, 齐建川, 杨航, 李瑞琴, 冯宇珍, 张思, 和 许明. 赋能工作记忆以大型语言模型代理, arXiv preprintarXiv:2312.17259, 2024. URLhttps://arxiv.org/abs/2312.17259.

[356] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, N. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *International Joint Conference on Artificial Intelligence*, 2024.

[357] Xiaojun Guo, Ang Li, Yifei Wang, Stefanie Jegelka, and Yisen Wang. G1: Teaching llms to reason on graphs with reinforcement learning. *arXiv preprint arXiv:2505.18499*, 2025.

[358] Yuan Guo, Tingjia Miao, Zheng Wu, Pengzhou Cheng, Ming Zhou, and Zhuosheng Zhang. Atomic-to-compositional generalization for mobile agents with a new benchmark and scheduling system, arXiv preprint arXiv:2506.08972, 2025. URL https://arxiv.org/abs/2506.08972v1.

[359] Zhicheng Guo, Sijie Cheng, Hao Wang, Shihao Liang, Yujia Qin, Peng Li, Zhiyuan Liu, Maosong Sun, and Yang Liu. Stabletoolbench: Towards stable large-scale benchmarking on tool learning of large language models. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[360] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation, arXiv preprint arXiv:2410.05779, 2024. URL https://arxiv.org/abs/2410.05779v3.

[361] Sharut Gupta, Chenyu Wang, Yifei Wang, T. Jaakkola, and Stefanie Jegelka. In-context symmetries: Self-supervised learning through contextual world models. *Neural Information Processing Systems*, 2024.

[362] Tanmay Gupta, Luca Weihs, and Aniruddha Kembhavi. Codenav: Beyond tool-use to using real-world codebases with llm agents, arXiv preprint arXiv:2406.12276, 2024. URL https://arxiv.org/abs/2406.12276v1.

[363] I Gur, H Furuta, A Huang, M Safdari, and Y Matsuo. . . . A real-world webagent with planning, long context understanding, and program synthesis. 2023. URL https://arxiv.org/abs/2307.12856.

[364] Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, D. Eck, and Aleksandra Faust. A real-world webagent with planning, long context understanding, and program synthesis. *International Conference on Learning Representations*, 2023.

[365] Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Jordi Bayarri-Planas, Adrián Tormos, Daniel Hinjos, Pablo Bernabeu Perez, Anna Arias-Duart, Pablo A. Martin-Torres, Lucia Urcelay-Ganzabal, Marta Gonzalez-Mallo, S. Álvarez Napagao, Eduard Ayguad'e-Parra, and Ulises Cortés Dario Garcia-Gasulla. Aloe: A family of fine-tuned open healthcare llms, arXiv preprint arXiv:2405.01886, 2024. URL https://arxiv.org/abs/2405.01886v1.

[366] Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Neural Information Processing Systems*, 2024.

[367] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. *International Conference on Machine Learning*, 2020.

[368] K. Gödel, B. Meltzer, and R. Schlegel. On formally undecidable propositions of principia mathematica and related systems. arXiv preprint, 1966.

[356] 郭太成, 陈秀英, 王亚琪, 常瑞迪, 鲍时超, N. Chawla, Olaf Wiest, 张祥亮. 基于大型语言模型的多智能体: 进展与挑战的调查。国际人工智能联合会议, 2024。

[357] 郭晓军, 李昂, 王逸飞, Stefanie Jegelka, 王奕森. G1: 基于强化学习的图推理教学llms。*arXiv*预印本 *arXiv:2505.18499*, 2025。

[358] 郭园, 谭廷嘉, 吴铮, 程鹏舟, 周明, 张卓升. 移动智能体的原子到组合泛化: 新基准和调度系统, arXiv预印本 arXiv:2506.08972, 2025. URLhttps://arxiv.org/abs/2506.08972v1.

[359] 郭志成, 成思捷, 王浩, 梁时豪, 秦宇嘉, 李鹏, 刘志远, 孙毛松, 和刘杨. Stabletoolbench: 大型语言模型工具学习的稳定大规模基准测试. 计算语言学协会年会, 2024.

[360] 郭子睿, 夏良浩, 余艳华, 郝图, 和黄超. Lightrag: 简单快速的检索增强生成, arXiv 预印本 arXiv:2410.05779, 2024. URLhttps://arxiv.org/abs/2410.05779v3.

[361] Sharut Gupta, Chenyu Wang, Yifei Wang, T. Jaakkola, 和 Stefanie Jegelka. 上下文对称性: 通过上下文世界模型的自监督学习. 神经信息处理系统, 2024.

[362] Tanmay Gupta、Luca Weihs 和 Aniruddha Kembhavi. Codenav：超越工具使用，使用 LLM 代理与真实代码库，arXiv 预印本 arXiv:2406.12276，2024。URLhttps://arxiv.org/abs/2406.12276v1.

[363] I Gur, H Furuta, A Huang, M Safdari, and Y Matsuo. . . . 一个具有规划、长上下文理解和程序合成真实世界网络代理。2023。URLhttps://arxiv.org/abs/2307.12856.

Izzeddin Gur、Hiroki Furuta、Austin Huang、Mustafa Safdari、Yutaka Matsuo、D. Eck 和 Aleksandra Faust。一个具有规划、长上下文理解和程序合成的现实世界网络代理。国际学习表征会议, 2023。

[365] Ashwin Kumar Gururajan、Enrique Lopez-Cuena、Jordi Bayarri-Planas、Adrián Tormos、Daniel Hinjos、Pablo Bernabeu Perez、Anna Arias-Duart、Pablo A. Martin-Torres、Lucia Urcelay-Ganzabal、Marta Gonzalez-Mallo、S. Álvarez Napagao、Eduard Ayguad'e-Parra 和 Ulises Cortés Dario Garcia-Gasulla. Aloe：一个经过微调的开源医疗大型语言模型家族，arXiv 预印本 arXiv:2405.01886，2024年。URLhttps://arxiv.org/abs/2405.01886v1.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga 和 Yu Su. Hipporag: 基于神经生物学的大型语言模型的长期记忆. 神经信息处理系统, 2024.

[367] Kelvin Guu、Kenton Lee、Zora Tung、Panupong Pasupat 和 Ming-Wei Chang。Realm：检索增强语言模型的预训练。机器学习国际会议，2020。

[368] K.哥德尔，B.梅尔策尔，和R.施莱格尔。论数学原理及相关系统的不可判定命题。arXiv预印本，1966年。

[369] Idan Habler, Ken Huang, Vineeth Sai Narajala, and Prashant Kulkarni. Building a secure agentic ai application leveraging a2a protocol, arXiv preprint arXiv:2504.16902, 2025. URL https://arxiv.org/abs/2504.16902v2.

[370] John Halloran. Mcp safety training: Learning to refuse falsely benign mcp exploits using improved preference alignment, arXiv preprint arXiv:2505.23634, 2025. URL https://arxiv.org/abs/2505.23634v1.

[371] Tae Jun Ham, Yejin Lee, Seong Hoon Seo, Soo-Uck Kim, Hyunji Choi, Sungjun Jung, and Jae W. Lee. Elsa: Hardware-software co-design for efficient, lightweight self-attention mechanism in neural networks. *International Symposium on Computer Architecture*, 2021.

[372] Feijiang Han, Licheng Guo, Hengtao Cui, and Zhiyuan Lyu. Question tokens deserve more attention: Enhancing large language models without training through step-by-step reading and question attention recalibration, arXiv preprint arXiv:2504.09402, 2025. URL https://arxiv.org/abs/2504.09402v1.

[373] Han Han, Tong Zhu, Xiang Zhang, Mengsong Wu, Hao Xiong, and Wenliang Chen. Nestools: A dataset for evaluating nested tool learning abilities of large language models. *International Conference on Computational Linguistics*, 2024.

[374] Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A. Rossi, Subhabrata Mukherjee, Xianfeng Tang, Qi He, Zhigang Hua, Bo Long, Tong Zhao, Neil Shah, Amin Javari, Yinglong Xia, and Jiliang Tang. Retrieval-augmented generation with graphs (graphrag), arXiv preprint arXiv:2501.00309, 2025. URL https://arxiv.org/abs/2501.00309.

[375] Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyun Zhao, Shiqing Ma, and Zhenyu Chen. Token-budget-aware llm reasoning, arXiv preprint arXiv:2412.18547, 2024. URL https://arxiv.org/abs/2412.18547v5.

[376] Yuanning Han, Ziyi Qiu, Jiale Cheng, and Ray Lc. When teams embrace ai: Human collaboration strategies in generative prompting in a creative design task. *International Conference on Human Factors in Computing Systems*, 2024.

[377] R. Hankache, Kingsley Nketia Acheampong, Liang Song, Marek Brynda, Raad Khraishi, and Greig A. Cowan. Evaluating the sensitivity of llms to prior context, arXiv preprint arXiv:2506.00069, 2025. URL https://arxiv.org/abs/2506.00069v1.

[378] S Hao, T Liu, Z Wang, and Z Hu. Toolkengpt: Augmenting frozen language models with massive tools via tool embeddings. 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8fd1a81c882cd45f64958da6284f4a3f-Abstract-Conference.html.

[379] Mohanakrishnan Hariharan. Semantic mastery: Enhancing llms with advanced natural language understanding, arXiv preprint arXiv:2504.00409, 2025. URL https://arxiv.org/abs/2504.00409v1.

[380] Mareike Hartmann and Alexander Koller. A survey on complex tasks for goal-directed interactive agents, arXiv preprint arXiv:2409.18538, 2024. URL https://arxiv.org/abs/2409.18538v1.

[369] Idan Habler、Ken Huang、Vineeth Sai Narajala 和 Prashant Kulkarni。利用 a2a 协议构建安全的代理式 AI 应用，arXiv 预印本 arXiv:2504.16902，2025。URL https://arxiv.org/abs/2504.16902v2。

[370] 约翰·霍洛兰. Mcp安全培训：学习拒绝虚假的良性mcp利用，arXiv预印本arXiv:2505.23634，2025年。URLhttps://arxiv.org/abs/2505.23634v1.

{T1}韩泰俊, 李艺珍, Seong Hoon Seo, 金秀国, Choi Hyunji, Jung Sungjun, 和 Lee Jae W. Elsa：神经网络中高效、轻量级自注意力机制的系统设计。计算机体系结构国际会议, 2021.

[372] 韩飞江, 郭立成, 崔恒涛, 和吕志远. 问题标记值得更多关注：通过逐步阅读和问题注意力重新校准来增强大型语言模型而不进行训练, arXiv 预印本 arXiv:2504.09402, 2025. URLhttps://arxiv.org/abs/2504.09402v1.

[373] 韩汉, 朱通, 张翔, 吴梦松, 熊浩, 和陈文亮. Nestools：用于评估大型语言模型嵌套工具学习能力的数据集. 计算语言学国际会议, 2024.

[374] 韩浩宇, 王宇, Harry Shomer, 郭凯, 丁家源, 雷永嘉, Mahantesh Halappanavar, Ryan A.Rossi,Subhabrata Mukherjee, 唐先峰, QiHe, 华志刚, 龙波, 赵通, Neil Shah, Amin Javari, 夏英龙, 和唐继良. 基于图的检索增强生成 (graphrag), arXiv 预印本 arXiv:2501.00309, 2025. URLhttps://arxiv.org/abs/2501.00309.

[375] 韩挺旭, 王振庭, 方春荣, 赵时云, 马世庆, 和陈振宇. Token预算感知的LLM推理, arXiv预印本 arXiv:2412.18547, 2024. URLhttps://arxiv.org/abs/2412.18547v5.

[376] 韩宇宁，邱子奕，程嘉乐，和 Ray Lc. 当团队拥抱人工智能：在创意设计任务中生成式提示下的人类协作策略。国际人机交互会议, 2024.

[377] R. Hankache、Kingsley Nketia Acheampong、Liang Song、Marek Brynda、Raad Khraishi 和 Greig A. Cowan。评估大型语言模型对先验上下文的敏感性，arXiv 预印本 arXiv:2506.00069，2025 年。URL https://arxiv.org/abs/2506.00069v1。

[378] S Hao, T Liu, Z Wang, and Z Hu. Toolkengpt: 通过工具嵌入增强冻结的语言模型。2023。URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8fd1a81c882cd45f64958da6284f4a3f-Abstract-Conference.html.

[379] Mohanakrishnan Hariharan. 语义掌握：通过高级自然语言理解增强llms，arXiv预印本 arXiv:2504.00409，2025。URL https://arxiv.org/abs/2504.00409v1.

[380] Mareike Hartmann和Alexander Koller。关于目标导向交互智能体复杂任务的研究，arXiv预印本 arXiv:2409.18538，2024。URL https://arxiv.org/abs/2409.18538v1.

[381] A. Hassani, A. Medvedev, P. D. Haghighi, Sea Ling, A. Zaslavsky, and P. Jayaraman. Context definition and query language: Conceptual specification, implementation, and evaluation. *Italian National Conference on Sensors*, 2019.

[382] Kostas Hatalis, Despina Christou, Joshua Myers, Steven Jones, Keith Lambert, Adam Amos-Binks, Zohreh Dannenhauer, and Dustin Dannenhauer. Memory matters: The need to improve long-term memory in llm-agents. *Proceedings of the AAAI Symposium Series*, 2024.

[383] Kostas Hatalis, Despina Christou, and Vyshnavi Kondapalli. Review of case-based reasoning for llm agents: Theoretical foundations, architectural components, and cognitive integration, arXiv preprint arXiv:2504.06943, 2025. URL https://arxiv.org/abs/2504.06943v2.

[384] Jacky He, Guiran Liu, Binrong Zhu, Hanlu Zhang, Hongye Zheng, and Xiaokai Wang. Context-guided dynamic retrieval for improving generation quality in rag models, arXiv preprint arXiv:2504.19436, 2025. URL https://arxiv.org/abs/2504.19436v1.

[385] Jianben He, Xingbo Wang, Shiyi Liu, Guande Wu, Claudio Silva, and Huamin Qu. Poem: Interactive prompt optimization for enhancing multimodal reasoning of large language models. *IEEE Pacific Visualization Symposium*, 2024.

[386] Junqing He, Liang Zhu, Rui Wang, Xi Wang, Gholamreza Haffari, and Jiaxing Zhang. Madial-bench: Towards real-world evaluation of memory-augmented dialogue generation. *North American Chapter of the Association for Computational Linguistics*, 2024.

[387] Shawn He, Surangika Ranathunga, Stephen Cranefield, and B. Savarimuthu. Norm violation detection in multi-agent systems using large language models: A pilot study. *COINE*, 2024.

[388] Shengtao He. Achieving tool calling functionality in llms using only prompt engineering without fine-tuning, arXiv preprint arXiv:2407.04997, 2024. URL https://arxiv.org/abs/2407.04997v1.

[389] Wenchong He, Liqian Peng, Zhe Jiang, and Alex Go. You only fine-tune once: Many-shot in-context fine-tuning for large language model, arXiv preprint arXiv:2506.11103, 2025. URL https://arxiv.org/abs/2506.11103v1.

[390] Xu He, Di Wu, Yan Zhai, and Kun Sun. Sentinelagent: Graph-based anomaly detection in multi-agent systems, arXiv preprint arXiv:2505.24201, 2025. URL https://arxiv.org/abs/2505.24201v1.

[391] Yang He, Xiao Ding, Bibo Cai, Yufei Zhang, Kai Xiong, Zhouhao Sun, Bing Qin, and Ting Liu. Self-route: Automatic mode switching via capability estimation for efficient reasoning. arXiv preprint, 2025.

[392] Yu He, Yingxi Li, Colin White, and Ellen Vitercik. Dsr-bench: Evaluating the structural reasoning abilities of llms via data structures. arXiv preprint, 2025.

[393] Zexue He, Leonid Karlinsky, Donghyun Kim, Julian McAuley, Dmitry Krotov, and Rogério Feris. Camelot: Towards large language models with training-free consolidated associative memory, arXiv preprint arXiv:2402.13449, 2024. URL https://arxiv.org/abs/2402.13449v1.

[394] James B. Heald, M. Lengyel, and D. Wolpert. Contextual inference in learning and memory. *Trends in Cognitive Sciences*, 2022.

[381] A. Hassani, A. Medvedev, P. D. Haghighi, Sea Ling, A. Zaslavsky, andP. Jayaraman. 上下文定义和查询语言：概念规范、实现和评估。 意大利传感器国家会议, 2019。

[382] Kostas Hatalis, Despina Christou, Joshua Myers, Steven Jones, Keith Lambert, Adam Amos-Binks, Zohreh Dannenhauer, and Dustin Dannenhauer. 记忆很重要：改进 llm-agent 长期记忆的必要性。 *AAAI* 专题研讨会系列论文集, 2024。

[383] Kostas Hatalis, Despina Christou, 和 Vyshnavi Kondapalli. 基于案例推理的 llm-agent 评述：理论基础、架构组件和认知集成，arXiv 预印本 arXiv:2504.06943，2025。URLhttps://arxiv.org/abs/2504.06943v2。

[384] Jacky He, Guiran Liu, Binrong Zhu, Hanlu Zhang, HongyeZheng, and Xiaokai Wang. 基于上下文的动态检索以提高rag模型的生成质量，arXiv预印本 arXiv:2504.19436，2025。
URLhttps://arxiv.org/abs/2504.19436v1.

[385] Jianben He, Xingbo Wang, Shiyi Liu, Guande Wu, Claudio Silva, 和 Huamin Qu. 诗：交互式提示优化以增强大型语言模型的多模态推理。 *IEEE*太平洋可视化 *symposium*, 2024.

[386] 何俊庆, 朱亮, 王瑞, 王熙, 哈菲里·格哈姆雷扎, 张嘉翔. Madial-bench: 面向真实世界评估记忆增强对话生成. 美国计算语言学协会北美分会, 2024.

[387] Shawn He, Surangika Ranathunga, Stephen Cranefield, 和 B. Savarimuthu. 基于大型语言模型的多人智能体系统违规检测：一项初步研究。*COINE*,2024.

[388] 何胜涛. 仅通过提示工程而不进行微调在大型语言模型中实现工具调用功能, arXiv 预印本 arXiv:2407.04997, 2024. URL https://arxiv.org/abs/2407.04997v1.

[389] 何文冲, 彭丽倩, 姜哲, 和 Alex Go. 你只需微调一次：大型语言模型的许多样本上下文微调, arXiv 预印本 arXiv:2506.11103, 2025. URL https://arxiv.org/abs/2506.11103v1.

[390] 何旭, 吴迪, 齐岩, 和 孙坤. Sentinelagent：多智能体系统中的基于图的异常检测, arXiv 预印本 arXiv:2505.24201, 2025. URLhttps://arxiv.org/abs/2505.24201v1.

[391] 杨鹤, 丁晓, 蔡比博, 张宇飞, 熊凯, 孙周浩, 秦冰, 和刘婷. 自我路由：通过能力估计实现高效推理的自动模式切换. arXiv preprint, 2025.

[392] 何宇, 李颖曦, 白科尔, 和艾伦·维特里克. Dsr-bench：通过数据结构评估大语言模型的结构推理能力. arXiv preprint, 2025.

[393] 何哲雪, 列昂尼德·卡尔金斯基, 金东勋, 朱利安·麦克阿莱, 德米特里·克罗托夫, 和罗杰里奥·费里斯. Camelot：面向具有无训练整合联想记忆的大语言模型, arXiv preprint arXiv:2402.13449, 2024.
URLhttps://arxiv.org/abs/2402.13449v1.

[394] 詹姆斯·B·希尔德, M·朗吉埃尔, 和 D·沃尔珀特. 学习和记忆中的情境推理. 认知科学趋势, 2022.

[395] Shekoofeh Hedayati, Ryan E. O'Donnell, and Brad Wyble. A model of working memory for latent representations. *Nature Human Behaviour*, 2021.

[396] Tooraj Helmi. Modeling response consistency in multi-agent llm systems: A comparative analysis of shared and separate context approaches, arXiv preprint arXiv:2504.07303, 2025. URL https://arxiv.org/abs/2504.07303v1.

[397] Arshia Hemmat, Kianoosh Vadaei, Mohammad Hassan Heydari, and Afsaneh Fatemi. Leveraging retrieval-augmented generation for persian university knowledge retrieval. *Conference on Information and Knowledge Technology*, 2024.

[398] M. Herrera, Marco Pérez-Hernández, A. Kumar Parlikad, and J. Izquierdo. Multi-agent systems and complex networks: Review and applications in systems engineering. *Processes*, 2020.

[399] Nora A. Herweg, A. Sharan, M. Sperling, A. Brandt, A. Schulze-Bonhage, and M. Kahana. Reactivated spatial context guides episodic recall. *Journal of Neuroscience*, 2018.

[400] Thomas F. Heston and Charya Khun. Prompt engineering in medical education. *International Medical Education*, 2023.

[401] Dollaya Hirunyasiri, Danielle R. Thomas, Jionghao Lin, K. Koedinger, and Vincent Aleven. Comparative analysis of gpt-4 and human graders in evaluating human tutors giving praise to students. *Human-AI Math Tutoring@AIED*, 2023.

[402] Thomas Hoang. Gnn: Graph neural network and large language model for data discovery, arXiv preprint arXiv:2408.13609, 2024. URL https://arxiv.org/abs/2408.13609v2.

[403] W. Hoek and M. Wooldridge. Towards a logic of rational agency. *Logic Journal of the IGPL*, 2003.

[404] Aidan Hogan, E. Blomqvist, Michael Cochez, C. d'Amato, Gerard de Melo, C. Gutierrez, J. E. L. Gayo, S. Kirrane, S. Neumaier, A. Polleres, Roberto Navigli, A. Ngomo, S. M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. *ACM Computing Surveys*, 2020.

[405] Nithin Holla, Pushkar Mishra, H. Yannakoudakis, and Ekaterina Shutova. Meta-learning with sparse experience replay for lifelong language learning, arXiv preprint arXiv:2009.04891, 2020. URL https://arxiv.org/abs/2009.04891v2.

[406] Chuanyang Hong and Qingyun He. Enhancing memory retrieval in generative agents through llm-trained cross attention networks. *Frontiers in Psychology*, 2025.

[407] M. Hong, Sean M. Polyn, and Lisa K. Fazio. Examining the episodic context account: does retrieval practice enhance memory for context? *Cognitive Research*, 2019.

[408] Sirui Hong, Xiawu Zheng, Jonathan P. Chen, Yuheng Cheng, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Z. Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. Metagpt: Meta programming for multi-agent collaborative framework. arXiv preprint, 2023.

[409] Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. Metagpt: Meta programming for a multi-agent collaborative framework, arXiv preprint arXiv:2308.00352, 2024. URL https://arxiv.org/abs/2308.00352.

[395] Shekoofeh Hedayati, Ryan E. O' Donnell, and Brad Wyble. 一种用于潜在表示的工作记忆模型. *Nature Human Behaviour*, 2021.

[396] Tooraj Helmi. 多智能体 LLM 系统中响应一致性的建模：共享和独立上下文方法的比较分析，arXiv 预印本 arXiv:2504.07303, 2025. URLhttps://arxiv.org/abs/2504.07303v1.

[397] Arshia Hemmat, Kianoosh Vadaei, Mohammad Hassan Heydari, and Afsaneh Fatemi. 利用检索增强生成进行波斯大学知识检索. 信息与知识技术会议, 2024.

[398] M. Herrera, Marco Pérez-Hernández, A. Kumar Parlikad, and J. Izquierdo. 多智能体系统和复杂网络：系统工程中的应用综述. *Processes*, 2020.

[399] Nora A. Herweg、A. Sharan、M. Sperling、A.Brandt、A. Schulze-Bonhage和M. Kahana。重新激活的空间上下文指导情景回忆。神经科学杂志，2018。

[400] 托马斯·F·海斯顿和查里亚·昆。医学教育中的提示工程。国际医学教育, 2023.

Dollaya Hirunyasiri，Danielle R. Thomas，Jionghao Lin，K. Koedinger 和 Vincent Aleven。对 GPT-4 和人类评分员在评估人类导师对学生给予表扬方面的比较分析。《Human-AI 数学辅导@AIED》，2023。

[402] 托马斯·黄. Gnn: 图神经网络和大型语言模型用于数据发现，arXiv 预印本 arXiv:2408.13609，2024年。URLhttps://arxiv.org/abs/2408.13609v2.

[403] W. Hoek 和 M. Wooldridge. 迈向理性代理的逻辑. *LogicJournal of the IGPL*, 2003.

[404] Aidan Hogan, E. Blomqvist, Michael Cochez, C. d' Amato, Gerard de Melo, C. Gutierrez, J. E. L. Gayo, S. Kirrane, S. Neumaier, A. Polleres, Roberto Navigli, A. Ngomo, S. M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab 和 Antoine Zimmermann. 知识图谱. *ACMComputing Surveys*, 2020.

[405] Nithin Holla, Pushkar Mishra, H. Yannakoudakis 和 Ekaterina Shutova. 基于稀疏经验回放的元学习用于终身语言学习, arXiv 预印本 arXiv:2009.04891, 2020. URLhttps://arxiv.org/abs/2009.04891v2.

[406] Chuanyang Hong 和 Qingyun He. 通过 llm 训练的交叉注意力网络增强生成式代理的记忆检索。 *Frontiers inPsychology*,2025。

[407] M. Hong, Sean M. Polyn, 和 Lisa K. Fazio. 检验情景上下文解释：提取练习是否增强了上下文记忆? *CognitiveResearch*, 2019。

[408] Sirui Hong, Xiawu Zheng, Jonathan P. Chen, Yuheng Cheng, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Z. Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, 和 Chenglin Wu. Metagpt：多智能体协作框架的元编程。arXiv preprint, 2023。

[409] Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, 和 Jürgen Schmidhuber. 多智能体协作框架的元编程，arXiv preprint arXiv:2308.00352, 2024。URLhttps://arxiv.org/abs/2308.00352.

[410] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A visual language model for gui agents. *Computer Vision and Pattern Recognition*, 2023.

[411] Xiangyu Hong, Che Jiang, Biqing Qi, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. On the token distance modeling ability of higher rope attention dimension. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[412] Yubin Hong, Chaofan Li, Jingyi Zhang, and Yingxia Shao. Fg-rag: Enhancing query-focused summarization with context-aware fine-grained graph rag. arXiv preprint, 2025.

[413] Thanapapas Horsuwan, Piyawat Lertvittayakumjorn, Kasidis Kanwatchara, B. Kijsirikul, and P. Vateekul. Meta lifelong-learning with selective and task-aware adaptation. *IEEE Access*, 2024.

[414] A. N. Hoskin, A. Bornstein, K. Norman, and J. Cohen. Refresh my memory: Episodic memory reinstatements intrude on working memory maintenance. *Cognitive, Affective, & Behavioral Neuroscience*, 2017.

[415] Timothy M. Hospedales, Antreas Antoniou, P. Micaelli, and A. Storkey. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[416] Peyman Hosseini, Ignacio Castro, Iacopo Ghinassi, and Matthew Purver. Efficient solutions for an intriguing failure of llms: Long context window does not mean llms can analyze long sequences flawlessly. *International Conference on Computational Linguistics*, 2024.

[417] Haowen Hou, Fei Ma, Binwen Bai, Xinxin Zhu, and F. Yu. Enhancing and accelerating large language models via instruction-aware contextual compression, arXiv preprint arXiv:2408.15491, 2024. URL https://arxiv.org/abs/2408.15491v1.

[418] Wenjun Hou, Yi Cheng, Kaishuai Xu, Yan Hu, Wenjie Li, and Jiangming Liu. Memory-augmented multimodal llms for surgical vqa via self-contained inquiry, arXiv preprint arXiv:2411.10937v1, 2024. URL https://arxiv.org/abs/2411.10937v1.

[419] Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. Model context protocol (mcp): Landscape, security threats, and future research directions, arXiv preprint arXiv:2503.23278, 2025. URL https://arxiv.org/abs/2503.23278v2.

[420] Zejiang Hou, Julian Salazar, and George Polovets. Meta-learning the difference: Preparing large language models for efficient adaptation. *Transactions of the Association for Computational Linguistics*, 2022.

[421] N. Houlsby, A. Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and S. Gelly. Parameter-efficient transfer learning for nlp. *International Conference on Machine Learning*, 2019.

[422] Marc W Howard and M. Kahana. A distributed representation of temporal context. arXiv preprint, 2002.

[423] Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, J. Zhao, and Hang Zhao. Chatdb: Augmenting llms with databases as their symbolic memory, arXiv preprint arXiv:2306.03901, 2023. URL https://arxiv.org/abs/2306.03901v2.

[410] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A visual language model for gui agents.计算机视觉与模式识别, 2023.

[411] Xiangyu Hong, Che Jiang, Biqing Qi, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. On the token distance modeling ability of higher rope attention dimension. 自然语言处理经验方法会议, 2024.

[412] Hong Yubin, Li Chaofan, Zhang Jingyi, and Shao Yingxia. Fg-rag: 基于上下文感知的细粒度图rag增强查询聚焦摘要. arXiv预印本, 2025.

[413] Thanapapas Horsuwan、Piyawat Lertvittayakumjorn、Kasidis Kanwatchara、B. Kijsirikul 和 P. Vateekul。元终身学习与选择性和任务感知的适应。*IEEEAccess*, 2024.

[414] A. N. Hoskin, A. Bornstein, K. Norman, and J. Cohen. 刷新我的记忆：情景记忆重新陈述干扰工作记忆的维持。 认知、情感与行为神经科学, 2017。

[415] Timothy M. Hospedales, Antreas Antoniou, P. Micaelli, and A. Storkey. 神经网络中的元学习：综述。 *IEEE* 模式分析与机器智能汇刊, 2020。

[416] Peyman Hosseini, Ignacio Castro, Iacopo Ghinassi, and Matthew Purver. 针对令人着迷的 LLM 失败的高效解决方案：长上下文窗口并不意味着 LLM 能够完美分析长序列。 国际计算语言学会议, 2024。

[417] Haowen Hou, Fei Ma, Binwen Bai, Xinxin Zhu, and F. Yu. 通过指令感知的上下文压缩来增强和加速大型语言模型，arXiv 预印本 arXiv:2408.15491，2024。URLhttps://arxiv.org/abs/2408.15491v1。

[418] 侯文俊, 程毅, 许凯帅, 胡岩, 李文杰, 和刘江明. 基于自包含查询的手术问答记忆增强多模态大模型, arXiv 预印本 arXiv:2411.10937v1, 2024. URL https://arxiv.org/abs/2411.10937v1.

[419] 侯新怡, 赵岩杰, 王沈奥, 和王浩宇. 模型上下文协议 (mcp): 景象、安全威胁和未来研究方向, arXiv 预印本 arXiv:2503.23278, 2025. URLhttps://arxiv.org/abs/2503.23278v2.

[420] 侯泽江, 朱利安·萨拉查, 和乔治·波洛维茨. 元学习差异: 为高效适应准备大型语言模型. 计算语言学协会汇刊, 2022.

[421] N. 侯尔斯比, A. 吉乌里乌, 斯坦尼斯瓦夫·亚斯特热布斯基, 布鲁娜·莫罗内, 昆汀·德·拉鲁西勒, 安德烈亚·盖斯蒙多, 莫娜·阿塔里亚扬, 和 S. 盖利. 参数高效的自然语言处理迁移学习. 机器学习国际会议, 2019.

[422] Marc W Howard 和 M. Kahana. 一种时间上下文的分布式表示. arXiv 预印本, 2002.

[423] Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, J. Zhao, 和 Hang Zhao. Chatdb: 将数据库作为其符号内存增强 llms, arXiv 预印本 arXiv:2306.03901, 2023. URL https://arxiv.org/abs/2306.03901v2.

[424] J. E. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2021.

[425] Junhao Hu, Wenrui Huang, Weidong Wang, Zhenwen Li, Tiancheng Hu, Zhixia Liu, XuSheng Chen, Tao Xie, and Yizhou Shan. Raas: Reasoning-aware attention sparsity for efficient llm reasoning, arXiv preprint arXiv:2502.11147, 2025. URL https://arxiv.org/abs/2502.11147v2.

[426] Junwei Hu, Weicheng Zheng, Yan Liu, and Yihan Liu. Optimizing token consumption in llms: A nano surge approach for code reasoning efficiency, arXiv preprint arXiv:2504.15989, 2025. URL https://arxiv.org/abs/2504.15989v2.

[427] Junyan Hu, Hanlin Niu, J. Carrasco, B. Lennox, and F. Arvin. Voronoi-based multi-robot autonomous exploration in unknown environments via deep reinforcement learning. *IEEE Transactions on Vehicular Technology*, 2020.

[428] Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[429] Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model, arXiv preprint arXiv:2408.09559, 2024. URL https://arxiv.org/abs/2408.09559.

[430] Nathan J. Hu, E. Mitchell, Christopher D. Manning, and Chelsea Finn. Meta-learning online adaptation of language models. *Conference on Empirical Methods in Natural Language Processing*, 2023.

[431] Shengxiang Hu, Guobing Zou, Song Yang, Yanglan Gan, Bofeng Zhang, and Yixin Chen. Large language model meets graph neural network in knowledge distillation. *AAAI Conference on Artificial Intelligence*, 2024.

[432] Siyuan Hu, Mingyu Ouyang, Difei Gao, and Mike Zheng Shou. The dawn of gui agent: A preliminary case study with claude 3.5 computer use. arXiv preprint, 2024.

[433] Ting Hu, Christoph Meinel, and Haojin Yang. Scaled prompt-tuning for few-shot natural language generation, arXiv preprint arXiv:2309.06759, 2023. URL https://arxiv.org/abs/2309.06759v1.

[434] Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas, arXiv preprint arXiv:2410.14255, 2024. URL https://arxiv.org/abs/2410.14255v2.

[435] Yuntong Hu, Zhengwu Zhang, and Liang Zhao. Beyond text: A deep dive into large language models' ability on understanding graph data, arXiv preprint arXiv:2310.04944, 2023. URL https://arxiv.org/abs/2310.04944v1.

[436] Yilun Hua and Yoav Artzi. Talk less, interact better: Evaluating in-context conversational adaptation in multimodal llms, arXiv preprint arXiv:2408.01417v1, 2024. URL https://arxiv.org/abs/2408.01417v1.

[424] J. E. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: 大型语言模型的低秩适配. 国际学习表征会议, 2021.

[425] Junhao Hu, Wenrui Huang, Weidong Wang, Zhenwen Li, Tiancheng Hu, Zhixia Liu, XuSheng Chen, Tao Xie, and Yizhou Shan. Raas: 用于高效 LLM 推理的推理感知注意力稀疏性, arXiv 预印本 arXiv:2502.11147, 2025. URLhttps://arxiv.org/abs/2502.11147v2.

[426] Junwei Hu, Weicheng Zheng, Yan Liu, and Yihan Liu. 优化 LLM 中的 token 消耗：用于代码推理效率的纳米脉冲方法, arXiv 预印本 arXiv:2504.15989, 2025. URLhttps://arxiv.org/abs/2504.15989v2.

[427] 胡俊岩, 牛汉林, J. Carrasco, B. Lennox, 和 F. Arvin. 基于Voronoi的多机器人自主探索未知环境通过深度强化学习. *IEEE Transactions onVehicular Technology*, 2020.

[428] 胡林梅, 刘泽艺, 赵子旺, 侯雷, 聂立强, 和 李娟子. 知识增强预训练语言模型综述. *IEEE Transactions on Knowledge andData Engineering*, 2022.

[429] 胡梦康, 陈天行, 陈启光, 姜耀, 邵文琪, 和 罗平. Hiagent: 大型语言模型用于解决长时程智能体任务的分层工作内存管理, arXiv预印本 arXiv:2408.09559, 2024. URLhttps://arxiv.org/abs/2408.09559.

[430] 胡纳森 J. Hu, E. Mitchell, Christopher D. Manning, 和 Chelsea Finn. 在线元学习语言模型的适应. 自然语言处理经验方法会议, 2023.

[431] 胡胜祥, 邹国兵, 杨宋, 甘阳兰, 张伯峰, 和陈奕欣. 大语言模型在知识蒸馏中与图神经网络相遇. *AAAI*人工智能会议, 2024.

[432] 胡思源, 欧阳明宇, 高迪飞, 和郑书铭. 图智能体的黎明：基于claude 3.5计算机使用的初步案例研究. arXiv预印本, 2024.

[433] 胡婷, Christoph Meinel, 和杨浩金. 针对少样本自然语言生成的扩展提示微调, arXiv预印本 arXiv:2309.06759, 2023. URLhttps://arxiv.org/abs/2309.06759v1.

[434] 胡翔, 傅红宇, 王景, 王奕峰, 李志坤, 许仁军, 陆宇, 金耀初, 潘丽丽, 和蓝振中. Nova：一种迭代规划和搜索方法，用于增强大语言模型生成想法的新颖性和多样性, arXiv预印本 arXiv:2410.14255, 2024. URLhttps://arxiv.org/abs/2410.14255v2.

[435] 云通胡，张正武，赵亮。超越文本：深入探讨大型语言模型对图数据理解的能力，arXiv预印本 arXiv:2310.04944，2023。URLhttps://arxiv.org/abs/2310.04944v1。

[436] Yilun Hua 和 Yoav Artzi. 少说多互动：评估多模态大语言模型中的情境对话适应，arXiv 预印本 arXiv:2408.01417v1，2024。URL https://arxiv.org/abs/2408.01417v1.

[437] Brandon Huang, Chancharik Mitra, Assaf Arbelle, Leonid Karlinsky, Trevor Darrell, and Roei Herzig. Multimodal task vectors enable many-shot multimodal in-context learning. *Neural Information Processing Systems*, 2024.

[438] Chengkai Huang, Hongtao Huang, Tong Yu, Kaige Xie, Junda Wu, Shuai Zhang, Julian J. McAuley, Dietmar Jannach, and Lina Yao. A survey of foundation model-powered recommender systems: From feature-based, generative to agentic paradigms, arXiv preprint arXiv:2504.16420, 2025. URL https://arxiv.org/abs/2504.16420v1.

[439] Chengkai Huang, Junda Wu, Yu Xia, Zixu Yu, Ruhan Wang, Tong Yu, Ruiyi Zhang, Ryan A. Rossi, B. Kveton, Dongruo Zhou, Julian J. McAuley, and Lina Yao. Towards agentic recommender systems in the era of multimodal large language models, arXiv preprint arXiv:2503.16734, 2025. URL https://arxiv.org/abs/2503.16734v1.

[440] Chengrui Huang, Shen Gao, Zhengliang Shi, Dongsheng Wang, and Shuo Shang. Ttpa: Token-level tool-use preference alignment training framework with fine-grained evaluation, arXiv preprint arXiv:2505.20016, 2025. URL https://arxiv.org/abs/2505.20016v1.

[441] Chensen Huang, Guibo Zhu, Xuepeng Wang, Yifei Luo, Guojing Ge, Haoran Chen, Dong Yi, and Jinqiao Wang. Recurrent context compression: Efficiently expanding the context window of llm, arXiv preprint arXiv:2406.06110, 2024. URL https://arxiv.org/abs/2406.06110v1.

[442] Jing Huang, X. Ruan, Naigong Yu, Qingwu Fan, Jiaming Li, and Jianxian Cai. A cognitive model based on neuromodulated plasticity. *Computational Intelligence and Neuroscience*, 2016.

[443] Ken Huang, Akram Sheriff, Vineeth Sai Narajala, and Idan Habler. Agent capability negotiation and binding protocol (acnbp), arXiv preprint arXiv:2506.13590, 2025. URL https://arxiv.org/abs/2506.13590v1.

[444] Le Huang, Hengzhi Lan, Zijun Sun, Chuan Shi, and Ting Bai. Emotional rag: Enhancing role-playing agents through emotional retrieval. *2024 IEEE International Conference on Knowledge Graph (ICKG)*, 2024.

[445] Lisheng Huang, Yichen Liu, Jinhao Jiang, Rongxiang Zhang, Jiahao Yan, Junyi Li, and Wayne Xin Zhao. Manusearch: Democratizing deep search in large language models with a transparent and open multi-agent framework, arXiv preprint arXiv:2505.18105, 2025. URL https://arxiv.org/abs/2505.18105v1.

[446] Shiting Huang, Zhen Fang, Zehui Chen, Siyu Yuan, Junjie Ye, Yu Zeng, Lin Chen, Qi Mao, and Feng Zhao. Critictool: Evaluating self-critique capabilities of large language models in tool-calling error scenarios, arXiv preprint arXiv:2506.13977, 2025. URL https://arxiv.org/abs/2506.13977v1.

[447] Sirui Huang, Yanggan Gu, Xuming Hu, Zhonghao Li, Qing Li, and Guandong Xu. Reasoning factual knowledge in structured data with large language models, arXiv preprint arXiv:2408.12188, 2024. URL https://arxiv.org/abs/2408.12188v1.

[448] Sirui Huang, Hanqian Li, Yanggan Gu, Xuming Hu, Qing Li, and Guandong Xu. Hyperg: Hypergraph-enhanced llms for structured knowledge, arXiv preprint arXiv:2502.18125, 2025. URL https://arxiv.org/abs/2502.18125v1.

[437] 黄 Brandon, Mitra Chancharik, Arbelle Assaf, Karlinsky Leonid, Darrell Trevor 和 Herzig Roei. 多模态任务向量支持多示例多模态情境学习。 神经信息处理系统, 2024。

[438] 黄 Chengkai, 黄 Hongtao, 余 Tong, 谢 Kaige, 吴 Junda, 张 Shuai, McAuley Julian J., Jannach Dietmar, 和 姚Lina. 基于基础模型推荐系统的调查：从基于特征、生成到代理范式，arXiv 预印本 arXiv:2504.16420，2025。URLhttps://arxiv.org/abs/2504.16420v1.

[439] Chengkai Huang, Junda Wu, Yu Xia, Zixu Yu, Ruhan Wang, Tong Yu, Ruiyi Zhang, Ryan A. Rossi, B. Kveton, Dongruo Zhou, Julian J. McAuley, and Lina Yao. Towards agentic recommender systems in the era of multimodal large language models, arXiv preprint arXiv:2503.16734, 2025. URL https://arxiv.org/abs/2503.16734v1.

[440] 黄 Chengrui, 高 Shen, 石 Zhengliang, 王 Dongsheng, 和 上 Shuo. Ttpa：基于 token 级工具使用偏好对齐训练框架，具有细粒度评估，arXiv 预印本 arXiv:2505.20016，2025。URLhttps://arxiv.org/abs/2505.20016v1.

[441] 黄晨森, 朱贵波, 王雪鹏, 罗逸飞, 葛国敬, 陈浩然, 易东, 王金桥. 循环上下文压缩：高效扩展llm的上下文窗口, arXiv预印本arXiv:2406.06110, 2024. URLhttps://arxiv.org/abs/2406.06110v1.

[442] 黄静, 阮雪, 余乃功, 范庆武, 李佳明, 和蔡建贤. 基于神经调节可塑性的认知模型. 计算智能与神经科学, 2016.

[443] 黄肯, 阿克兰·谢里夫, Vineeth Sai Narajala, 和Idan Habler. 代理能力协商和绑定协议 (acnbp), arXiv预印本arXiv:2506.13590, 2025. URL https://arxiv.org/abs/2506.13590v1.

[444] 刘黄, 兰恒志, 孙子军, 石川, 和 白婷. 情感编织: 通过情感检索增强角色扮演代理. *2024 IEEE* 国际知识图谱会议 *(ICKG)*, 2024.

[445] 黄立升, 刘奕辰, 姜金浩, 张荣祥, 颜嘉豪, 李俊毅, 和 赵新伟. Manusearch: 通过透明和开放的多代理框架在大语言模型中实现深度搜索的民主化, arXiv 预印本 arXiv:2505.18105, 2025. URLhttps://arxiv.org/abs/2505.18105v1.

[446] 黄石庭, 方振, 陈哲辉, 袁思宇, 叶俊杰, 曾宇, 陈林, 毛琪, 和 赵峰. Critictool: 在工具调用错误场景中评估大语言模型的自批评能力, arXiv 预印本 arXiv:2506.13977, 2025. URLhttps://arxiv.org/abs/2506.13977v1.

[447] 黄思睿, 古杨甘, 胡旭明, 李忠浩, 李清, 和徐冠东. 基于大型语言模型的结构化数据推理事实性知识, arXiv preprint arXiv:2408.12188, 2024. URL https://arxiv.org/abs/2408.12188v1.

[448] 黄思睿, 李汉倩, 古杨甘, 胡旭明, 李清, 和徐冠东. Hyperg: 基于超图的增强型大型语言模型用于结构化知识, arXiv preprint arXiv:2502.18125, 2025. URL https://arxiv.org/abs/2502.18125v1.

[449] Wenlong Huang, P. Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *International Conference on Machine Learning*, 2022.

[450] Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. Recommender ai agent: Integrating large language models for interactive recommendations, arXiv preprint arXiv:2308.16505, 2024. URL https://arxiv.org/abs/2308.16505.

[451] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey, arXiv preprint arXiv:2402.02716, 2024. URL https://arxiv.org/abs/2402.02716v1.

[452] Y Huang, J Shi, Y Li, C Fan, S Wu, and Q Zhang.... Metatool benchmark for large language models: Deciding whether to use tools and which to use. 2023. URL https://arxiv.org/abs/2310.03128.

[453] Yunpeng Huang, Jingwei Xu, Zixu Jiang, Junyu Lai, Zenan Li, Yuan Yao, Taolue Chen, Lijuan Yang, Zhou Xin, and Xiaoxing Ma. Advancing transformer architecture in long-context large language models: A comprehensive survey, arXiv preprint arXiv:2311.12351, 2023. URL https://arxiv.org/abs/2311.12351v2.

[454] Zeyi Huang, Yuyang Ji, Anirudh Sundara Rajan, Zefan Cai, Wen Xiao, Junjie Hu, and Yong Jae Lee. Visualtoolagent (vista): A reinforcement learning framework for visual tool selection, arXiv preprint arXiv:2505.20289, 2025. URL https://arxiv.org/abs/2505.20289v1.

[455] Ziheng Huang, S. Gutierrez, Hemanth Kamana, and S. Macneil. Memory sandbox: Transparent and interactive memory management for conversational agents. *ACM Symposium on User Interface Software and Technology*, 2023.

[456] Ziheng Huang, Sebastian Gutierrez, Hemanth Kamana, and Stephen MacNeil. Memory sandbox: Transparent and interactive memory management for conversational agents, arXiv preprint arXiv:2308.01542, 2023. URL https://arxiv.org/abs/2308.01542.

[457] Alexis Huet, Zied Ben-Houidi, and Dario Rossi. Episodic memories generation and evaluation benchmark for large language models. *International Conference on Learning Representations*, 2025.

[458] Dom Huh and Prasant Mohapatra. Multi-agent reinforcement learning: A comprehensive survey, arXiv preprint arXiv:2312.10256, 2023. URL https://arxiv.org/abs/2312.10256v2.

[459] Eunjeong Hwang, Yichao Zhou, James Bradley Wendt, Beliz Gunel, Nguyen Vo, Jing Xie, and Sandeep Tata. Enhancing incremental summarization with structured representations. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[460] Thorsten Händler. Balancing autonomy and alignment: A multi-dimensional taxonomy for autonomous llm-powered multi-agent architectures. arXiv preprint, 2023.

[461] Michael Iannelli, Sneha Kuchipudi, and Vera Dvorak. Sla management in reconfigurable multi-agent rag: A systems approach to question answering, arXiv preprint arXiv:2412.06832, 2024. URL https://arxiv.org/abs/2412.06832v2.

[462] IBM. What is agent communication protocol (acp)? https://www.ibm.com/think/topics/agent-communication-protocol, 2025. [Online; accessed 17-July-2025].

[449] 黄文龙, P. Abbeel, Deepak Pathak, 和 Igor Mordatch. 语言模型作为零样本规划器：为具身智能体提取可执行知识. 国际机器学习会议, 2022.

[450] 黄旭, 连建勋, 雷雨轩, 姚静, 连德福, 和 谢行. 推荐人工智能智能体：整合大型语言模型进行交互式推荐, arXiv 预印本 arXiv:2308.16505, 2024. URLhttps://arxiv.org/abs/2308.16505.

[451] 黄旭, 刘伟文, 陈晓龙, 王兴梅, 王浩, 连德福, 王亚胜, 唐瑞明, 和 陈恩红. 理解 LLM 智能体的规划：调查, arXiv 预印本 arXiv:2402.02716, 2024. URLhttps://arxiv.org/abs/2402.02716v1.

[452] 黄勇, 石杰, 李岩, 范晨, 吴森, 张强. . . . 大型语言模型的元工具基准：决定是否使用工具以及使用哪些工具. 2023. URL https://arxiv.org/abs/2310.03128.

[453] 黄云鹏, 许景伟, 姜子旭, 赖俊宇, 李泽楠, 姚远, 陈涛路, 杨丽娟, 肖欣, 马晓星. 推进长上下文大语言模型中的 Transformer 架构：一项综合调查, arXiv 预印本 arXiv:2311.12351, 2023. URL https://arxiv.org/abs/2311.12351v2.

[454] 黄泽怡, 李宇阳, Anirudh Sundara Rajan, 蔡泽帆, 肖文, 胡俊杰, 李永才. Visualtoolagent (vista): 一种用于视觉工具选择的强化学习框架, arXiv 预印本 arXiv:2505.20289, 2025. URLhttps://arxiv.org/abs/2505.20289v1.

[455] 黄志恒, S. Gutierrez, Hemanth Kamana, 和 S. Macneil. 内存沙盒：对话代理的透明和交互式内存管理. *ACM* 用户界面软件与技术研讨会, 2023.

[456] 黄志恒, Sebastian Gutierrez, Hemanth Kamana, 和 Stephen MacNeil. 内存沙盒：对话代理的透明和交互式内存管理, arXiv 预印本 arXiv:2308.01542, 2023. URLhttps://arxiv.org/abs/2308.01542.

[457] Alexis Huet, Zied Ben-Houidi, and Dario Rossi. Episodic memories generation and evaluation benchmark for large language models. *International Conference on Learning Representations*, 2025.

[458] Dom Huh 和 Prasant Mohapatra. 多智能体强化学习：一项综合调查, arXiv 预印本 arXiv:2312.10256, 2023. URLhttps://arxiv.org/abs/2312.10256v2.

[459] EunjeongHwang, YichaoZhou, James BradleyWendt, Beliz Gunel, Nguyen Vo, Jing Xie, 和 Sandeep Tata. 使用结构化表示增强增量式摘要. 自然语言处理经验方法会议, 2024.

[460] Thorsten Händler. 平衡自主性与一致性：面向自主 LLM 驱动的多智能体架构的多维度分类法。arXiv 预印本, 2023年。

[461] Michael Iannelli, Sneha Kuchipudi, 和 Vera Dvorak. 可重构多智能体问答中的 SLA 管理：面向问题回答的系统方法，arXiv 预印本 arXiv:2412.06832，2024年。URLhttps://arxiv.org/abs/2412.06832v2。

[462] IBM. 什么是智能体通信协议 (ACP)? https://www.ibm.com/think/topics/agent-communication-protocol, 2025年。 [在线；访问于 17-July-2025]。

[463] T Inaba, H Kiyomaru, F Cheng, and S Kurohashi. Multitool-cot: Gpt-3 can use multiple external tools with chain of thought prompting. 2023. URL https://arxiv.org/abs/2305.16896.

[464] G. Indiveri and Shih-Chii Liu. Memory and information processing in neuromorphic systems. *Proceedings of the IEEE*, 2015.

[465] V. Ioannidis, Xiang Song, Da Zheng, Houyu Zhang, Jun Ma, Yi Xu, Belinda Zeng, Trishul M. Chilimbi, and G. Karypis. Efficient and effective training of language and graph neural network models, arXiv preprint arXiv:2206.10781, 2022. URL https://arxiv.org/abs/2206.10781v1.

[466] Yoichi Ishibashi, Taro Yano, and M. Oyamada. Can large language models invent algorithms to improve themselves?: Algorithm discovery for recursive self-improvement through reinforcement learning, arXiv preprint arXiv:2410.15639, 2024. URL https://arxiv.org/abs/2410.15639v5.

[467] Shadi Iskander, Nachshon Cohen, Zohar S. Karnin, Ori Shapira, and Sofia Tolmach. Quality matters: Evaluating synthetic data for tool-using llms. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[468] Z. Ismail and N. Sariff. A survey and analysis of cooperative multi-agent robot systems: Challenges and directions. *Applications of Mobile Robots*, 2018.

[469] Yusuf Izmirlioglu, Loc Pham, Tran Cao Son, and Enrico Pontelli. A survey of multi-agent systems for smartgrids. *Energies*, 2024.

[470] Jace.AI. Jace.ai web agent, 2024. URL https://www.jace.ai/. Accessed: 2025-07-14.

[471] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Neural Information Processing Systems*, 2018.

[472] Tejas Jade and Alex Yartsev. Chatgpt for automated grading of short answer questions in mechanical ventilation, arXiv preprint arXiv:2505.04645, 2025. URL https://arxiv.org/abs/2505.04645v1.

[473] A. Jafarpour, L. Fuentemilla, A. Horner, W. Penny, and E. Duzel. Replay of very early encoding representations during recollection. *Journal of Neuroscience*, 2014.

[474] A. Jaiswal, Nurendra Choudhary, Ravinarayana Adkathimar, M. P. Alagappan, G. Hiranandani, Ying Ding, Zhangyang Wang, E-Wen Huang, and Karthik Subbian. All against some: Efficient integration of large language models for message passing in graph neural networks, arXiv preprint arXiv:2407.14996, 2024. URL https://arxiv.org/abs/2407.14996v1.

[475] H. Jaleel, Jane J. Stephan, and Sinan Naji. Multi-agent systems: A review study. *Ibn AL- Haitham Journal For Pure and Applied Sciences*, 2020.

[476] Lawrence Jang, Yinheng Li, Charles Ding, Justin Lin, Paul Pu Liang, Dan Zhao, Rogerio Bonatti, and K. Koishida. Videowebarena: Evaluating long context multimodal agents with video understanding web tasks. *International Conference on Learning Representations*, 2024.

[477] R. Janik. Aspects of human memory and large language models, arXiv preprint arXiv:2311.03839, 2023. URL https://arxiv.org/abs/2311.03839v3.

[478] Shumaila Javaid, Hamza Fahim, Bin He, and Nasir Saeed. Large language models for uavs: Current state and pathways to the future. *IEEE Open Journal of Vehicular Technology*, 2024.

[463] T Inaba, H Kiyomaru, F Cheng, and S Kurohashi. Multitool-cot: GPT-3 can use multiple external tools with chain of thought prompting. 2023. URLhttps://arxiv.org/abs/2305.16896.

[464] G. IndiveriandShih-ChiiLiu. 神经形态系统中的记忆与信息处理。*IEEE*汇刊, 2015.

[465] V. Ioannidis, Xiang Song, Da Zheng, Houyu Zhang, Jun Ma, Yi Xu, Belinda Zeng, Trishul M. Chilimbi, and G. Karypis. 高效且有效的语言和图神经网络模型训练，arXiv 预印本 arXiv:2206.10781，2022。URLhttps://arxiv.org/abs/2206.10781v1.

[466] 市桥芳一、矢野太郎和Oyamada M. 大型语言模型能否发明算法来提升自身？通过强化学习实现递归自我提升的算法发现，arXiv预印本arXiv:2410.15639，2024年。URLhttps://arxiv.org/abs/2410.15639v5。

[467] Shadi Iskander, Nachshon Cohen, Zohar S. Karnin, Ori Shapira, 和 Sofia Tolmach. 质量很重要：评估用于工具使用型大语言模型的合成数据。 自然语言处理经验方法会议, 2024。

[468] Z. Ismail and N. Sariff. A survey and analysis of cooperative multi-agent robot systems: Challenges and directions. *Applications of Mobile Robots*, 2018.

[469] Yusuf Izmirlioglu, Loc Pham, Tran Cao Son, 和 Enrico Pontelli. 智能电网的多智能体系统综述。 *Energies*, 2024。

[470] Jace.AI. Jace.ai 网络代理, 2024. URLhttps://www.jace.ai/. 访问时间：2025-07-14 .

[471] Arthur Jacot, Franck Gabriel, 和 Clément Hongler. 神经切线核：神经网络中的收敛和泛化。 神经信息处理系统, 2018。

[472] Tejas Jade 和 Alex Yartsev. 用于机械通气简答题自动评分的 ChatGPT，arXiv 预印本 arXiv:2505.04645, 2025. URL https://arxiv.org/abs/2505.04645v1.

[473] A. Jafarpour, L. Fuentemilla, A. Horner, W. Penny, and E. Duzel. 回忆期间重放非常早期的编码表征. *Journal ofNeuroscience*, 2014.

[474] A. Jaiswal, Nurendra Choudhary, Ravinarayana Adkathimar, M. P. Alagappan, G. Hiranandani, Ying Ding, Zhangyang Wang, E-Wen Huang, and Karthik Subbian. 全对部分：用于图神经网络中消息传递的高效大型语言模型集成, arXiv 预印本 arXiv:2407.14996, 2024. URLhttps://arxiv.org/abs/2407.14996v1.

[475] H. Jaleel, Jane J. Stephan, and Sinan Naji. 多智能体系统：一项综述研究. *Ibn AL- HaithamJournal For Pureand Applied Sciences*, 2020.

[476] Lawrence Jang, Yinheng Li, Charles Ding, Justin Lin, Paul Pu Liang, Dan Zhao, Rogerio Bonatti 和 K. Koishida. Videowebarena: 使用视频理解网络任务评估长上下文多模态智能体。国际学习表征会议, 2024.

[477] R. Janik. 人类记忆与大语言模型，arXiv 预印本 arXiv:2311.03839，2023。URLhttps://arxiv.org/abs/2311.03839v3.

[478] Shumaila Javaid、Hamza Fahim、Bin He 和 Nasir Saeed。无人机的大型语言模型：当前状态和通往未来的途径。《IEEE 交通运输技术开放期刊》，2024年。

[479] T. S. Jayram, Younes Bouhadjar, Ryan L. McAvoy, Tomasz Kornuta, Alexis Asseman, K. Rocki, and A. Ozcan. Learning to remember, forget and ignore using attention control in memory, arXiv preprint arXiv:1809.11087, 2018. URL https://arxiv.org/abs/1809.11087v1.

[480] Cheonsu Jeong. A study on the mcp x a2a framework for enhancing interoperability of llm-based autonomous agents, arXiv preprint arXiv:2506.01804, 2025. URL https://arxiv.org/abs/2506.01804v2.

[481] Gang Ji and J. Bilmes. Multi-speaker language modeling. *North American Chapter of the Association for Computational Linguistics*, 2004.

[482] Ke Ji, Junying Chen, Anningzhe Gao, Wenya Xie, Xiang Wan, and Benyou Wang. Llms could autonomously learn without external supervision, arXiv preprint arXiv:2406.00606, 2024. URL https://arxiv.org/abs/2406.00606v2.

[483] Shaoxiong Ji, Shirui Pan, E. Cambria, P. Marttinen, and Philip S. Yu. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[484] Bowen Jiang, Runchuan Zhu, Jiang Wu, Zinco Jiang, Yifan He, Junyuan Gao, Jia Yu, Rui Min, Yinfan Wang, Haote Yang, et al. Evaluating large language model with knowledge oriented language specific simple question answering. 2025.

[485] Caigao Jiang, Siqiao Xue, James Zhang, Lingyue Liu, Zhibo Zhu, and Hongyan Hao. Learning large-scale universal user representation with sparse mixture of experts, arXiv preprint arXiv:2207.04648, 2022. URL https://arxiv.org/abs/2207.04648v1.

[486] Feibo Jiang, Li Dong, Yubo Peng, Kezhi Wang, Kun Yang, Cunhua Pan, D. Niyato, and O. Dobre. Large language model enhanced multi-agent systems for 6g communications. *IEEE wireless communications*, 2023.

[487] J Jiang, K Zhou, WX Zhao, Y Song, and C Zhu.... Kg-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph. 2024. URL https://arxiv.org/abs/2402.11163.

[488] Jinhao Jiang, Kun Zhou, Wayne Xin Zhao, and Ji rong Wen. Unikgqa: Unified retrieval and reasoning for solving multi-hop question answering over knowledge graph. *International Conference on Learning Representations*, 2022.

[489] Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji rong Wen. Structgpt: A general framework for large language model to reason over structured data. *Conference on Empirical Methods in Natural Language Processing*, 2023.

[490] Song Jiang, Zahra Shakeri, Aaron Chan, Maziar Sanjabi, Hamed Firooz, Yinglong Xia, Bugra Akyildiz, Yizhou Sun, Jinchao Li, Qifan Wang, and Asli Celikyilmaz. Resprompt: Residual connection prompting advances multi-step reasoning in large language models. *North American Chapter of the Association for Computational Linguistics*, 2023.

[491] Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. *Conference on Empirical Methods in Natural Language Processing*, 2023.

[479] T. S. Jayram, Younes Bouhadjar, Ryan L. McAvoy, Tomasz Kornuta, Alexis Asseman, K. Rocki, and A. Ozcan. 学习记忆、遗忘和忽略的注意力控制方法，arXiv 预印本 arXiv:1809.11087，2018。URLhttps://arxiv.org/abs/1809.11087v1.

[480] Cheonsu Jeong. 关于增强基于 LLM 的自主代理互操作性的 mcp x a2a 框架的研究，arXiv 预印本 arXiv:2506.01804，2025。URLhttps://arxiv.org/abs/2506.01804v2.

[481] Gang Ji 和 J. Bilmes. 多说话人语言建模。计算语言学协会北美分会，2004。

[482] Ke Ji, Junying Chen, Anningzhe Gao, Wenya Xie, Xiang Wan, 和 Benyou Wang. LLM 可以在没有外部监督的情况下自主学习，arXiv 预印本 arXiv:2406.00606，2024。URLhttps://arxiv.org/abs/2406.00606v2.

[483] 肖雄姬, 潘仕睿, E. Cambria, P. Marttinen, 和 Philip S. Yu. 知识图谱综述：表示、获取和应用. *IEEE TransactionsonNeural Networks and LearningSystems*, 2020.

[484] 姜 Bowen, 朱 Runchuan, 吴 Jiang, 蒋 Zinco, 何 Yifan, 高 Junyuan, 余 Jia, 闵 Rui, 王 Yinfan, 杨 Haote, 等. 基于知识导向的语言特定简单问答评估大型语言模型. 2025.

[485] 蒋 Caigao, 薛 Siqiao, 张 James, 刘 Lingyue, 朱 Zhibo, 和 郝 Hongyan. 基于稀疏专家混合学习大规模通用用户表示, arXiv 预印本 arXiv:2207.04648, 2022. URLhttps://arxiv.org/abs/2207.04648v1.

[486] 蒋 Feibo, 蒋 LiDong, 彭 Yubo, 王Kezhi, 杨 Kun, 潘 Cunhua, D.Niyato, 和 O.Dobre. 用于 6G 通信的大型语言模型增强多智能体系统. *IEEE* 无线通信, 2023.

[487] J Jiang, K Zhou, WX Zhao, Y Song, and C Zhu. . . . Kg-agent: 一个用于知识图谱复杂推理的高效自主代理框架. 2024. URL https://arxiv.org/abs/2402.11163.

[488] 蒋金浩，周坤，赵文新，文继荣。Unikgqa：统一检索和推理，用于解决知识图谱上的多跳问答。国际学习表征会议，2022。

[489] 蒋金浩，周坤，董子灿，叶克明，赵文新，温继荣。Structgpt：一个用于大型语言模型在结构化数据上进行推理的通用框架。自然语言处理经验方法会议，2023。

宋江，扎哈拉·沙基里，亚伦·陈，马兹亚尔·桑贾比，哈迈德·菲鲁兹， Yinglong 夏，Bugra Akıldız，易周·孙，金超·李，王启帆，以及阿斯拉·切利克伊尔马兹。Resprompt：残差连接提示推动了大型语言模型中的多步推理。美国计算语言学协会北美分会，2023年。

[491] 郑宝江, 弗兰克·徐, 高路宇, 孙志庆, 刘倩, Jane Dwivedi-Yu, 杨毅明, Jamie Callan, 以及 Graham Neubig. 主动检索增强生成. 自然语言处理经验方法会议, 2023.

[492] Zhonglin Jiang, Qian Tang, and Zequn Wang. Generative reliability-based design optimization using in-context learning capabilities of large language models, arXiv preprint arXiv:2503.22401, 2025. URL https://arxiv.org/abs/2503.22401v1.

[493] Zhuoxuan Jiang, Haoyuan Peng, Shanshan Feng, Fan Li, and Dongsheng Li. Llms can find mathematical reasoning mistakes by pedagogical chain-of-thought. *International Joint Conference on Artificial Intelligence*, 2024.

[494] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues?, arXiv preprint arXiv:2310.06770, 2024. URL https://arxiv.org/abs/2310.06770.

[495] B Jin, C Xie, J Zhang, KK Roy, Y Zhang, and Z Li.... Graph chain-of-thought: Augmenting large language models by reasoning on graphs. 2024. URL https://arxiv.org/abs/2404.07103.

[496] Bowen Jin, Yu Zhang, Qi Zhu, and Jiawei Han. Heterformer: Transformer-based deep node representation learning on heterogeneous text-rich networks. *Knowledge Discovery and Data Mining*, 2022.

[497] Feihu Jin, Jiajun Zhang, and Chengqing Zong. Parameter-efficient tuning for large language model without calculating its gradients. *Conference on Empirical Methods in Natural Language Processing*, 2023.

[498] Haolin Jin, Linghan Huang, Haipeng Cai, Jun Yan, Bo Li, and Huaming Chen. From llms to llm-based agents for software engineering: A survey of current, challenges and future, arXiv preprint arXiv:2408.02479, 2024. URL https://arxiv.org/abs/2408.02479v2.

[499] Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia yuan Chang, Huiyuan Chen, and Xia Hu. Llm maybe longlm: Self-extend llm context window without tuning. *International Conference on Machine Learning*, 2024.

[500] Jiajie Jin, Yutao Zhu, Xinyu Yang, Chenghao Zhang, and Zhicheng Dou. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. *The Web Conference*, 2024.

[501] Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinform.*, 2023.

[502] Tian Jin, W. Yazar, Zifei Xu, Sayeh Sharify, and Xin Wang. Self-selected attention span for accelerating large language model inference, arXiv preprint arXiv:2404.09336, 2024. URL https://arxiv.org/abs/2404.09336v1.

[503] Weiqiang Jin, Hongyang Du, Biao Zhao, Xingwu Tian, Bohang Shi, and Guang Yang. A comprehensive survey on multi-agent cooperative decision-making: Scenarios, approaches, challenges and perspectives. arXiv preprint, 2025.

[504] Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, Di Zhang, Wenwu Ou, Kun Gai, and Yadong Mu. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. *International Conference on Learning Representations*, 2023.

[492] 钟林江、唐倩和王泽群。利用大型语言模型的上下文学习能力进行生成可靠性设计优化，arXiv预印本arXiv:2503.22401，2025。URLhttps://arxiv.org/abs/2503.22401v1。

[493] 蒋卓轩，彭浩源，冯珊珊，李帆和李东升。Llmscan通过教学思维链发现数学推理错误。国际人工智能联合会议，2024.

[494] 卡洛斯·E·希门内斯，约翰·杨，亚历山大·韦蒂格，姚顺宇，裴克欣，奥菲尔·普雷斯，和卡齐克·纳拉辛汉。Swe-bench：语言模型能否解决现实世界的GitHub问题？，arXiv预印本arXiv:2310.06770，2024年。URLhttps://arxiv.org/abs/2310.06770。

[495] B Jin, C Xie, J Zhang, KK Roy, Y Zhang, and Z Li. . . . 图链思维：通过图推理增强大型语言模型。2024。URL https://arxiv.org/abs/2404.07103.

[496] Bowen Jin, Yu Zhang, Qi Zhu, and Jiawei Han. Heterformer：基于Transformer的异构文本丰富网络深度节点表示学习。 知识发现与数据挖掘，2022。

[497] Feihu Jin, Jiajun Zhang, and Chengqing Zong。无梯度计算的大型语言模型参数高效微调。自然语言处理经验方法会议，2023。

[498] Haolin Jin, Linghan Huang, Haipeng Cai, Jun Yan, Bo Li, and Huaming Chen。从llms到基于llms的软件工程代理：当前、挑战和未来的调查，arXiv预印本arXiv:2408.02479，2024。URLhttps://arxiv.org/abs/2408.02479v2.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia yuan Chang, Huiyuan Chen, and Xia Hu. LLM maybe longLM: Self-extend LLM context window without tuning. 机器学习国际会议，2024.

[500] Jiajie Jin, Yutao Zhu, Xinyu Yang, Chenghao Zhang, and Zhicheng Dou. Flashrag: 一种用于高效检索增强生成研究的模块化工具包. *The WebConference*，2024.

Qiao Jin, Yang Yifan, Chen Qingyu, and Lu Zhiyong. Genegpt: 使用领域工具增强大型语言模型以改进生物医学信息获取. *Bioinform.*，2023.

[502] 天津，W. Yazar，Zifei Xu，Sayeh Sharify，和 Xin Wang. 自适应注意力时长以加速大型语言模型推理，arXiv 预印本 arXiv:2404.09336，2024. URL https://arxiv.org/abs/2404.09336v1.

[503] 金伟强，杜红阳，赵标，田兴武，石博文，杨光. 多智能体协同决策的全面综述：场景、方法、挑战与展望. arXiv 预印本，2025.

[504] 金杨，许坤，许坤，陈立伟，廖超，谭建超，黄取哲，陈斌，雷晨毅，刘安，宋成儒，雷晓强，张迪，欧文武，赵坤，牟亚东. 在 LLM 中进行统一语言 - 视觉预训练， 采用动态离散视觉标记化. 国际学习表征会议，2023.

[505] Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, and Srijan Kumar. Mm-soc: Benchmarking multimodal large language models in social media platforms. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[506] Yiyang Jin, Kunzhao Xu, Hang Li, Xueting Han, Yanmin Zhou, Cheng Li, and Jing Bai. Reveal: Self-evolving code agents via iterative generation-verification, arXiv preprint arXiv:2506.11442, 2025. URL https://arxiv.org/abs/2506.11442v1.

[507] Jeff Johnson, Matthijs Douze, and H. Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 2017.

[508] Jeff A. Johnson and Daniel H. Bullock. Fragility in ais using artificial neural networks. *Communications of the ACM*, 2023.

[509] Zhao Kaiya, Michelangelo Naim, J. Kondic, Manuel Cortes, Jiaxin Ge, Shuying Luo, Guangyu Robert Yang, and Andrew Ahn. Lyfe agents: Generative agents for low-cost real-time social interactions, arXiv preprint arXiv:2310.02172, 2023. URL https://arxiv.org/abs/2310.02172v1.

[510] Kurmanbek Kaiyrbekov, Nic Dobbins, and Sean D. Mooney. Automated survey collection with llm-based conversational agents, arXiv preprint arXiv:2504.02891, 2025. URL https://arxiv.org/abs/2504.02891v1.

[511] A. Kakas, P. Mancarella, F. Sadri, Kostas Stathis, and Francesca Toni. Computational logic foundations of kgp agents. *Journal of Artificial Intelligence Research*, 2008.

[512] Vikas Kamra, Lakshya Gupta, Dhruv Arora, and Ashwin Kumar Yadav. Enhancing document retrieval using ai and graph-based rag techniques. *2024 5th International Conference on Communication, Computing & Industry 6.0 (C2I6)*, 2024.

[513] Eser Kandogan, Nikita Bhutani, Dan Zhang, Rafael Li Chen, Sairam Gurajada, and Estevam R. Hruschka. Orchestrating agents and data for enterprise: A blueprint architecture for compound ai, arXiv preprint arXiv:2504.08148, 2025. URL https://arxiv.org/abs/2504.08148v1.

[514] Haoyu Kang, Yuzhou Zhu, Yukun Zhong, Ke Wang Central South University, Dalian University of Technology, Nanjing University, and Xidian University. Sakr: Enhancing retrieval-augmented generation via streaming algorithm and k-means clustering. arXiv preprint, 2024.

[515] Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory os of ai agent, arXiv preprint arXiv:2506.06326, 2025. URL https://arxiv.org/abs/2506.06326v1.

[516] Jikun Kang, Wenqi Wu, Filippos Christianos, Alex J. Chan, Fraser Greenlee, George Thomas, Marvin Purtorab, and Andy Toulis. Lm2: Large memory models, arXiv preprint arXiv:2502.06049, 2025. URL https://arxiv.org/abs/2502.06049v1.

[517] Sungmin Kang, Gabin An, and S. Yoo. A quantitative and qualitative evaluation of llm-based explainable fault localization. *Proc. ACM Softw. Eng.*, 2023.

[518] Guy Kaplan, Matanel Oren, Yuval Reif, and Roy Schwartz. From tokens to words: On the inner lexicon of llms. *International Conference on Learning Representations*, 2024.

[519] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering. *Conference on Empirical Methods in Natural Language Processing*, 2020.

[505] Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, 和 Srijan Kumar. Mm-soc: 在社交媒体平台上对多模态大型语言模型进行基准测试。 计算语言学协会年会, 2024。

[506] Yiyang Jin, Kunzhao Xu, Hang Li, Xueting Han, Yanmin Zhou, Cheng Li, 和 Jing Bai. Reveal: 通过迭代生成-验证进行自我进化的代码代理，arXiv 预印本 arXiv:2506.11442，2025。
URLhttps://arxiv.org/abs/2506.11442v1.

[507] Jeff Johnson, Matthijs Douze, 和 H. Jégou. 使用 GPU 的十亿规模相似性搜索。 *IEEE* 大数据汇刊, 2017。

[508] Jeff A. Johnson 和 Daniel H. Bullock. 使用人工神经网络的 AI 脆弱性。 *ACM* 通讯, 2023。

[509] 赵凯亚, 米开朗基罗·奈姆, J. Kondic, 曼努埃尔·科尔特斯, 贾鑫, 罗舒莹, 杨光宇罗伯特, 和 安德鲁·安. Lyfe agents: 用于低成本实时社交交互的生成式代理, arXiv preprint arXiv:2310.02172, 2023.
URLhttps://arxiv.org/abs/2310.02172v1.

[510] 库尔曼别克·凯伊尔别科夫, 尼克·多宾斯, 和 肖恩·D·穆尼. 基于llm的对话代理的自动化调查收集, arXiv preprint arXiv:2504.02891, 2025. URL https://arxiv.org/abs/2504.02891v1.

[511] A. Kakas, P.Mancarella, F. Sadri, 科斯塔斯·斯塔西斯, 和 弗朗塞斯卡·托尼. kgp代理的计算逻辑基础. 人工智能研究杂志, 2008.

[512] Vikas Kamra、Lakshya Gupta、Dhruv Arora 和 Ashwin Kumar Yadav。使用 AI 和基于图的 RAG 技术增强文档检索。*2024* 年第 *5* 届通信、计算与工业 *6.0* 国际会议 *(C2I6)*，2024 年。

[513] Eser Kandogan、Nikita Bhutani、Dan Zhang、Rafael Li Chen、Sairam Gurajada 和 Estevam R. Hruschka。为企业编排代理和数据：复合 AI 的蓝图架构，arXiv 预印本 arXiv:2504.08148，2025 年。
URLhttps://arxiv.org/abs/2504.08148v1。

[514] Haoyu Kang、Yuzhou Zhu、Yukun Zhong、Ke Wang 中南大学、大连理工大学、南京大学和西安电子科技大学。Sakr：通过流式算法和 k-means 聚类增强检索增强生成。arXiv 预印本，2024 年。

[515] Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. AI agent's memory operating system, arXiv preprint arXiv:2506.06326, 2025. URLhttps://arxiv.org/abs/2506.06326v1.

[516] Jikun Kang, Wenqi Wu, Filippos Christianos, Alex J. Chan, Fraser Greenlee, George Thomas, Marvin Purtorab, and Andy Toulis. Lm2: 大型内存模型，arXiv 预印本 arXiv:2502.06049，2025年。URL https://arxiv.org/abs/2502.06049v1.

[517] Kang Sungmin, An Gabin, 和 Yoo S. 基于LLM的可解释故障定位的定量和定性评估。 *ACM Software Engineering* 会议录, 2023.

[518] Guy Kaplan、Matanel Oren、Yuval Reif 和 Roy Schwartz。从 token 到单词：关于大型语言模型的内部词汇。国际学习表征会议，2024年。

[519] 弗拉基米尔·卡普钦，巴拉斯·奥古兹，徐世文，帕特里克·刘易斯，武立德尔·余，谢尔盖·埃杜诺夫，陈丹琪，以及俞文涛。开放域问答的密集段落检索。自然语言处理经验方法会议, 2020。

[520] Zdeněk Kasner and Ondrej Dusek. Beyond traditional benchmarks: Analyzing behaviors of open llms on data-to-text generation. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[521] Kiran Kate, Tejaswini Pedapati, Kinjal Basu, Yara Rizk, Vijil Chenthamarakshan, Subhajit Chaudhury, Mayank Agarwal, and Ibrahim Abdelaziz. Longfunceval: Measuring the effectiveness of long context models for function calling, arXiv preprint arXiv:2505.10570, 2025. URL https://arxiv.org/abs/2505.10570v1.

[522] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and Franccois Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. *International Conference on Machine Learning*, 2020.

[523] Richard Katrix, Quentin Carroway, Rowan Hawkesbury, and Matthias Heathfield. Context-aware semantic recomposition mechanism for large language models, arXiv preprint arXiv:2501.17386, 2025. URL https://arxiv.org/abs/2501.17386v2.

[524] Amirhossein Kazemnejad, Inkit Padhi, K. Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Neural Information Processing Systems*, 2023.

[525] T. Kelley, R. Thomson, and Jonathan Milton. Standard model of mind: Episodic memory. *Biologically Inspired Cognitive Architectures*, 2018.

[526] Daan Kepel and Konstantina Valogianni. Autonomous prompt engineering in large language models, arXiv preprint arXiv:2407.11000, 2024. URL https://arxiv.org/abs/2407.11000v1.

[527] R. Kesner. Neurobiological foundations of an attribute model of memory. arXiv preprint, 2013.

[528] A. Khan, Md Toufique Hasan, Kai-Kristian Kemell, Jussi Rasku, and Pekka Abrahamsson. Developing retrieval augmented generation (rag) based llm systems from pdfs: An experience report, arXiv preprint arXiv:2410.15944, 2024. URL https://arxiv.org/abs/2410.15944v1.

[529] Muhammad Tayyab Khan, Lequn Chen, Ye Han Ng, Wenhe Feng, Nicholas Yew Jin Tan, and Seung Ki Moon. Leveraging vision-language models for manufacturing feature recognition in cad designs, arXiv preprint arXiv:2411.02810, 2024. URL https://arxiv.org/abs/2411.02810v1.

[530] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and M. Lewis. Generalization through memorization: Nearest neighbor language models. *International Conference on Learning Representations*, 2019.

[531] Elahe Khatibi, Ziyu Wang, and Amir M. Rahmani. Cdf-rag: Causal dynamic feedback for adaptive retrieval-augmented generation. arXiv preprint, 2025.

[532] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Neural Information Processing Systems*, 2020.

[533] Tushar Khot, H. Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. *International Conference on Learning Representations*, 2022.

[520] Zdeněk Kasner 和 Ondrej Dusek. 超越传统基准：分析开放大型语言模型在数据到文本生成中的行为。 计算语言学协会年会, 2024。

[521] Kiran Kate, Tejaswini Pedapati, Kinjal Basu, Yara Rizk, Vijil Chenthamarakshan, Subhajit Chaudhury, Mayank Agarwal, 和 Ibrahim Abdelaziz. Longfunceval：测量长上下文模型在函数调用中的有效性，arXiv 预印本 arXiv:2505.10570, 2025。URLhttps://arxiv.org/abs/2505.10570v1.

[522] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, 和 Franccois Fleuret. Transformer 是 RNN：具有线性注意力的快速自回归 Transformer。 机器学习国际会议, 2020。

[523] Richard Katrix, Quentin Carroway, Rowan Hawkesbury, and Matthias Heathfield. 大型语言模型的上下文感知语义重组机制，arXiv 预印本 arXiv:2501.17386，2025。URLhttps://arxiv.org/abs/2501.17386v2.

[524] Amirhossein Kazemnejad, Inkit Padhi, K. Ramamurthy, Payel Das, and Siva Reddy. 位置编码对 Transformer 长度泛化的影响。 神经信息处理系统, 2023。

[525] T. Kelley, R. Thomson, and Jonathan Milton. Standard model of mind: Episodic memory. *Biologically Inspired Cognitive Architectures*, 2018.

[526] Daan Kepel and Konstantina Valogianni. 大型语言模型中的自主提示工程，arXiv 预印本 arXiv:2407.11000, 2024。URLhttps://arxiv.org/abs/2407.11000v1.

[527] R. Kesner. 记忆属性模型的神经生物学基础。arXiv 预印本，2013 .

[528] A. Khan, Md Toufique Hasan, Kai-Kristian Kemell, Jussi Rasku, and Pekka Abrahamsson. 开发基于 PDF 的检索增强生成 (rag) 的 LLM 系统：经验报告，arXiv preprint arXiv:2410.15944，2024。URLhttps://arxiv.org/abs/2410.15944v1.

[529] Muhammad Tayyab Khan, Lequn Chen, Ye Han Ng, Wenhe Feng, Nicholas Yew Jin Tan, and Seung Ki Moon. 利用视觉语言模型进行 CAD 设计中的制造特征识别，arXiv preprint arXiv:2411.02810，2024。URLhttps://arxiv.org/abs/2411.02810v1.

[530] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, 和 M. Lewis。通过记忆进行泛化：最近邻语言模型。国际学习表示会议，2019。

[531] Elahe Khatibi、Ziyu Wang 和 Amir M. Rahmani。Cdf-rag：用于自适应检索增强生成的因果动态反馈。arXiv 预印本，2025。

[532] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, Dilip Krishnan. 有监督对比学习. 神经信息处理系统, 2020.

[533] Tushar Khot、H. Trivedi、Matthew Finlayson、Yao Fu、Kyle Richardson、Peter Clark 和 Ashish Sabharwal。分解式提示：一种解决复杂任务的模块化方法。国际学习表征会议, 2022。

[534] Sambhav Khurana, Xiner Li, Shurui Gui, and Shuiwang Ji. A hierarchical language model for interpretable graph reasoning, arXiv preprint arXiv:2410.22372, 2024. URL https://arxiv.org/abs/2410.22372v1.

[535] Daehee Kim, Deokhyung Kang, Sangwon Ryu, and Gary Geunbae Lee. Ontology-free general-domain knowledge graph-to-text generation dataset synthesis using large language model, arXiv preprint arXiv:2409.07088, 2024. URL https://arxiv.org/abs/2409.07088v1.

[536] Geunwoo Kim, P. Baldi, and S. McAleer. Language models can solve computer tasks. *Neural Information Processing Systems*, 2023.

[537] Jaeyeon Kim, Injune Hwang, and Kyogu Lee. Learning semantic information from raw audio signal using both contextual and phonetic representations. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2024.

[538] Jang-Hyun Kim, Junyoung Yeom, Sangdoo Yun, and Hyun Oh Song. Compressed context memory for online language model interaction. *International Conference on Learning Representations*, 2023.

[539] Jiho Kim, Yeonsu Kwon, Yohan Jo, and Edward Choi. Kg-gpt: A general framework for reasoning on knowledge graphs using large language models. *Conference on Empirical Methods in Natural Language Processing*, 2023.

[540] Jiin Kim, Byeongjun Shin, Jinha Chung, and Minsoo Rhu. The cost of dynamic reasoning: Demystifying ai agents and test-time scaling from an ai infrastructure perspective, arXiv preprint arXiv:2506.04301, 2025. URL https://arxiv.org/abs/2506.04301v1.

[541] Sehoon Kim, Suhong Moon, Ryan Tabrizi, Nicholas Lee, Michael W. Mahoney, Kurt Keutzer, and A. Gholami. An llm compiler for parallel function calling. *International Conference on Machine Learning*, 2023.

[542] Taewoon Kim, Michael Cochez, Vincent Francois-Lavet, Mark Neerincx, and Piek Vossen. A machine with short-term, episodic, and semantic memory systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):48–56, 2023. ISSN 2159-5399. doi: 10.1609/aaai.v37i1.25075. URL http://dx.doi.org/10.1609/aaai.v37i1.25075.

[543] Lukas Kirchdorfer, Robert Blümel, T. Kampik, Han van der Aa, and Heiner Stuckenschmidt. Discovering multi-agent systems for resource-centric business process simulation. *Process Science*, 2025.

[544] J. Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, J. Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2016.

[545] Louis Kirsch, James Harrison, Jascha Narain Sohl-Dickstein, and Luke Metz. General-purpose in-context learning by meta-learning transformers, arXiv preprint arXiv:2212.04458, 2022. URL https://arxiv.org/abs/2212.04458v2.

[546] Yuval Kirstain, Patrick Lewis, Sebastian Riedel, and Omer Levy. A few more examples may be worth billions of parameters. *Conference on Empirical Methods in Natural Language Processing*, 2021.

[534] Sambhav Khurana, Xiner Li, Shurui Gui, and Shuiwang Ji. 一种用于可解释图推理的层次化语言模型，arXiv 预印本 arXiv:2410.22372，2024年。URL https://arxiv.org/abs/2410.22372v1.

[535] 金大熙, 康德兴, 柳尚元, 和 李根培. 基于大型语言模型的领域无关知识图谱到文本生成数据集合成, arXiv 预印本 arXiv:2409.07088, 2024. URLhttps://arxiv.org/abs/2409.07088v1.

金根宇、P. Baldi 和 S. McAleer。语言模型可以解决计算机任务。神经信息处理系统, 2023.

[537] 金佳延、黄仁均和李奎国。利用上下文和语音表示从原始音频信号中学习语义信息。*IEEE*国际声学、语音与信号处理会议, 2024.

[538] Jang-Hyun Kim, Junyoung Yeom, Sangdoo Yun, and Hyun Oh Song. Compressed context memory for online language model interaction. *International Conference on Learning Representations*, 2023.

[539] 金基浩, 韩延苏, 赵耀汉, 和 崔东洙. Kg-gpt: 使用大型语言模型进行知识图谱推理的通用框架. 自然语言处理经验方法会议, 2023.

[540] 金锦, 沈丙俊, 韩镇河, 和 李珉洙. 动态推理的成本：从人工智能基础设施视角揭示人工智能代理和测试时扩展, arXiv 预印本 arXiv:2506.04301, 2025. URLhttps://arxiv.org/abs/2506.04301v1.

[541] 金世勋, 文秀洪, 泰伦·塔布里齐, 李尼古拉斯, 马洪·迈克尔·W., 库尔特·基茨, 和 A. 诺拉米. 用于并行函数调用的llm编译器. 机器学习国际会议, 2023.

[542] 金泰雄, 科切兹·迈克尔, 弗朗索瓦-拉韦·文森特, 尼尔林克斯·马克, 和 维克·皮克. 具有短期、情景和语义记忆系统的机器. 人工智能协会会议论文集, 37(1):48–56, 2023. ISSN2159-5399. doi: 10.1609/aaai.v37i1.25075. URLhttp://dx.doi.org/10.1609/aaai.v37i1.25075.

[543] Lukas Kirchdorfer, Robert Blümel, T. Kampik, Han van der Aa, and Heiner Stuckenschmidt. 发现面向资源中心的业务流程模拟的多智能体系统。 *Process Science*, 2025。

[544] J. Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, J. Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. 克服神经网络中的灾难性遗忘。 美国国家科学院院报, 2016。

[545] Louis Kirsch, James Harrison, Jascha Narain Sohl-Dickstein, and Luke Metz. 通过元学习Transformer实现通用情境学习，arXiv预印本 arXiv:2212.04458，2022。URLhttps://arxiv.org/abs/2212.04458v2.

[546] Yuval Kirstain, Patrick Lewis, Sebastian Riedel, and Omer Levy. 还有一些例子可能值得数十亿个参数。 自然语言处理经验方法会议, 2021。

[547] Andrew Kiruluta, Preethi Raju, and Priscilla Burity. Breaking quadratic barriers: A non-attention llm for ultra-long context horizons, arXiv preprint arXiv:2506.01963, 2025. URL https://arxiv.org/abs/2506.01963v1.

[548] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *International Conference on Learning Representations*, 2020.

[549] Vincent Koc, Jacques Verre, Douglas Blank, and Abigail Morgan. Mind the metrics: Patterns for telemetry-aware in-ide ai application development using the model context protocol (mcp), arXiv preprint arXiv:2506.11019, 2025. URL https://arxiv.org/abs/2506.11019v1.

[550] Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 2017.

[551] Jing Yu Koh, R. Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. *International Conference on Machine Learning*, 2023.

[552] Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks, arXiv preprint arXiv:2401.13649, 2024. URL https://arxiv.org/abs/2401.13649v2.

[553] Takeshi Kojima, S. Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Neural Information Processing Systems*, 2022.

[554] Yilun Kong, Jingqing Ruan, Yihong Chen, Bin Zhang, Tianpeng Bao, Shiwei Shi, Guoqing Du, Xiaoru Hu, Hangyu Mao, Ziyue Li, Xingyu Zeng, and Rui Zhao. Tptu-v2: Boosting task planning and tool usage of large language model-based agents in real-world systems, arXiv preprint arXiv:2311.11315, 2023. URL https://arxiv.org/abs/2311.11315.

[555] P. Korzyński, G. Mazurek, Pamela Krzypkowska, and Artur Kurasiński. Artificial intelligence prompt engineering as a new digital competence: Analysis of generative ai technologies such as chatgpt. *Entrepreneurial Business and Economics Review*, 2023.

[556] Oliver Kramer. Cognitive prompts using guilford's structure of intellect model. arXiv preprint, 2025.

[557] Oliver Kramer. Conceptual metaphor theory as a prompting paradigm for large language models, arXiv preprint arXiv:2502.01901, 2025. URL https://arxiv.org/abs/2502.01901v1.

[558] Oliver Kramer and Jill Baumann. Unlocking structured thinking in language models with cognitive prompting. *ESANN 2025 proceesdings*, 2024.

[559] K. Kravari and Nick Bassiliades. A survey of agent platforms. *Journal of Artificial Societies and Social Simulation*, 2015.

[560] Prashant Krishnan, Zilong Wang, Yangkun Wang, and Jingbo Shang. Towards few-shot entity recognition in document images: A graph neural network approach robust to image manipulation. *International Conference on Language Resources and Evaluation*, 2023.

[561] W. Kruijne, S. Bohté, P. Roelfsema, and C. Olivers. Flexible working memory through selective gating and attentional tagging. *bioRxiv*, 2019.

[547] Andrew Kiruluta, Preethi Raju, and Priscilla Burity. 突破二次障碍：一种非注意力大语言模型用于超长上下文范围，arXiv 预印本 arXiv:2506.01963，2025。URL https://arxiv.org/abs/2506.01963v1.

[548] Nikita Kitaev, LukaszKaiser, andAnselm Levskaya. Reformer：高效的 Transformer。学习表示国际会议，2020。

[549] Vincent Koc, Jacques Verre, Douglas Blank, and Abigail Morgan. 注意指标：使用模型上下文协议（mcp）进行感知 IDE 人工智能应用程序开发的模式，arXiv 预印本 arXiv:2506.11019，2025。URLhttps://arxiv.org/abs/2506.11019v1.

[550] Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. Narrativeqa 阅读理解挑战。计算语言学协会事务，2017。

[551] Jing Yu Koh, R. Salakhutdinov, and Daniel Fried. 将语言模型与图像结合用于多模态输入和输出. 机器学习国际会议,2023.

[552] Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Gra- ham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. Visualwebarena: 在真实视觉网络任务上评估多模态智能体，arXiv 预印本 arXiv:2401.13649，2024. URLhttps://arxiv.org/abs/2401.13649v2.

[553] Takeshi Kojima, S. Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 大型语言模型是零样本推理者. 神经信息处理系统，2022.

[554] Yilun Kong, Jingqing Ruan, Yihong Chen, Bin Zhang, Tianpeng Bao, Shiwei Shi, Guoqing Du, Xiaoru Hu, Hangyu Mao, Ziyue Li, Xingyu Zeng, and Rui Zhao. Tptu-v2: 提升基于大型语言模型智能体在现实系统中的任务规划和工具使用能力，arXiv 预印本 arXiv:2311.11315，2023. URLhttps://arxiv.org/abs/2311.11315.

[555] P. Korzyński, G. Mazurek, Pamela Krzypkowska, 和 Artur Kurasiński. 人工智能提示工程作为一种新的数字能力：分析 ChatGPT 等生成式人工智能技术。*Entrepreneurial Businessand Economics Review*，2023。

[556] Oliver Kramer. 基于吉尔福特智力结构模型的认知提示。arXiv 预印本，2025。

[557] Oliver Kramer. 概念隐喻理论作为大型语言模型的提示范式，arXiv 预印本 arXiv:2502.01901，2025。URLhttps://arxiv.org/abs/2502.01901v1。

[558] Oliver Kramer 和 Jill Baumann. 使用认知提示解锁语言模型中的结构化思维。*ESANN 2025 proceesdings*，2024。

[559] K. Kravari 和 Nick Bassiliades. 智能体平台的综述。*Journal of Artificial Societies andSocialSimulation*，2015。

[560] Prashant Krishnan, Zilong Wang, Wang Yangkun, 和 Shang Jingbo. 面向文档图像中的少样本实体识别：一种对图像操作具有鲁棒性的图神经网络方法。国际语言资源与评估会议，2023。

[561] W. Kruijne, S.Bohté, P. Roelfsema, 和 C.Olivers. 通过选择性门控和注意力标记实现灵活的工作记忆。*bioRxiv*，2019。

[562] L. Krupp, Daniel Geissler, P. Lukowicz, and Jakob Karolus. Towards sustainable web agents: A plea for transparency and dedicated metrics for energy consumption, arXiv preprint arXiv:2502.17903, 2025. URL https://arxiv.org/abs/2502.17903v1.

[563] M. Kuhail, Jose Berengueres, Fatma Taher, Sana Z. Khan, and Ansah Siddiqui. Designing a haptic boot for space with prompt engineering: Process, insights, and implications. *IEEE Access*, 2024.

[564] Amandeep Kumar, Muzammal Naseer, Sanath Narayan, R. Anwer, Salman H. Khan, and Hisham Cholakkal. Multi-modal generation via cross-modal in-context learning, arXiv preprint arXiv:2405.18304v1, 2024. URL https://arxiv.org/abs/2405.18304v1.

[565] Rajeev Kumar, Harishankar Kumar, and Kumari Shalini. Detecting and mitigating bias in llms through knowledge graph-augmented training. *2025 International Conference on Artificial Intelligence and Data Engineering (AIDE)*, 2025.

[566] Taeyoon Kwon, Dongwook Choi, Sunghwan Kim, Hyojun Kim, Seungjun Moon, Beong woo Kwak, Kuan-Hao Huang, and Jinyoung Yeo. Embodied agents meet personalization: Exploring memory utilization for personalized assistance, arXiv preprint arXiv:2505.16348, 2025. URL https://arxiv.org/abs/2505.16348v1.

[567] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Haotong Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. *Symposium on Operating Systems Principles*, 2023.

[568] T. Lai, Quan Hung Tran, Trung Bui, and D. Kihara. A gated self-attention memory network for answer selection. *Conference on Empirical Methods in Natural Language Processing*, 2019.

[569] Divya Lamba. The role of prompt engineering in improving language understanding and generation. *International Journal For Multidisciplinary Research*, 2024.

[570] Xiaochong Lan, Jie Feng, Jia Lei, Xinlei Shi, and Yong Li. Benchmarking and advancing large language models for local life services, arXiv preprint arXiv:2506.02720, 2025. URL https://arxiv.org/abs/2506.02720v1.

[571] LangChain Team. Memory in langgraph. https://langchain-ai.github.io/langgraph/concepts/memory/, 2025. Accessed: 2025-07-17.

[572] Samuel T. Langlois, Oghenetekevwe Akoroda, Estefany Carrillo, J. Herrmann, S. Azarm, Huan Xu, and Michael W. Otte. Metareasoning structures, problems, and modes for multiagent systems: A survey. *IEEE Access*, 2020.

[573] B. Lattimer, Varun Gangal, Ryan McDonald, and Yi Yang. Sparse rewards can self-train dialogue agents, arXiv preprint arXiv:2409.04617, 2024. URL https://arxiv.org/abs/2409.04617v2.

[574] Pak Kin Lau and Stuart Michael McManus. Mining asymmetric intertextuality, arXiv preprint arXiv:2410.15145, 2024. URL https://arxiv.org/abs/2410.15145v1.

[575] Hung Le, T. Tran, and S. Venkatesh. Self-attentive associative memory. *International Conference on Machine Learning*, 2020.

[562] L. Krupp、Daniel Geissler、P. Lukowicz 和 Jakob Karolus。迈向可持续的网页代理：呼吁透明度和专门的能源消耗指标，arXiv 预印本 arXiv:2502.17903，2025 年。URLhttps://arxiv.org/abs/2502.17903v1。

[563] M. Kuhail, Jose Berengueres, Fatma Taher, Sana Z. Khan, and Ansah Siddiqui. 设计用于太空的触觉靴：通过提示工程进行设计：流程、见解和启示。IEEE Access, 2024.

[564]Amandeep Kumar、Muzammal Naseer、Sanath Narayan、R. Anwer、Salman H. Khan 和 Hisham Cholakkal。跨模态情境学习实现多模态生成，arXiv 预印本 arXiv:2405.18304v1，2024。URLhttps://arxiv.org/abs/2405.18304v1。

[565] Rajeev Kumar, Harishankar Kumar, 和 Kumari Shalini. 通过知识图谱增强训练检测和缓解大型语言模型中的偏差。 *2025* 国际人工智能与数据工程会议 *(AIDE)*, 2025。

[566] Taeyoon Kwon, Dongwook Choi, Sunghwan Kim, Hyojun Kim, Seungjun Moon, Beong woo Kwak, Kuan-Hao Huang, 和 Jinyoung Yeo. 体现代理遇见个性化：探索个性化辅助的记忆利用率，arXiv 预印本 arXiv:2505.16348, 2025。 URL https://arxiv.org/abs/2505.16348v1.

[567] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Haotong Zhang, 和 Ion Stoica. 通过分页注意力高效管理大型语言模型服务的内存。 操作系统原理研讨会，2023。

[568] T. Lai, Quan Hung Tran, Trung Bui, and D. Kihara. 基于门控自注意力记忆网络的答案选择. 自然语言处理经验方法会议, 2019.

[569] Divya Lamba. 提示工程在提升语言理解和生成中的作用.多学科研究国际期刊, 2024.

[570] Xiaochong Lan, Jie Feng, Jia Lei, Xinlei Shi, and Yong Li. 面向本地生活服务的超大语言模型基准测试与进展, arXiv 预印本 arXiv:2506.02720, 2025. URL https://arxiv.org/abs/2506.02720v1.

[571] LangChain 团队. langgraph 中的记忆. https://langchain-ai.github.io/langgraph/concepts/memory/, 2025. 访问时间: 2025-07-17.

[572] Samuel T. Langlois, Oghenetekevwe Akoroda, Estefany Carrillo, J. Herrmann, S. Azarm, Huan Xu, and Michael W. Otte. 多智能体系统的元推理结构、问题和模式：一项调查. *IEEEAccess*, 2020.

[573] B. Lattimer、Varun Gangal、Ryan McDonald 和 Yi Yang。稀疏奖励可以自训练对话代理，arXiv 预印本 arXiv:2409.04617，2024 年。URLhttps://arxiv.org/abs/2409.04617v2。

[574] Pak Kin Lau 和 Stuart Michael McManus. 开采非对称互文性, arXiv 预印本 arXiv:2410.15145, 2024. URLhttps://arxiv.org/abs/2410.15145v1.

[575] Hung Le, T. Tran, 和 S. Venkatesh. 自注意力关联记忆. 机器学习国际会议, 2020.

[576] Dohyun Lee, Seungil Chad Lee, Chanwoo Yang, Yujin Baek, and Jaegul Choo. Exploring in-context example generation for machine translation, arXiv preprint arXiv:2506.00507, 2025. URL https://arxiv.org/abs/2506.00507v1.

[577] Dongyub Lee, Eunhwan Park, Hodong Lee, and Heuiseok Lim. Ask, assess, and refine: Rectifying factual consistency and hallucination in llms with metric-guided feedback learning. *Conference of the European Chapter of the Association for Computational Linguistics*, 2024.

[578] Eunhae Lee. Towards ethical personal ai applications: Practical considerations for ai assistants with long-term memory, arXiv preprint arXiv:2409.11192, 2024. URL https://arxiv.org/abs/2409.11192v1.

[579] Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. Prompted llms as chatbot modules for long open-domain conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.277. URL http://dx.doi.org/10.18653/v1/2023.findings-acl.277.

[580] Heejun Lee, Geon Park, Youngwan Lee, Jina Kim, Wonyoung Jeong, Myeongjae Jeon, and Sung Ju Hwang. A training-free sub-quadratic cost transformer model serving framework with hierarchically pruned attention, arXiv preprint arXiv:2406.09827, 2024. URL https://arxiv.org/abs/2406.09827v3.

[581] Ho-Jun Lee, Junho Kim, Hyunjun Kim, and Yonghyun Ro. Refocus: Reinforcement-guided frame optimization for contextual understanding, arXiv preprint arXiv:2506.01274v1, 2025. URL https://arxiv.org/abs/2506.01274v1.

[582] Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and E. Brunskill. Supervised pretraining can learn in-context reinforcement learning. *Neural Information Processing Systems*, 2023.

[583] Namkyeong Lee, E. Brouwer, Ehsan Hajiramezanali, Chanyoung Park, and Gabriele Scalia. Rag-enhanced collaborative llm agents for drug discovery, arXiv preprint arXiv:2502.17506, 2025. URL https://arxiv.org/abs/2502.17506v2.

[584] Shinbok Lee, Gaeun Seo, Daniel Lee, Byeongil Ko, Sunghee Jung, and M. Shin. Functionchat-bench: Comprehensive evaluation of language models' generative capabilities in korean tool-use dialogs. arXiv preprint, 2024.

[585] Younghun Lee, Sungchul Kim, Ryan A. Rossi, Tong Yu, and Xiang Chen. Learning to reduce: Towards improving performance of large language models on structured data, arXiv preprint arXiv:2407.02750, 2024. URL https://arxiv.org/abs/2407.02750v1.

[586] Younghun Lee, Sungchul Kim, Tong Yu, Ryan A. Rossi, and Xiang Chen. Learning to reduce: Optimal representations of structured data in prompting large language models, arXiv preprint arXiv:2402.14195, 2024. URL https://arxiv.org/abs/2402.14195v1.

[587] Yu-Ting Lee, Hui-Ying Shih, Fu-Chieh Chang, and Pei-Yuan Wu. An explanation of intrinsic self-correction via linear representations and latent concepts, arXiv preprint arXiv:2505.11924, 2025. URL https://arxiv.org/abs/2505.11924v1.

[576] 李多贤, 李承吉, 杨灿宇, 白宇珍, 和 蔡哲圭. 探索机器翻译的上下文示例生成, arXiv preprint arXiv:2506.00507, 2025. URLhttps://arxiv.org/abs/2506.00507v1.

[577] 李东雨, 朴恩孕, 李浩东, 和 李徽硕. 问, 评估, 和 完善: 使用度量引导反馈学习纠正大语言模型的事实一致性和幻觉. 欧洲计算语言学协会欧洲分会会议, 2024.

[578] 李恩海. 迈向道德个人人工智能应用: 具有长期记忆的人工智能助手的实际考虑, arXiv preprint arXiv:2409.11192, 2024. URLhttps://arxiv.org/abs/2409.11192v1.

[579] 李吉文，哈特曼·沃尔克，朴钟浩，帕帕伊奥卢斯·迪米特里斯，和李康佑。提示式LLM作为长开放域对话的聊天机器人模块。在 计算语言学协会发现：*ACL 2023*。计算语言学协会，2023。doi: 10.18653/ v1/2023.findings-acl.277.
URLhttp://dx.doi.org/10.18653/v1/2023.findings-acl.277。

[580] 李希俊, 朴贤, 李永万, 金晶娜, 靳永勇, 全明宰, 和黄成柱. 一种具有分层剪枝注意力的无训练亚二次成本 Transformer 模型服务框架, arXiv 预印本 arXiv:2406.09827, 2024.
URLhttps://arxiv.org/abs/2406.09827v3.

[581] Ho-Jun Lee, Junho Kim, Hyunjun Kim, and Yonghyun Ro. Refocus: 强化学习引导的框架优化用于上下文理解，arXiv preprint arXiv:2506.01274v1, 2025. URL https://arxiv.org/abs/2506.01274v1.

[582] Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, ChelseaFinn, OfirNachum, andE. Brunskill. 监督预训练可以学习上下文强化学习。 *Neural Information ProcessingSystems*, 2023.

[583] Namkyeong Lee, E. Brouwer, Ehsan Hajiramezanali, Chanyoung Park, and Gabriele Scalia. Rag- 增强型协同 LLM 代理用于药物发现，arXiv preprint arXiv:2502.17506, 2025.
URLhttps://arxiv.org/abs/2502.17506v2.

[584] Shinbok Lee, Gaeun Seo, Daniel Lee, Byeongil Ko, Sunghee Jung, 和 M. Shin. Functionchat-bench: 对韩语工具使用对话中语言模型生成能力的综合评估。arXiv 预印本，2024年。

[585] Younghun Lee, Sungchul Kim, Ryan A. Rossi, Tong Yu, 和 Xiang Chen. 学习减少：提高大型语言模型在结构化数据上的性能，arXiv 预印本 arXiv:2407.02750，2024年。URLhttps://arxiv.org/abs/2407.02750v1.

[586] Younghun Lee, Sungchul Kim, Tong Yu, Ryan A. Rossi, 和 Xiang Chen. 学习减少：在提示大型语言模型中结构化数据的最佳表示，arXiv 预印本 arXiv:2402.14195，2024年。URLhttps://arxiv.org/abs/2402.14195v1.

[587] 李宇婷，施惠莹，张福杰，和吴培远。通过线性表示和潜在概念解释内在自校正，arXiv 预印本 arXiv:2505.11924，2025。URL https://arxiv.org/abs/2505.11924v1.

[588] Melissa Lehman and Kenneth J. Malmberg. A buffer model of memory encoding and temporal correlations in retrieval. *Psychology Review*, 2013.

[589] Yiming Lei, Zhizheng Yang, Zeming Liu, Haitao Leng, Shaoguo Liu, Tingting Gao, Qingjie Liu, and Yunhong Wang. Contextqformer: A new context modeling method for multi-turn multi-modal conversations, arXiv preprint arXiv:2505.23121v1, 2025. URL https://arxiv.org/abs/2505.23121v1.

[590] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *Conference on Empirical Methods in Natural Language Processing*, 2021.

[591] Patrick Lewis, Ethan Perez, Aleksandara Piktus, F. Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, M. Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Neural Information Processing Systems*, 2020.

[592] Bohan Li, Yutai Hou, and Wanxiang Che. Data augmentation approaches in natural language processing: A survey. *AI Open*, 2021.

[593] Chaozhuo Li, Bochen Pang, Yuming Liu, Hao Sun, Zheng Liu, Xing Xie, Tianqi Yang, Yanling Cui, Liangjie Zhang, and Qi Zhang. Adsgnn: Behavior-graph augmented relevance modeling in sponsored search. *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

[594] Chengpeng Li, Zhengyang Tang, Ziniu Li, Mingfeng Xue, Keqin Bao, Tian Ding, Ruoyu Sun, Benyou Wang, Xiang Wang, Junyang Lin, and Dayiheng Liu. Cort: Code-integrated reasoning within thinking, arXiv preprint arXiv:2506.09820, 2025. URL https://arxiv.org/abs/2506.09820v2.

[595] Chengshu Li, Jacky Liang, Andy Zeng, Xinyun Chen, Karol Hausman, Dorsa Sadigh, Sergey Levine, Fei-Fei Li, Fei Xia, and Brian Ichter. Chain of code: Reasoning with a language model-augmented code emulator. *International Conference on Machine Learning*, 2023.

[596] Chuanhao Li, Runhan Yang, Tiankai Li, Milad Bafarassat, Kourosh Sharifi, Dirk Bergemann, and Zhuoran Yang. Stride: A tool-assisted llm agent framework for strategic and interactive decision-making, arXiv preprint arXiv:2405.16376, 2024. URL https://arxiv.org/abs/2405.16376v2.

[597] Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory, arXiv preprint arXiv:2211.05110, 2022. URL https://arxiv.org/abs/2211.05110.

[598] Daniel Li and Lincoln Murr. Humaneval on latest gpt models - 2024. arXiv preprint, 2024.

[599] Fu Li, Xueying Wang, Bin Li, Yunlong Wu, Yanzhen Wang, and Xiaodong Yi. A study on training and developing large language models for behavior tree generation, arXiv preprint arXiv:2401.08089, 2024. URL https://arxiv.org/abs/2401.08089v1.

[600] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[588] Melissa Lehman 和 Kenneth J. Malmberg. 一种记忆编码和检索中时间相关性的缓冲模型. 心理学评论, 2013.

[589] Yiming Lei, Zhizheng Yang, Zeming Liu, Haitao Leng, Shaoguo Liu, Tingting Gao, Qingjie Liu, 和 Yunhong Wang. Contextqformer: 一种用于多轮多模态对话的新上下文建模方法, arXiv 预印本 arXiv:2505.23121v1, 2025. URLhttps://arxiv.org/abs/2505.23121v1.

[590] Brian Lester, Rami Al-Rfou, 和 Noah Constant. 参数高效提示微调的规模力量. 自然语言处理经验方法会议, 2021.

[591] Patrick Lewis, Ethan Perez, Aleksandara Piktus, F. Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, M. Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, 和 Douwe Kiela. 用于知识密集型 NLP 任务的检索增强生成 . 神经信息处理系统 , 2020.

[592] Bohan Li, Yutai Hou, and Wanxiang Che. 自然语言处理中的数据增强方法：一项调查. *AI Open*,2021.

[593] Chaozhuo Li, Bochen Pang, Yuming Liu, Hao Sun, Zheng Liu, Xing Xie, Tianqi Yang, Yanling Cui, Liangjie Zhang, andQiZhang. Adsgnn: 行为图增强的相关性建模在竞价搜索中. *Annual International ACM SIGIR Conference on Research andDevelopment inInformationRetrieval*, 2021.

[594] Chengpeng Li, Zhengyang Tang, Ziniu Li, Mingfeng Xue, Keqin Bao, Tian Ding, Ruoyu Sun, Benyou Wang, Xiang Wang, Junyang Lin, and Dayiheng Liu. Cort: 思考中的代码集成推理, arXiv 预印本 arXiv:2506.09820, 2025. URLhttps://arxiv.org/abs/2506.09820v2.

[595] 程舒, 李杰克, 曾安迪, 陈新云, 卡罗尔·豪斯曼, 多尔萨·萨迪格, 谢尔盖·列文, 李飞飞, 谢菲, 布莱恩·伊彻. 代码链：使用语言模型增强的代码模拟器进行推理. 机器学习国际会议, 2023.

[596] 李传浩, 杨润寒, 李天凯, 米拉德·巴法拉斯阿特, 库鲁什·沙里菲, 迪尔克·伯格曼, 和杨卓然. Stride：一个用于战略和交互式决策的工具辅助llm代理框架, arXiv预印本 arXiv:2405.16376, 2024. URLhttps://arxiv.org/abs/2405.16376v2.

[597] 李达亮, 安克特·辛格·拉瓦特, 曼齐尔·扎黑尔, 王欣, 米哈伊尔·卢卡西克, 安德烈亚斯·维特, 余斐, 卡姆卡尔·桑吉夫. 具有可控工作内存的大语言模型, arXiv预印本 arXiv:2211.05110, 2022. URLhttps://arxiv.org/abs/2211.05110.

[598] Daniel Li和Lincoln Murr. Humaneval在最新GPT模型上的表现 - 2024. arXiv预印本, 2024.

[599] 付丽, 王雪莹, 李斌, 吴云龙, 王彦臻, 以及易晓东. 关于行为树生成的大型语言模型训练与发展研究, arXiv 预印本 arXiv:2401.08089, 2024. URLhttps://arxiv.org/abs/2401.08089v1.

李国豪、哈桑·阿卜杜勒·卡德尔·哈穆德、哈尼·伊塔尼、德米特里·基兹布林和伯纳德·加内姆。Camel：用于“心灵”探索大型语言模型社会的交流代理。在《第37届神经信息处理系统大会》，2023。

[601] Guozheng Li, Peng Wang, Jiajun Liu, Yikai Guo, Ke Ji, Ziyu Shang, and Zijie Xu. Meta in-context learning makes large language models better zero and few-shot relation extractors. *International Joint Conference on Artificial Intelligence*, 2024.

[602] Haoyang Li, Jing Zhang, Cuiping Li, and Hong Chen. Resdsql: Decoupling schema linking and skeleton parsing for text-to-sql. *AAAI Conference on Artificial Intelligence*, 2023.

[603] Jia Li, Ge Li, Yongming Li, and Zhi Jin. Structured chain-of-thought prompting for code generation. *ACM Transactions on Software Engineering and Methodology*, 2023.

[604] Jia Li, Xiangguo Sun, Yuhan Li, Zhixun Li, Hong Cheng, and Jeffrey Xu Yu. Graph intelligence with large language models and prompt learning. *Knowledge Discovery and Data Mining*, 2024.

[605] Jiahao Nick Li, Yan Xu, Tovi Grossman, Stephanie Santosa, and Michelle Li. Omniactions: Predicting digital actions in response to real-world multimodal sensory inputs with llms. *International Conference on Human Factors in Computing Systems*, 2024.

[606] Jiarui Li, Ye Yuan, and Zehua Zhang. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases, arXiv preprint arXiv:2403.10446, 2024. URL https://arxiv.org/abs/2403.10446v1.

[607] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat seng Chua, Siliang Tang, and Yueting Zhuang. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. *International Conference on Learning Representations*, 2023.

[608] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and S. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Neural Information Processing Systems*, 2021.

[609] Junnan Li, Dongxu Li, S. Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning*, 2023.

[610] Kun Li, Tianhua Zhang, Yunxiang Li, Hongyin Luo, Abdalla Moustafa, Xixin Wu, James Glass, and Helen M. Meng. Generate, discriminate, evolve: Enhancing context faithfulness via fine-grained sentence-level self-evolution, arXiv preprint arXiv:2503.01695, 2025. URL https://arxiv.org/abs/2503.01695v1.

[611] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wen Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding, arXiv preprint arXiv:2305.06355v2, 2023. URL https://arxiv.org/abs/2305.06355v2.

[612] M Li, Y Zhao, B Yu, F Song, H Li, H Yu, and Z Li.... Api-bank: A comprehensive benchmark for tool-augmented llms. 2023. URL https://arxiv.org/abs/2304.08244.

[613] Michelle M. Li, Ben Y. Reis, Adam Rodman, Tianxi Cai, Noa Dagan, Ran D. Balicer, Joseph Loscalzo, Isaac S. Kohane, and M. Zitnik. One patient, many contexts: Scaling medical ai through contextual intelligence, arXiv preprint arXiv:2506.10157, 2025. URL https://arxiv.org/abs/2506.10157v1.

[601] 李国正, 王鹏, 刘嘉俊, 郭奕凯, 贾纪, 尚子宇, 和徐子杰. 元上下文学习使大型语言模型成为更好的零和少样本关系抽取器. 国际人工智能联合会议, 2024.

[602] 李浩阳, 张静, 李翠萍, 和陈红. Resdsql: 解耦模式链接和骨架解析用于文本到SQL. *AAAI*人工智能会议, 2023.

[603] 李嘉, 李格, 李永明, 和金智. 结构化思维链提示用于代码生成.*ACM*软件工程与方法学汇刊, 2023.

[604] Jia Li, Xiangguo Sun, Yuhan Li, Zhixun Li, Hong Cheng, and Jeffrey Xu Yu. Graph intelligence with large language models and prompt learning. *Knowledge Discovery and Data Mining*, 2024.

[605] 李嘉豪, 许岩, Tovi Grossman, Stephanie Santosa, 和李美琪. Omniactions: 使用LLM预测对真实世界多模态感官输入的数字动作. 人机交互国际会议, 2024.

[606] 李家瑞, 袁叶, 和 张泽华. 使用rag增强llm事实准确性以对抗幻觉：私有知识库中特定领域查询的案例研究, arXiv预印本 arXiv:2403.10446, 2024. URLhttps://arxiv.org/abs/2403.10446v1.

[607] 李峻程, 潘凯航, 葛志奇, 高明和, 张汉旺, 季伟, 张文桥, Chua Tat seng, 唐思亮, 和 庄玉婷. 微调多模态llm以遵循零样本示范性指令. 学习表示国际会议, 2023.

[608] 李俊南, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, 熊才明, 和 S. Hoi. 先对齐后融合：视觉和语言表示学习与动量蒸馏.神经信息处理系统, 2021.

[609] 李俊南, 李东旭, S. Savarese, 和 Steven C. H. Hoi. Blip-2: 使用冻结图像编码器和大型语言模型进行语言-图像预训练的引导. 国际机器学习会议, 2023.

[610] 李坤, 张天华, 李云翔, 罗红寅, Abdalla Moustafa, 吴希欣, James Glass, 和 Helen M. Meng. 生成、区分、进化：通过细粒度句子级自进化增强上下文忠实度, arXiv 预印本 arXiv:2503.01695, 2025. URLhttps://arxiv.org/abs/2503.01695v1.

[611] 李坤昌, 何一南, 王毅, 李一卓, 王文, 罗平, 王亚丽, 王黎明, 和 邱宇. Videochat: 以聊天为中心的视频理解, arXiv 预印本 arXiv:2305.06355v2, 2023. URL https://arxiv.org/abs/2305.06355v2.

[612] M Li, Y Zhao, B Yu, F Song, H Li, H Yu, and Z Li.... Api-bank: A comprehensive benchmark for tool-augmented llms. 2023. URL https://arxiv.org/abs/2304.08244.

[613] Michelle M. Li, Ben Y. Reis, Adam Rodman, Tianxi Cai, Noa Dagan, Ran D. Balicer, Joseph Loscalzo, Isaac S. Kohane, and M. Zitnik. One patient, many contexts: Scaling medical ai through contextual intelligence, arXiv preprint arXiv:2506.10157, 2025. URL https://arxiv.org/abs/2506.10157v1.

[614] Ming Li, Keyu Chen, Ziqian Bi, Ming Liu, Benji Peng, Qian Niu, Junyu Liu, Jinlang Wang, Sen Zhang, Xuanhe Pan, Jiawei Xu, and Pohsun Feng. Surveying the mllm landscape: A meta-review of current surveys, arXiv preprint arXiv:2409.18991, 2024. URL https://arxiv.org/abs/2409.18991v1.

[615] Minghao Li, Feifan Song, Yu Bowen, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. Api-bank: A comprehensive benchmark for tool-augmented llms. *Conference on Empirical Methods in Natural Language Processing*, 2023.

[616] Qiaomu Li and Ying Xie. From glue-code to protocols: A critical analysis of a2a and mcp integration for scalable agent systems, arXiv preprint arXiv:2505.03864, 2025. URL https://arxiv.org/abs/2505.03864v1.

[617] Rongsheng Li, Jin Xu, Zhixiong Cao, Hai-Tao Zheng, and Hong-Gee Kim. Extending context window in large language models with segmented base adjustment for rotary position embeddings. *Applied Sciences*, 2024.

[618] Shuaike Li, Kai Zhang, Qi Liu, and Enhong Chen. Mindbridge: Scalable and cross-model knowledge editing via memory-augmented modality, arXiv preprint arXiv:2503.02701v1, 2025. URL https://arxiv.org/abs/2503.02701v1.

[619] Shuaiyi Li, Zhisong Zhang, Yang Deng, Chenlong Deng, Tianqing Fang, Hongming Zhang, Haitao Mi, Dong Yu, and Wai Lam. Incomes: Integrating compression and selection mechanisms into llms for efficient model editing, arXiv preprint arXiv:2505.22156, 2025. URL https://arxiv.org/abs/2505.22156v1.

[620] Siheng Li, Cheng Yang, Zesen Cheng, Lemao Liu, Mo Yu, Yujiu Yang, and Wai Lam. Large language models can self-improve in long-context reasoning, arXiv preprint arXiv:2411.08147, 2024. URL https://arxiv.org/abs/2411.08147v1.

[621] X Li, H Zou, and P Liu. Torl: Scaling tool-integrated rl. 2025. URL https://arxiv.org/abs/2503.23383.

[622] Xiaopeng Li, Pengyue Jia, Derong Xu, Yi Wen, Yingyi Zhang, Wenlin Zhang, Wanyu Wang, Yichao Wang, Zhaochen Du, Xiangyang Li, Yong Liu, Huifeng Guo, Ruiming Tang, and Xiangyu Zhao. A survey of personalization: From rag to agent, arXiv preprint arXiv:2504.10147, 2025. URL https://arxiv.org/abs/2504.10147v1.

[623] Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability, arXiv preprint arXiv:2504.21776, 2025. URL https://arxiv.org/abs/2504.21776v1.

[624] Xin Li, Qizhi Chu, Yubin Chen, Yang Liu, Yaoqi Liu, Zekai Yu, Weize Chen, Cheng Qian, Chuan Shi, and Cheng Yang. Graphteam: Facilitating large language model-based graph analysis via multi-agent collaboration, arXiv preprint arXiv:2410.18032v4, 2024. URL https://arxiv.org/abs/2410.18032v4.

[625] Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 2024.

[614] 明李，陈键宇，毕子谦，刘明，彭本基，牛倩，刘军宇，王金浪，张森，潘轩鹤，徐佳伟，冯博森. 综述大语言模型领域：当前调查的元综述，arXiv 预印本 arXiv:2409.18991，2024 年。URLhttps://arxiv.org/abs/2409.18991v1.

李明浩, 宋飞帆, Bowen Yu, 余海阳, 李周军, 黄飞, 李永斌. Api-bank: 一个用于工具增强型大语言模型的综合基准. 自然语言处理经验方法会议, 2023.

[616] 李乔木和谢颖. 从粘合代码到协议：对可扩展代理系统中a2a和mcp集成的批判性分析，arXiv预印本arXiv:2505.03864，2025年。URL https://arxiv.org/abs/2505.03864v1.

[617] 戎胜 李, 徐金, 曹志雄, 郑海涛, 和 金鸿吉. 通过分段基础调整扩展大型语言模型的上下文窗口用于旋转位置嵌入. 应用科学, 2024.

[618] 李帅科, 张凯, 刘奇, 和 陈恩宏. Mindbridge: 通过记忆增强模态实现可扩展和跨模型知识编辑, arXiv 预印本 arXiv:2503.02701v1, 2025. URL https://arxiv.org/abs/2503.02701v1.

[619] 李帅毅, 张志松, 邓阳, 邓晨龙, 方天庆, 张红明, 米海涛, 余东, 和 林伟. Incomes: 将压缩和选择机制集成到 llms 中以实现高效的模型编辑, arXiv 预印本 arXiv:2505.22156, 2025. URL https://arxiv.org abs/2505.22156v1.

[620] 李思恒, 杨成, 程泽森, 刘乐毛, 余墨, 杨宇宇, 和 林伟蓝. 大型语言模型可以在长上下文推理中自我改进, arXiv 预印本 arXiv:2411.08147, 2024. URLhttps://arxiv.org/abs/2411.08147v1.

[621] 李晓, 邹华, 和 刘鹏. Torl: 扩展工具集成 RL. 2025. URL https://arxiv.org/abs/2503.23383.

[622] 李晓鹏, 贾鹏越, 许德荣, 温怡, 张颖颖, 张文林, 王万宇, 王奕超, 杜赵晨, 李向阳, 刘勇, 郭惠峰, 唐瑞明, 和 赵祥宇. 关于个性化的调查：从 RAG 到代理, arXiv 预印本 arXiv:2504.10147, 2025. URLhttps://arxiv.org/abs/2504.10147v1.

[623] 李小溪, 金嘉杰, 董冠廷, 钱宏进, 朱宇韬, 吴永康, 温继荣, 和窦志成. Webthinker: 用深度研究能力赋能大型推理模型, arXiv preprint arXiv:2504.21776, 2025. URLhttps://arxiv.org/abs/2504.21776v1.

[624] 李欣, 储启之, 陈宇斌, 刘杨, 刘瑶琪, 余泽凯, 陈伟泽, 钱程, 石川, 和杨成. Graphteam: 通过多智能体协作促进基于大型语言模型的图分析, arXiv preprint arXiv:2410.18032v4, 2024. URL https://arxiv.org/abs/2410.18032v4.

[625] 李新怡, 王赛, 曾思琪, 吴宇, 和杨怡. 关于基于llm的多智能体系统: 工作流、基础设施和挑战. *Vicinagearth*, 2024.

[626] Xinze Li, Zhenghao Liu, Chenyan Xiong, Shi Yu, Yu Gu, Zhiyuan Liu, and Ge Yu. Structure-aware language model pretraining improves dense retrieval on structured data. *Annual Meeting of the Association for Computational Linguistics*, 2023.

[627] Yang Li, Jiacong He, Xiaoxia Zhou, Yuan Zhang, and Jason Baldridge. Mapping natural language instructions to mobile ui action sequences. *Annual Meeting of the Association for Computational Linguistics*, 2020.

[628] Yinghao Li, R. Ramprasad, and Chao Zhang. A simple but effective approach to improve structured language model output for information extraction. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[629] Yuan Li, Yixuan Zhang, and Lichao Sun. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents, arXiv preprint arXiv:2310.06500, 2023. URL https://arxiv.org/abs/2310.06500.

[630] Yucheng Li. Unlocking context constraints of llms: Enhancing context efficiency of llms with self-information-based content filtering, arXiv preprint arXiv:2304.12102, 2023. URL https://arxiv.org/abs/2304.12102v1.

[631] Yucheng Li, Bo Dong, Chenghua Lin, and Frank Guerin. Compressing context to enhance inference efficiency of large language models. *Conference on Empirical Methods in Natural Language Processing*, 2023.

[632] Zhaoxin Li, Xiaoming Zhang, Haifeng Zhang, and Chengxiang Liu. Refining interactions: Enhancing anisotropy in graph neural networks with language semantics, arXiv preprint arXiv:2504.01429, 2025. URL https://arxiv.org/abs/2504.01429v1.

[633] Zhecheng Li, Yiwei Wang, Bryan Hooi, Yujun Cai, Naifan Cheung, Nanyun Peng, and Kai-Wei Chang. Think carefully and check again! meta-generation unlocking llms for low-resource cross-lingual summarization. 2024.

[634] Zhecheng Li, Yiwei Wang, Bryan Hooi, Yujun Cai, Nanyun Peng, and Kai-Wei Chang. Drs: Deep question reformulation with structured output. In *Association for Computational Linguistics ACL, 2025.*, 2024.

[635] Zhecheng Li, Yiwei Wang, Bryan Hooi, Yujun Cai, Zhen Xiong, Nanyun Peng, and Kai-Wei Chang. Vulnerability of llms to vertically aligned text manipulations. In *Association for Computational Linguistics ACL, 2025.*, 2024.

[636] Zhecheng Li, Guoxian Song, Yujun Cai, Zhen Xiong, Junsong Yuan, and Yiwei Wang. Texture or semantics? vision-language models get lost in font recognition. In *Conference on Language Modeling COLM, 2025.*, 2025.

[637] Zhiyu Li, Shichao Song, Hanyu Wang, Simin Niu, Ding Chen, Jiawei Yang, Chenyang Xi, Huayi Lai, Jihao Zhao, Yezhaohui Wang, Junpeng Ren, Zehao Lin, Jiahao Huo, Tianyi Chen, Kai Chen, Ke-Rong Li, Zhiqiang Yin, Qingchen Yu, Bo Tang, Hongkang Yang, Zhiyang Xu, and Feiyu Xiong. Memos: An operating system for memory-augmented generation (mag) in large language models, arXiv preprint arXiv:2505.22101, 2025. URL https://arxiv.org/abs/2505.22101v1.

[638] Zhong-Zhi Li, Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. Lans: A layout-aware neural solver for plane geometry problem. 2023.

[626] 李新泽, 刘正浩, 熊晨岩, 余石, 古宇, 刘志远, 和余格. 基于结构感知的语言模型预训练提升结构化数据上的密集检索. 计算语言学协会年会, 2023.

[627] 李杨, 何家聪, 周晓霞, 张远, 和Jason Baldridge. 将自然语言指令映射到移动UI操作序列. 计算语言学协会年会, 2020.

[628] 李英浩, R. Ramprasad, 和张超. 一种简单但有效的方法来改进结构化语言模型输出以用于信息抽取. 自然语言处理经验方法会议, 2024.

[629] 李元, 张艺轩, 和孙立超. 元代理: 通过协作生成代理模拟人类行为交互以实现基于LLM的任务导向协调, arXiv预印本arXiv:2310.06500, 2023. URLhttps://arxiv.org/abs/2310.06500.

[630] 李玉成. 解锁大语言模型的上下文约束：基于自信息内容过滤提升大语言模型的上下文效率, arXiv 预印本 arXiv:2304.12102，2023。URLhttps://arxiv.org/abs/2304.12102v1.

[631] 李玉成, 董波, 林成华, 和 Frank Guerin. 压缩上下文以提升大语言模型的推理效率。自然语言处理经验方法会议, 2023。

[632] 李赵欣, 张小明, 张海峰, 和 刘成翔. 优化交互：利用语言语义增强图神经网络的各向异性，arXiv 预印本 arXiv:2504.01429，2025。URLhttps://arxiv.org/abs/2504.01429v1.

[633] 李哲成, 王一伟, Bryan Hooi, 蔡宇君, 钟乃凡, 彭南云, 和 Chang Kai-Wei. 仔细思考并再次检查! 元生成解锁低资源跨语言摘要的大语言模型。2024。

[634] 李哲成, 王一伟, 胡斌浩, 蔡宇君, 彭南云, 和 张凯伟. Drs: 基于结构化输出的深度问题重述. 在 计算语言学协会 *ACL, 2025.*, 2024.

[635] 李哲成, 王一伟, 胡斌浩, 蔡宇君, 邢振, 彭南云, 和 张凯伟. 大型语言模型对垂直对齐文本操作的脆弱性. 在 计算语言学协会 *ACL, 2025.*, 2024.

[636] 李哲成, 宋国先, 蔡宇君, 邢振, 袁俊松, 和 王一伟. 纹理还是语义？视觉语言模型在字体识别中迷失. 在 语言建模会议 *COLM, 2025.*, 2025.

[637] 李志宇, 宋时超, 王汉宇, 牛思敏, 陈丁, 杨嘉伟, 西晨阳, 赖华怡, 赵继浩, 王叶浩, 任俊鹏, 林泽浩, 胡嘉豪, 陈天一, 陈凯, 李克荣, 殷志强, 余清尘, 唐波, 杨宏康, 许志阳, 和 邢飞宇. Memos: 大型语言模型中增强记忆生成 (mag) 的操作系统, arXiv 预印本 arXiv:2505.22101, 2025.
URLhttps://arxiv.org/abs/2505.22101v1.

[638] 李中志、张明良、尹飞、刘成林。Lans：一种用于平面几何问题的布局感知神经求解器。2023。

[639] Zhong-Zhi Li, Ming-Liang Zhang, Fei Yin, Zhi-Long Ji, Jin-Feng Bai, Zhen-Ru Pan, Fan-Hu Zeng, Jian Xu, Jia-Xin Zhang, and Cheng-Lin Liu. Cmmath: A chinese multi-modal math skill evaluation benchmark for foundation models. 2024.

[640] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. From system 1 to system 2: A survey of reasoning large language models. 2025.

[641] Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. Structrag: Boosting knowledge intensive reasoning of llms via inference-time hybrid information structurization. *International Conference on Learning Representations*, 2024.

[642] Zinuo Li, Xian Zhang, Yongxin Guo, Mohammed Bennamoun, F. Boussaid, Girish Dwivedi, Luqi Gong, and Qiuhong Ke. Watch and listen: Understanding audio-visual-speech moments with multimodal llm, arXiv preprint arXiv:2505.18110v2, 2025. URL https://arxiv.org/abs/2505.18110v2.

[643] Zixuan Li, Jing Xiong, Fanghua Ye, Chuanyang Zheng, Xun Wu, Jianqiao Lu, Zhongwei Wan, Xiaodan Liang, Chengming Li, Zhenan Sun, Lingpeng Kong, and Ngai Wong. Uncertaintyrag: Span-level uncertainty enhanced long-context modeling for retrieval-augmented generation, arXiv preprint arXiv:2410.02719, 2024. URL https://arxiv.org/abs/2410.02719v1.

[644] Zonglin Li, Ruiqi Guo, and Surinder Kumar. Decoupled context processing for context augmented language modeling. *Neural Information Processing Systems*, 2022.

[645] Zongqian Li, Yinhong Liu, Yixuan Su, and Nigel Collier. Prompt compression for large language models: A survey. *North American Chapter of the Association for Computational Linguistics*, 2024.

[646] Wen li Yu and Junfeng Zhao. Quantum multi-agent reinforcement learning as an emerging ai technology: A survey and future directions. *International Conferences on Computing Advancements*, 2023.

[647] Guannan Liang and Qianqian Tong. Llm-powered ai agent systems and their applications in industry, arXiv preprint arXiv:2505.16120, 2025. URL https://arxiv.org/abs/2505.16120v1.

[648] Jintao Liang, Gang Su, Huifeng Lin, You Wu, Rui Zhao, and Ziyue Li. Reasoning rag via system 1 or system 2: A survey on reasoning agentic retrieval-augmented generation for industry challenges, arXiv preprint arXiv:2506.10408, 2025. URL https://arxiv.org/abs/2506.10408v1.

[649] Xinnian Liang, Bing Wang, Huijia Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. Scm: Enhancing large language model with self-controlled memory framework, arXiv preprint arXiv:2304.13343, 2023. URL https://arxiv.org/abs/2304.13343v4.

[650] Xuechen Liang, Meiling Tao, Yinghui Xia, Tianyu Shi, Jun Wang, and Jingsong Yang. Self-evolving agents with reflective and memory-augmented abilities, arXiv preprint arXiv:2409.00872, 2024. URL https://arxiv.org/abs/2409.00872v2.

[651] Xuechen Liang, Meiling Tao, Yinghui Xia, Jianhui Wang, Kun Li, Yijin Wang, Jingsong Yang, Tianyu Shi, Yuantao Wang, Miao Zhang, and Xueqian Wang. Mars: Memory-enhanced agents with reflective self-improvement, arXiv preprint arXiv:2503.19271, 2025. URL https://arxiv.org/abs/2503.19271v2.

[639] 钟志立, 张明良, 尹飞, 季志龙, 白金峰, 潘振如, 曾帆胡, 许健, 张嘉欣, 刘成林. Cmmath: 一个面向基础模型的中文多模态数学技能评估基准. 2024.

[640] 钟志立, 张杜珍, 张明良, 张嘉欣, 刘增岩, 姚宇轩, 许浩天, 郑俊豪, 王培杰, 陈秀怡, 等. 从系统1到系统2：推理大语言模型综述. 2025.

[641] 李卓群, 陈宣昂, 余海阳, 林红宇, 陆瑶杰, 唐巧宇, 黄飞, 韩先培, 孙乐, 李永斌. Structrag: 通过推理时混合信息结构化提升大语言模型的知识密集型推理能力. 学习表征国际会议, 2024.

[642] 李子诺, 张献, 郭永新, Mohammed Bennamoun, F. Boussaid, Girish Dwivedi, Gong Luqi, 和 Ke Qiuhong. 观察与聆听：使用多模态LLM理解视听语音时刻, arXiv预印本 arXiv:2505.18110v2, 2025. URL https://arxiv.org/abs/2505.18110v2.

[643] 李子轩, 熊静, 叶方华, 郑传阳, 吴迅, 陆建桥, 万中伟, 梁晓丹, 李成明, 孙振安, 孔令鹏, 及 Ngai Wong. Uncertaintyrag: 用于检索增强生成的基于片段级不确定性的长上下文建模, arXiv 预印本 arXiv:2410.02719, 2024. URLhttps://arxiv.org/abs/2410.02719v1.

[644] 李宗林, 郭瑞琪, 和 苏瑞·库马尔. 解耦的上下文处理用于上下文增强语言建模. 神经信息处理系统, 2022.

[645] 李宗倩, 刘银红, 苏一轩, 和 Nigel Collier. 大型语言模型的提示压缩：一项调查. 美国计算语言学协会北美分会, 2024.

[646] 余文莉 和 赵军峰. 量子多智能体强化学习作为一种新兴的人工智能技术：一项调查和未来方向. 国际计算进展会议, 2023.

[647] 梁冠南 和 葛倩倩. 基于大型语言模型的AI智能体系统及其在工业中的应用, arXiv预印本 arXiv:2505.16120, 2025. URLhttps://arxiv.org/abs/2505.16120v1.

[648] 梁金涛, 苏刚, 林惠峰, 吴友, 赵瑞, 和 李子越. 系统一或系统二通过推理RAG：一项关于推理智能体检索增强生成用于行业挑战的调查, arXiv预印本 arXiv:2506.10408, 2025. URLhttps://arxiv.org/abs/2506.10408v1.

[649] 李新年, 王兵, 黄会嘉, 吴双知, 吴培浩, 陆陆, 马泽军, 和 李周军. Scm: 基于自控记忆框架增强大型语言模型, arXiv 预印本 arXiv:2304.13343, 2023. URLhttps://arxiv.org/abs/2304.13343v4.

[650] 梁学臣, 陶美玲, 夏英会, 石天宇, 王军, 和 杨景松. 具备反思和记忆增强能力的自进化智能体, arXiv 预印本 arXiv:2409.00872, 2024. URL https://arxiv.org/abs/2409.00872v2.

[651] 梁学臣, 陶美玲, 夏英会, 王建辉, 李坤, 王一锦, 杨景松, 石天宇, 王元涛, 张苗, 和 王雪倩. Mars: 具备记忆增强和反思自改进的智能体, arXiv 预印本 arXiv:2503.19271, 2025. URLhttps://arxiv.org/abs/2503.19271v2.

[652] Yanbiao Liang, Huihong Shi, Haikuo Shao, and Zhongfeng Wang. Accllm: Accelerating long-context llm inference via algorithm-hardware co-design, arXiv preprint arXiv:2505.03745, 2025. URL https://arxiv.org/abs/2505.03745v1.

[653] Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yangyiwen Ou, Shuai Lu, Lei Ji, Shaoguang Mao, Yun Wang, Linjun Shou, Ming Gong, and Nan Duan. Taskmatrix.ai: Completing tasks by connecting foundation models with millions of apis. *Intelligent Computing*, 2023.

[654] Yiming Liang, Ge Zhang, Xingwei Qu, Tianyu Zheng, Jiawei Guo, Xinrun Du, Zhenzhu Yang, Jiaheng Liu, Chenghua Lin, Lei Ma, Wenhao Huang, and Jiajun Zhang. I-sheep: Self-alignment of llm from scratch through an iterative self-enhancement paradigm. arXiv preprint, 2024.

[655] Bingli Liao and Danilo Vasconcellos Vargas. Beyond kv caching: Shared attention for efficient llms. *Neurocomputing*, 2024.

[656] Xiaoxuan Liao, Binrong Zhu, Jacky He, Guiran Liu, Hongye Zheng, and Jia Gao. A fine-tuning approach for t5 using knowledge graphs to address complex tasks, arXiv preprint arXiv:2502.16484, 2025. URL https://arxiv.org/abs/2502.16484v1.

[657] David Lillis. Internalising interaction protocols as first-class programming elements in multi agent systems, arXiv preprint arXiv:1711.02634, 2017. URL https://arxiv.org/abs/1711.02634v1.

[658] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. Kagnet: Knowledge-aware graph networks for commonsense reasoning. *Conference on Empirical Methods in Natural Language Processing*, 2019.

[659] Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. *Annual Meeting of the Association for Computational Linguistics*, 2021.

[660] Bin Lin, Chen Zhang, Tao Peng, Hanyu Zhao, Wencong Xiao, Minmin Sun, Anmin Liu, Zhipeng Zhang, Lanbo Li, Xiafei Qiu, Shen Li, Zhigang Ji, Tao Xie, Yong Li, and Wei Lin. Infinite-llm: Efficient llm service for long context with distattention and distributed kvcache, arXiv preprint arXiv:2401.02669, 2024. URL https://arxiv.org/abs/2401.02669.

[661] Jianhao Lin, Lexuan Sun, and Yixin Yan. Simulating macroeconomic expectations using llm agents, arXiv preprint arXiv:2505.17648, 2025. URL https://arxiv.org/abs/2505.17648v2.

[662] Lei Lin, Jiayi Fu, Pengli Liu, Qingyang Li, Yan Gong, Junchen Wan, Fuzheng Zhang, Zhongyuan Wang, Di Zhang, and Kun Gai. Just ask one more time! self-agreement improves reasoning of language models in (almost) all scenarios. *Annual Meeting of the Association for Computational Linguistics*, 2023.

[663] Matthieu Lin, Jenny Sheng, Andrew Zhao, Shenzhi Wang, Yang Yue, Yiran Wu, Huan Liu, Jun Liu, Gao Huang, and Yong-Jin Liu. Training of scaffolded language models with language supervision: A survey, arXiv preprint arXiv:2410.16392, 2024. URL https://arxiv.org/abs/2410.16392v2.

[664] Qiqiang Lin, Muning Wen, Qiuying Peng, Guanyu Nie, Junwei Liao, Jun Wang, Xiaoyun Mo, Jiamu Zhou, Cheng Cheng, Yin Zhao, and Weinan Zhang. Hammer: Robust function-calling for on-device language models via function masking, arXiv preprint arXiv:2410.04587, 2024. URL https://arxiv.org/abs/2410.04587v2.

[652] Yanbiao Liang, Huihong Shi, Haikuo Shao, and Zhongfeng Wang. Accllm: 通过算法-硬件协同设计加速长上下文 llm 推理, arXiv preprint arXiv:2505.03745, 2025. URLhttps://arxiv.org/abs/2505.03745v1.

[653] Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yangyiwen Ou, Shuai Lu, Lei Ji, Shaoguang Mao, Yun Wang, Linjun Shou, Ming Gong, and Nan Duan. Taskmatrix.ai: 通过连接数百万个 api 完成任务的基础模型. 智能计算, 2023.

[654] Yiming Liang, Ge Zhang, Xingwei Qu, Tianyu Zheng, Jiawei Guo, Xinrun Du, Zhenzhu Yang, Jiaheng Liu, Chenghua Lin, Lei Ma, Wenhao Huang, and Jiajun Zhang. I-sheep: 通过迭代自增强范式从头开始实现 llm 自对齐. arXiv preprint, 2024.

[655] Bingli Liao和Danilo Vasconcellos Vargas.超越kv缓存：共享注意力机制用于高效的llms。*Neurocomputing*, 2024.

[656] Xiaoxuan Liao, Binrong Zhu, Jacky He, Guiran Liu, Hongye Zheng,和Jia Gao.使用知识图谱对t5进行微调以解决复杂任务，arXiv预印本arXiv:2502.16484，2025。URLhttps://arxiv.org/abs/2502.16484v1.

[657] David Lillis.将交互协议作为一等编程元素内部化于多智能体系统中，arXiv预印本arXiv:1711.02634，2017。URL https://arxiv.org/abs/1711.02634v1.

[658] Bill Yuchen Lin, Xinyue Chen, Jamin Chen,和XiangRen. Kagnet：用于常识推理的知识感知图网络。自然语言处理经验方法会议, 2019。

[659] 林玉川, 李思颖, 乔晓阳, 和 任翔. 超越英语的常识: 评估和改进用于常识推理的多语言语言模型. 计算语言学协会年会,2021.

[660] 林斌, 张晨, 彭涛, 赵汉宇, 肖文聪, 孙敏敏, 刘安民, 张志鹏, 李兰波, 邱夏菲, 李神, 季志刚, 谢涛, 李勇, 和 林伟. Infinite-llm: 基于长上下文的高效llm服务, arXiv预印本 arXiv:2401.02669, 2024. URLhttps://arxiv.org/abs/2401.02669.

[661] 林建浩, 孙乐轩, 和 严奕欣. 使用llm代理模拟宏观经济预期, arXiv预印本 arXiv:2505.17648, 2025. URLhttps://arxiv.org/abs/2505.17648v2.

[662] 雷林, 傅佳怡, 刘鹏丽, 李清阳, 龚岩, 万俊辰, 张福正, 王中元, 张迪, 赵坤. 再问一次！自我协议提高了语言模型在（几乎）所有场景中的推理能力。 计算语言学协会年会, 2023.

[663] Matthieu Lin, Jenny Sheng, Andrew Zhao, 王森芝, 阎岳, 吴依兰, 刘欢, 刘军, 黄高, 刘永进. 基于语言监督的支架语言模型训练：一项调查， arXiv 预印本 arXiv:2410.16392, 2024. URLhttps://arxiv.org/abs/2410.16392v2.

[664] 林齐强, 温木宁, 彭秋莹, 聂冠宇, 廖俊伟, 王俊, 莫晓云, 周嘉木, 程程, 赵寅, 张伟楠. 锤子：通过函数掩码实现设备上语言模型的鲁棒函数调用， arXiv 预印本 arXiv:2410.04587, 2024. URL https://arxiv.org/abs/2410.04587v2.

[665] Yu-Chen Lin, Akhilesh Kumar, Norman Chang, Wen-Liang Zhang, Muhammad Zakir, Rucha Apte, Haiyang He, Chao Wang, and Jyh-Shing Roger Jang. Novel preprocessing technique for data embedding in engineering code generation using large language model. *2024 IEEE LLM Aided Design Workshop (LAD)*, 2023.

[666] Yu-Hsuan Lin, Qian-Hui Chen, Yi-Jie Cheng, Jia-Ren Zhang, Yi-Hung Liu, Liang-Yu Hsia, and Yun-Nung Chen. Llm inference enhanced by external knowledge: A survey, arXiv preprint arXiv:2505.24377, 2025. URL https://arxiv.org/abs/2505.24377v1.

[667] Jack W Lindsey and Ashok Litwin-Kumar. Selective consolidation of learning and memory via recall-gated plasticity. *bioRxiv*, 2024.

[668] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 2016.

[669] Gili Lior, Yuval Shalev, Gabriel Stanovsky, and Ariel Goldstein. Computation or weight adaptation? rethinking the role of plasticity in learning. *bioRxiv*, 2024.

[670] Bingyang Liu, Haoyi Zhang, Xiaohan Gao, Zichen Kong, Xiyuan Tang, Yibo Lin, Runsheng Wang, and Ru Huang. Layoutcopilot: An llm-powered multi-agent collaborative framework for interactive analog layout design. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2024.

[671] E. Liu, Kelvin Guu, Panupong Pasupat, Tianlin Shi, and Percy Liang. Reinforcement learning on web interfaces using workflow-guided exploration. *International Conference on Learning Representations*, 2018.

[672] Guang-Da Liu, Haitao Mao, Bochuan Cao, Zhiyu Xue, K. Johnson, Jiliang Tang, and Rongrong Wang. On the intrinsic self-correction capability of llms: Uncertainty and latent concept, arXiv preprint arXiv:2406.02378, 2024. URL https://arxiv.org/abs/2406.02378v2.

[673] Guangyi Liu, Yongqi Zhang, Yong Li, and Quanming Yao. Dual reasoning: A gnn-llm collaborative framework for knowledge graph question answering, arXiv preprint arXiv:2406.01145, 2024. URL https://arxiv.org/abs/2406.01145v2.

[674] Guangyi Liu, Pengxiang Zhao, Liang Liu, Yaxuan Guo, Han Xiao, Weifeng Lin, Yuxiang Chai, Yue Han, Shuai Ren, Hao Wang, Xiaoyu Liang, Wenhao Wang, Tianze Wu, Linghao Li, Guanjing Xiong, Yong Liu, and Hongsheng Li. Llm-powered gui agents in phone automation: Surveying progress and prospects, arXiv preprint arXiv:2504.19838, 2025. URL https://arxiv.org/abs/2504.19838v2.

[675] Hanchao Liu, Rong-Zhi Li, Weimin Xiong, Ziyu Zhou, and Wei Peng. Workteam: Constructing workflows from natural language with multi-agents. *North American Chapter of the Association for Computational Linguistics*, 2025.

[676] Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *International Conference on Learning Representations*, 2023.

[677] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Neural Information Processing Systems*, 2023.

[665] 林宇辰, Akhilesh Kumar, 张正阳, 张文良, Muhammad Zakir, Rucha Apte, 何海阳, 王超, 和 蒋志兴 Roger Jang. 一种用于工程代码生成中数据嵌入的大语言模型新颖预处理技术. *2024 IEEE LLMAided DesignWorkshop (LAD)*, 2023.

[666] 林宇轩, 陈倩惠, 成艺杰, 张嘉仁, 刘奕宏, 夏亮宇, 和陈云农. 基于外部知识增强的大语言模型推理:一项调查, arXiv preprint arXiv:2505.24377, 2025. URLhttps://arxiv.org/abs/2505.24377v1.

[667] 杰克 W Lindsey 和 Ashok Litwin-Kumar. 通过回忆门控可塑性选择性整合学习和记忆. *bioRxiv*,2024.

[668] Tal Linzen, Emmanuel Dupoux, 和 Yoav Goldberg. 评估 LSTM 学习句法敏感依赖的能力。 *Association for Computational Linguistics* 的交易记录, 2016。

[669] Gili Lior, Yuval Shalev, Gabriel Stanovsky, and Ariel Goldstein. Computation or weight adaptation? rethinking the role of plasticity in learning. *bioRxiv*, 2024.

[670] Bingyang Liu, Haoyi Zhang, Xiaohan Gao, Zichen Kong, Xiyuan Tang, Yibo Lin, Runsheng Wang, 和 Ru Huang. Layoutcopilot: 一个由 LLM 驱动的多代理协作框架，用于交互式模拟布局设计。 *IEEE Transactions on Computer-Aided Design of Integrated Circuits andSystems*, 2024。

[671] E. Liu, Kelvin Guu, Panupong Pasupat, Tianlin Shi, 和 Percy Liang. 使用工作流引导探索在 Web 界面上进行强化学习。 *InternationalConference on LearningRepresentations*, 2018。

[672] Guang-Da Liu, Haitao Mao, Bochuan Cao, Zhiyu Xue, K. Johnson, Jiliang Tang, 和 Rongrong Wang. 关于 LLM 的内在自我纠正能力：不确定性和潜在概念，arXiv 预印本 arXiv:2406.02378，2024。URLhttps://arxiv.org/abs/2406.02378v2.

[673] 刘广毅, 张永奇, 李永, 姚全明. 双推理: 一种用于知识图谱问答的gnn-llm协同框架, arXiv预印本 arXiv:2406.01145, 2024. URLhttps://arxiv.org/abs/2406.01145v2.

[674] 刘广毅, 赵鹏翔, 刘亮, 郭亚轩, 肖汉, 林伟峰, 蔡宇翔, 韩越, 任帅, 王浩, 梁小宇, 王文豪, 吴天泽, 李凌浩, 熊冠景, 刘勇, 李红生. 基于llm的手机自动化gui代理: 调研进展与前景, arXiv预印本 arXiv:2504.19838, 2025. URLhttps://arxiv.org/abs/2504.19838v2.

[675] 刘汉超, 李荣志, 熊伟民, 周子宇, 彭伟. Workteam: 从自然语言构建工作流的多智能体系统. 美国计算语言学协会北美分会, 2025.

[676] 刘浩, Matei Zaharia, 和 Pieter Abbeel. 基于块状 Transformer 的环形注意力机制用于近乎无限的上下文. 国际学习表征会议, 2023.

[677] 刘浩天, 李春元, 吴庆阳, 和 李永才. 视觉指令微调. 神经信息处理系统会议, 2023.

[678] Jie Liu, Pan Zhou, Yingjun Du, Ah-Hwee Tan, Cees G. M. Snoek, J. Sonke, and E. Gavves. Capo: Cooperative plan optimization for efficient embodied multi-agent cooperation. *International Conference on Learning Representations*, 2024.

[679] Jun Liu, Ke Yu, Keliang Chen, Ke Li, Yuxinyue Qian, Xiaolian Guo, Haozhe Song, and Yinming Li. Acps: Agent collaboration protocols for the internet of agents, arXiv preprint arXiv:2505.13523, 2025. URL https://arxiv.org/abs/2505.13523v1.

[680] Junwei Liu, Kaixin Wang, Yixuan Chen, Xin Peng, Zhenpeng Chen, Lingming Zhang, and Yiling Lou. Large language model-based agents for software engineering: A survey, arXiv preprint arXiv:2409.02977, 2024. URL https://arxiv.org/abs/2409.02977v1.

[681] Kai Liu, Zhihang Fu, Chao Chen, Wei Zhang, Rongxin Jiang, Fan Zhou, Yao-Shen Chen, Yue Wu, and Jieping Ye. Enhancing llm's cognition via structurization. *Neural Information Processing Systems*, 2024.

[682] Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. Think-in-memory: Recalling and post-thinking enable llms with long-term memory. arXiv preprint, 2023.

[683] Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. Think-in-memory: Recalling and post-thinking enable llms with long-term memory, arXiv preprint arXiv:2311.08719, 2023. URL https://arxiv.org/abs/2311.08719.

[684] Na Liu, Liangyu Chen, Xiaoyu Tian, Wei Zou, Kaijiang Chen, and Ming Cui. From llm to conversational agent: A memory enhanced architecture with fine-tuning of large language models, arXiv preprint arXiv:2401.02777, 2024. URL https://arxiv.org/abs/2401.02777v2.

[685] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, F. Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 2023.

[686] Pei Liu, Xin Liu, Ruoyu Yao, Junming Liu, Siyuan Meng, Ding Wang, and Jun Ma. Hm-rag: Hierarchical multi-agent multimodal retrieval augmented generation. arXiv preprint, 2025.

[687] Shicheng Liu, Jialiang Xu, Wesley Tjangnaka, Sina J. Semnani, Chen Jie Yu, Gui D'avid, and Monica S. Lam. Suql: Conversational search over structured and unstructured data with large language models. *NAACL-HLT*, 2023.

[688] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. 2025.

[689] W Liu, X Huang, X Zeng, X Hao, S Yu, and D Li.... Toolace: Winning the points of llm function calling. 2024. URL https://arxiv.org/abs/2409.00920.

[690] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. *AAAI Conference on Artificial Intelligence*, 2019.

[678] Jie Liu, Pan Zhou, Yingjun Du, Ah-Hwee Tan, Cees G. M. Snoek, J. Sonke, and E. Gavves. Capo: 协作计划优化用于高效的具身多智能体合作。 国际学习表征会议, 2024。

[679] Jun Liu, Ke Yu, Keliang Chen, Ke Li, Yuxinyue Qian, Xiaolian Guo, Haozhe Song, andYinmingLi. Acps: 智能体互联网的智能体协作协议，arXiv 预印本 arXiv:2505.13523，2025。
URLhttps://arxiv.org/abs/2505.13523v1.

[680] Junwei Liu, Kaixin Wang, Yixuan Chen, Xin Peng, Zhenpeng Chen, Lingming Zhang, and Yil- ing Lou. 基于大型语言模型的软件工程智能体：一项调查，arXiv 预印本 arXiv:2409.02977，2024。
URLhttps://arxiv.org/abs/2409.02977v1.

[681] Kai Liu, Zhihang Fu, Chao Chen, Wei Zhang, Rongxin Jiang, Fan Zhou, Yao-Shen Chen, Yue Wu, and Jieping Ye. 通过结构化增强大语言模型的认知能力。 *Neural Information Processing Systems*, 2024。

[682] Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. 思维内存：回忆和后思考使大语言模型具有长期记忆。 arXiv preprint, 2023。

[683] Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. 思维内存：回忆和后思考使大语言模型具有长期记忆，arXiv preprint arXiv:2311.08719, 2023。
URLhttps://arxiv.org/abs/2311.08719.

[684] Na Liu, Liangyu Chen, Xiaoyu Tian, Wei Zou, Kaijiang Chen, and Ming Cui. 从 llm 到对话代理：一种具有大语言模型微调的记忆增强架构，arXiv 预印本 arXiv:2401.02777，2024。
URLhttps://arxiv.org/abs/2401.02777v2.

[685] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, F. Petroni, and Percy Liang. 迷失在中间：语言模型如何使用长上下文。 计算语言学协会汇刊, 2023。

[686] Pei Liu, Xin Liu, Ruoyu Yao, Junming Liu, Siyuan Meng, Ding Wang, and Jun Ma. Hm-rag：分层多智能体多模态检索增强生成。 arXiv 预印本，2025。

[687] 刘世成, 许嘉良, Tjangnaka Wesley, Sina J.Semnani, 程杰宇, 戴贵, 和 Lam MonicaS. 苏奎: 基于大型语言模型的结构化和非结构化数据对话式搜索。*NAACL-HLT*, 2023。

[688] 刘世宇, 韩宇成, 邢鹏, 尹福昆, 王瑞, 程伟, 廖家奇, 王英明, 傅洪浩, 韩春瑞, 等. Step1x-edit: 一种通用的图像编辑框架。 2025。

[689] W 刘, 黄 X, 曾 X, Hao X, Yu S, 和 Li D.. .. Toolace: 赢得 llm 函数调用的关键点。 2024.
URLhttps://arxiv.org/abs/2409.00920.

[690] 刘伟杰, 周鹏, 赵哲, 王志如, 巨奇, 邓浩堂, 和 王平. K-bert: 基于知识图谱的语言表示。*AAAI* 人工智能会议, 2019。

[691] Weiwen Liu, Xu Huang, Xingshan Zeng, Xinlong Hao, Shuai Yu, Dexun Li, Shuai Wang, Weinan Gan, Zhengying Liu, Yuanqing Yu, Zezhong Wang, Yuxian Wang, Wu Ning, Yutai Hou, Bin Wang, Chuhan Wu, Xinzhi Wang, Yong Liu, Yasheng Wang, Duyu Tang, Dandan Tu, Lifeng Shang, Xin Jiang, Ruiming Tang, Defu Lian, Qun Liu, and Enhong Chen. Toolace: Winning the points of llm function calling, arXiv preprint arXiv:2409.00920, 2024. URL https://arxiv.org/abs/2409.00920v1.

[692] Weiwen Liu, Jiarui Qin, Xu Huang, Xingshan Zeng, Yunjia Xi, Jianghao Lin, Chuhan Wu, Yasheng Wang, Lifeng Shang, Ruiming Tang, Defu Lian, Yong Yu, and Weinan Zhang. The real barrier to llm agent usability is agentic roi, arXiv preprint arXiv:2505.17767, 2025. URL https://arxiv.org/abs/2505.17767v1.

[693] Wentao Liu, Hanglei Hu, Jie Zhou, Yuyang Ding, Junsong Li, Jiayi Zeng, Mengliang He, Qin Chen, Bo Jiang, Aimin Zhou, and Liang He. Mathematical language models: A survey, arXiv preprint arXiv:2312.07622, 2023. URL https://arxiv.org/abs/2312.07622v4.

[694] Wentao Liu, Ruohua Zhang, Aimin Zhou, Feng Gao, and JiaLi Liu. Echo: A large language model with temporal episodic memory, arXiv preprint arXiv:2502.16090, 2025. URL https://arxiv.org/abs/2502.16090v1.

[695] Xu Liu, S. Ramirez, Petti T. Pang, C. Puryear, A. Govindarajan, K. Deisseroth, and S. Tonegawa. Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature*, 2012.

[696] Yang Liu, Xiaobin Tian, Zequn Sun, and Wei Hu. Finetuning generative large language models with discrimination instructions for knowledge graph completion. In *International Semantic Web Conference*, 2024.

[697] Yue Liu, Jiaying Wu, Yufei He, Hongcheng Gao, Hongyu Chen, Baolong Bi, Jiaheng Zhang, Zhiqi Huang, and Bryan Hooi. Efficient inference for large reasoning models: A survey, arXiv preprint arXiv:2503.23077, 2025. URL https://arxiv.org/abs/2503.23077v2.

[698] Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. A dynamic llm-powered agent network for task-oriented agent collaboration, arXiv preprint arXiv:2310.02170, 2023. URL https://arxiv.org/abs/2310.02170v2.

[699] Zijun Liu, Zhennan Wan, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. Scaling external knowledge input beyond context windows of llms via multi-agent collaboration, arXiv preprint arXiv:2505.21471, 2025. URL https://arxiv.org/abs/2505.21471v1.

[700] Zinan Liu, Haoran Li, Jingyi Lu, Gaoyuan Ma, Xu Hong, Giovanni Iacca, Arvind Kumar, Shaojun Tang, and Lin Wang. Nature's insight: A novel framework and comprehensive analysis of agentic reasoning through the lens of neuroscience. arXiv preprint, 2025.

[701] Zuxin Liu, Thai Hoang, Jianguo Zhang, Ming Zhu, Tian Lan, Shirley Kokane, Juntao Tan, Weiran Yao, Zhiwei Liu, Yihao Feng, Rithesh Murthy, Liangwei Yang, Silvio Savarese, Juan Carlos Niebles, Huan Wang, Shelby Heinecke, and Caiming Xiong. Apigen: Automated pipeline for generating verifiable and diverse function-calling datasets. *Neural Information Processing Systems*, 2024.

[702] Leo S. Lo. The art and science of prompt engineering: A new literacy in the information age. *Internet Reference Services Quarterly*, 2023.

[691] 刘伟文，黄旭，曾兴山，郝新龙，余帅，李德勋，王帅，甘伟男，刘正莹，余元庆，王泽中，王宇娴，宁武，侯宇泰，王斌，吴楚涵，王新芝，刘勇，王亚胜，唐都宇，涂丹丹，尚立峰，蒋欣，唐瑞明，连德福，刘群，陈恩红. Toolace: 赢得 llm 函数调用的分数，arXiv 预印本 arXiv:2409.00920, 2024. URLhttps://arxiv.org/abs/2409.00920v1.

[692] 刘伟文, 秦嘉瑞, 黄旭, 曾兴山, 西云嘉, 林江浩, 吴楚涵, 王亚胜, 尚立峰, 唐瑞明, 连德福, 余勇, 张伟楠. 大语言模型代理可用性的真正障碍是代理ROI, arXiv预印本arXiv:2505.17767, 2025. URLhttps://arxiv.org/abs/2505.17767v1.

[693] 刘文涛，胡汉雷，周杰，丁宇阳，李俊松，曾嘉怡，何梦亮，陈秦，江波，周爱民，何亮. 数学语言模型：综述，arXiv预印本arXiv:2312.07622，2023. URLhttps://arxiv.org/abs/2312.07622v4.

[694] 刘文涛，张若华，周爱民，高峰，刘佳丽. Echo：一个具有时间情景记忆的大型语言模型，arXiv预印本arXiv:2502.16090，2025. URLhttps://arxiv.org/abs/2502.16090v1.

刘旭，S. Ramírez，Petti T. Pang，C. Puryear，A. Govindarajan，K. Deisseroth 和 S. Tonegawa. 光遗传刺激海马体编码激活恐惧记忆回忆。*Nature*, 2012.

[696] 杨柳, 田晓斌, 孙泽群, 和 胡伟. 基于区分指令微调生成式大语言模型用于知识图谱补全. 在 国际语义网会议, 2024.

[697] 刘越, 吴佳颖, 何宇飞, 高洪成, 陈红宇, 毕宝龙, 张嘉恒, 黄志奇, 和 胡锦辉. 大推理模型的推理优化: 一份调查, arXiv 预印本 arXiv:2503.23077, 2025. URLhttps://arxiv.org/abs/2503.23077v2.

[698] 刘子君, 张岩哲, 李鹏, 杨柳, 和 杨迪一. 一种基于动态 llm 的智能体网络用于任务导向的智能体协作, arXiv 预印本 arXiv:2310.02170, 2023. URLhttps://arxiv.org/abs/2310.02170v2.

[699] 刘子君, 万振南, 李鹏, 阎明, 张继, 黄飞, 刘阳. 通过多智能体协作扩展大型语言模型上下文窗口之外的外部知识输入, arXiv 预印本 arXiv:2505.21471, 2025. URLhttps://arxiv.org/abs/2505.21471v1.

[700] 刘子南, 李浩然, 陆静怡, 马高远, 霍旭, Giovanni Iacca, Arvind Kumar, 唐少军, 王琳. 自然之洞见：通过神经科学视角对代理推理的新框架和综合分析. arXiv 预印本, 2025.

[701] 刘祖欣, Thai Hoang, 张建国, 朱明, 兰天, Shirley Kokane, 谭俊涛, 姚伟然, 刘志伟, 冯一浩, Rithesh Murthy, 杨亮伟, Silvio Savarese, Niebles Juan Carlos, 王欢, Shelby Heinecke, and 邢-caiming. Apigen：用于生成可验证和多样化函数调用数据集的自动化流程. 神经信息处理系统, 2024.

[702] Leo S. Lo. 提示工程的艺术与科学：信息时代的新素养. 互联网参考服务季刊, 2023.

[703] Joseph R. Loffredo and Suyeol Yun. Agent-enhanced large language models for researching political institutions, arXiv preprint arXiv:2503.13524, 2025. URL https://arxiv.org/abs/2503.13524v1.

[704] Jianqiao Lu, Wanjun Zhong, Wenyong Huang, Yufei Wang, Fei Mi, Baojun Wang, Weichao Wang, Lifeng Shang, and Qun Liu. Self: Self-evolution with language feedback, arXiv preprint arXiv:2310.00533, 2023. URL https://arxiv.org/abs/2310.00533v4.

[705] Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation, arXiv preprint arXiv:2308.08239, 2023. URL https://arxiv.org/abs/2308.08239.

[706] Junting Lu, Zhiyang Zhang, Fangkai Yang, Jue Zhang, Lu Wang, Chao Du, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. Axis: Efficient human-agent-computer interaction with api-first llm-based agents, arXiv preprint arXiv:2409.17140, 2025. URL https://arxiv.org/abs/2409.17140.

[707] Keer Lu, Xiaonan Nie, Zheng Liang, Da Pan, Shusen Zhang, Keshi Zhao, Weipeng Chen, Zenan Zhou, Guosheng Dong, Bin Cui, and Wentao Zhang. Datasculpt: Crafting data landscapes for long-context llms through multi-objective partitioning, arXiv preprint arXiv:2409.00997, 2024. URL https://arxiv.org/abs/2409.00997v2.

[708] Liqiang Lu, Yicheng Jin, Hangrui Bi, Zizhang Luo, Peng Li, Tao Wang, and Yun Liang. Sanger: A co-design framework for enabling sparse attention using reconfigurable architecture. *Micro*, 2021.

[709] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Y. Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *Neural Information Processing Systems*, 2023.

[710] Y Lu, H Yu, and D Khashabi. Gear: Augmenting language models with generalizable and efficient tool resolution. 2023. URL https://arxiv.org/abs/2307.08775.

[711] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *Annual Meeting of the Association for Computational Linguistics*, 2021.

[712] Yinquan Lu, H. Lu, Guirong Fu, and Qun Liu. Kelm: Knowledge enhanced pre-trained language representations with message passing on hierarchical relational graphs, arXiv preprint arXiv:2109.04223, 2021. URL https://arxiv.org/abs/2109.04223v2.

[713] Elias Lumer, Anmol Gulati, V. K. Subbiah, Pradeep Honaganahalli Basavaraju, and James A. Burke. Scalemcp: Dynamic and auto-synchronizing model context protocol tools for llm agents, arXiv preprint arXiv:2505.06416, 2025. URL https://arxiv.org/abs/2505.06416v1.

[714] Cheng Luo, Jiawei Zhao, Zhuoming Chen, Beidi Chen, and Anima Anandkumar. Mini-sequence transformer: Optimizing intermediate memory for long sequences training, arXiv preprint arXiv:2407.15892, 2024. URL https://arxiv.org/abs/2407.15892v4.

[715] Feng Luo, Yu-Neng Chuang, Guanchu Wang, Hoang Anh Duy Le, Shaochen Zhong, Hongyi Liu, Jiayi Yuan, Yang Sui, Vladimir Braverman, Vipin Chaudhary, and Xia Hu. Autol2s: Auto long-short reasoning for efficient large language models, arXiv preprint arXiv:2505.22662, 2025. URL https://arxiv.org/abs/2505.22662v1.

[703] 约瑟夫·R·洛弗罗多和徐永烈. 用于研究政治制度的代理增强大型语言模型，arXiv预印本arXiv:2503.13524，2025年。URL https://arxiv.org/abs/2503.13524v1.

[704] Jianqiao Lu, Wanjun Zhong, Wenyong Huang, Yufei Wang, Fei Mi, Baojun Wang, Weichao Wang, Lifeng Shang, and Qun Liu. Self: Self-evolution with language feedback, arXiv preprint arXiv:2310.00533, 2023. URL https://arxiv.org/abs/2310.00533v4.

[705] 陆俊儒，安思语，林明宝，Gabriele Pergola，何雨兰，尹迪，孙行，吴云胜. Memochat：调整LLM使用备忘录以实现一致的长距离开放域对话，arXiv预印本arXiv:2308.08239，2023年。URLhttps://arxiv.org/abs/2308.08239.

[706] 陆俊廷，张志阳，杨方凯，张决，王路，杜超，林清伟，Saravan Rajmohan，张冬梅，张琪。Axis：基于api-first LLM代理的高效人机交互，arXiv预印本arXiv:2409.17140，2025年。URLhttps://arxiv.org/abs/2409.17140。

[707] Keer Lu, Xiaonan Nie, Zheng Liang, Da Pan, Shusen Zhang, Keshi Zhao, Weipeng Chen, Zenan Zhou, Guosheng Dong, Bin Cui, and Wentao Zhang. Datasculpt: 通过多目标划分构建长上下文LLM的数据景观, arXiv预印本 arXiv:2409.00997, 2024. URLhttps://arxiv.org/abs/2409.00997v2.

[708] 陆丽强, 金奕成, 毕航瑞, 罗子张, 李鹏, 王涛, 梁云. Sanger: 一种使用可重构架构实现稀疏注意力的协同设计框架. *Micro*, 2021.

[709] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Y. Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: 大型语言模型的即插即用组合推理. *NeuralInformationProcessing Systems*, 2023.

[710] Y Lu, H Yu, and D Khashabi. Gear: 使用通用且高效的工具解析增强语言模型. 2023. URLhttps://arxiv.org/abs/2307.08775.

[711] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 非常有序的提示及其位置：克服少样本提示顺序敏感性. *Association for Computational Linguistics* 年会, 2021.

[712] Yinquan Lu, H. Lu, Guirong Fu, and Qun Liu. Kelm: 基于层次关系图的信使传递知识增强预训练语言表示, arXiv preprint arXiv:2109.04223, 2021. URLhttps://arxiv.org/abs/2109.04223v2.

[713] Elias Lumer, Anmol Gulati, V. K. Subbiah, Pradeep Honaganahalli Basavaraju, and James A. Burke. Scalemcp: 动态和自动同步的模型上下文协议工具，用于 LLM 代理，arXiv 预印本 arXiv:2505.06416, 2025. URLhttps://arxiv.org/abs/2505.06416v1.

[714] 程路，赵继伟，陈卓明，陈北迪，Anima Anandkumar. Mini-sequence transformer: 优化长序列训练的中间内存，arXiv 预印本 arXiv:2407.15892，2024年。URLhttps://arxiv.org/abs/2407.15892v4.

[715] 冯罗，庄宇能，王观初，黎黄安杜伊，钟少辰，刘红毅，袁佳怡，隋阳，布拉维曼弗拉基米尔，乔迪普因查德哈里，和胡夏。Autol2s：自动长短期推理以高效大型语言模型，arXiv预印本arXiv:2505.22662，2025年。URLhttps://arxiv.org/abs/2505.22662v1。

[716] Haitong Luo, Xuying Meng, Suhang Wang, Tianxiang Zhao, Fali Wang, Hanyun Cao, and Yujun Zhang. Enhance graph alignment for large language models, arXiv preprint arXiv:2410.11370v1, 2024. URL https://arxiv.org/abs/2410.11370v1.

[717] Haoran Luo, E. Haihong, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, and Anh Tuan Luu. Hypergraphrag: Retrieval-augmented generation with hypergraph-structured knowledge representation. arXiv preprint, 2025.

[718] Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. arXiv preprint, 2025.

[719] Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, Rongcheng Tu, Xiaoming Luo, Wei Ju, Zhiping Xiao, Yifan Wang, Mengxue Xiao, Chenwu Liu, Jingyang Yuan, Shichang Zhang, Yiqiao Jin, Fan Zhang, Xianhong Wu, Hanqing Zhao, Dacheng Tao, Philip S. Yu, and Ming Zhang. Large language model agent: A survey on methodology, applications and challenges, arXiv preprint arXiv:2503.21460, 2025. URL https://arxiv.org/abs/2503.21460v1.

[720] Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. *International Conference on Learning Representations*, 2023.

[721] Renjie Luo, Jiaxi Li, Chen Huang, and Wei Lu. Through the valley: Path to effective long cot training for small language models, arXiv preprint arXiv:2506.07712, 2025. URL https://arxiv.org/abs/2506.07712v1.

[722] Xindi Luo, Zequn Sun, Jing Zhao, Zhe Zhao, and Wei Hu. Knowla: Enhancing parameter-efficient finetuning with knowledgeable adaptation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.

[723] Yifan Luo, Yiming Tang, Chengfeng Shen, Zhennan Zhou, and Bin Dong. Prompt engineering through the lens of optimal control. *Journal of Machine Learning*, 2023.

[724] Panagiotis Lymperopoulos and Vasanth Sarathy. Tools in the loop: Quantifying uncertainty of llm question answering systems that use tools, arXiv preprint arXiv:2505.16113, 2025. URL https://arxiv.org/abs/2505.16113v1.

[725] Yougang Lyu, Xiaoyu Zhang, Lingyong Yan, M. D. Rijke, Zhaochun Ren, and Xiuying Chen. Deepshop: A benchmark for deep research shopping agents, arXiv preprint arXiv:2506.02839, 2025. URL https://arxiv.org/abs/2506.02839v1.

[726] Jianxiang Ma. Research on the role of llm in multi-agent systems: A survey. *Applied and Computational Engineering*, 2024.

[727] Jie Ma, Zhitao Gao, Qianyi Chai, Wangchun Sun, Pinghui Wang, Hongbin Pei, Jing Tao, Lingyun Song, Jun Liu, Chen Zhang, and Li zhen Cui. Debate on graph: a flexible and reliable reasoning framework for large language models, arXiv preprint arXiv:2409.03155, 2024. URL https://arxiv.org/abs/2409.03155v1.

[716] Haitong Luo, Xuying Meng, Suhang Wang, Tianxiang Zhao, Fali Wang, Hanyun Cao, and Yujun Zhang. 增强大型语言模型的图对齐, arXiv preprint arXiv:2410.11370v1, 2024. URLhttps://arxiv.org/abs/2410.11370v1.

[717] Haoran Luo, E. Haihong, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, and Anh Tuan Luu. Hypergraphrag: 基于超图结构知识表示的检索增强生成. arXiv preprint, 2025.

[718] Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. O1-pruner: 针对O1类推理剪枝的长度协调微调. arXiv preprint, 2025.

[719] 罗军宇, 张伟志, 袁叶, 赵宇升, 杨俊伟, 顾一阳, 吴博文, 陈宾琪, 乔子越, 龙清清, 屠荣成, 罗小明, 巨伟, 肖志平, 王奕帆, 肖梦雪, 刘晨武, 袁景阳, 张世昌, 金一桥, 张帆, 吴仙红, 赵汉清, 陶大程, Philip S. Yu, 张明. 大语言模型代理：方法论、应用与挑战的调查, arXiv preprint arXiv:2503.21460, 2025. URLhttps://arxiv.org/abs/2503.21460v1.

[720] 罗林浩, 李元芳, 哈菲里·格哈姆雷扎, 和 潘世瑞. 图上的推理：忠实且可解释的大语言模型推理. 学习表示国际会议, 2023.

[721] 罗仁杰, 李嘉熙, 黄晨, 和 陆伟. 通过山谷：小型语言模型有效长序列训练的路径, arXiv 预印本 arXiv:2506.07712, 2025. URLhttps://arxiv.org/abs/2506.07712v1.

[722] 罗信迪, 孙泽群, 赵静, 赵哲, 和魏辉. Knowla: 通过知识适应增强参数高效的微调. 在《*2024*年北美计算语言学协会分会会议论文集》, 2024.

[723] 罗一帆, 唐毅明, 沈成峰, 周振南, 董斌. 通过最优控制视角进行提示工程. 机器学习杂志,2023.

[724] Panagiotis Lymperopoulos 和 Vasanth Sarathy. 在回路中的工具：量化使用工具的大型语言模型问答系统的不确定性，arXiv 预印本 arXiv:2505.16113，2025。URL https://arxiv.org/abs/2505.16113v1.

[725] 刘友刚，张晓宇，严凌勇，M. D. Rijke，任赵春，和陈秀英. Deepshop：深度研究购物代理的基准，arXiv预印本arXiv:2506.02839，2025. URLhttps://arxiv.org/abs/2506.02839v1.

[726] Jianxiang Ma. 关于大语言模型在多智能体系统中的作用研究：综述. 应用与计算工程, 2024.

[727] Jie Ma, Zhitao Gao, Qianyi Chai, Wangchun Sun, Pinghui Wang, Hongbin Pei, Jing Tao, Lingyun Song, Jun Liu, Chen Zhang, and Li zhen Cui. Debate on graph: a flexible and reliable reasoning framework for large language models, arXiv preprint arXiv:2409.03155, 2024. URL https://arxiv.org/abs/2409.03155v1.

[728] Qun Ma, Xiao Xue, Deyu Zhou, Xiangning Yu, Donghua Liu, Xuwen Zhang, Zihan Zhao, Yifan Shen, Peilin Ji, Juanjuan Li, Gang Wang, and Wanpeng Ma. Computational experiments meet large language model based agents: A survey and perspective, arXiv preprint arXiv:2402.00262, 2024. URL https://arxiv.org/abs/2402.00262v1.

[729] Xin Ma, Yang Liu, Jingjing Liu, and Xiaoxu Ma. Mesa-extrapolation: A weave position encoding method for enhanced extrapolation in llms. *Neural Information Processing Systems*, 2024.

[730] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting for retrieval-augmented large language models, arXiv preprint arXiv:2305.14283, 2023. URL https://arxiv.org/abs/2305.14283v3.

[731] Xuezhe Ma, Xiaomeng Yang, Wenhan Xiong, Beidi Chen, Lili Yu, Hao Zhang, Jonathan May, Luke S. Zettlemoyer, Omer Levy, and Chunting Zhou. Megalodon: Efficient llm pretraining and inference with unlimited context length. *Neural Information Processing Systems*, 2024.

[732] Y Ma, Z Gou, J Hao, R Xu, S Wang, and L Pan.... Sciagent: Tool-augmented language models for scientific reasoning. 2024. URL https://arxiv.org/abs/2402.11451.

[733] Zhiyuan Ma, Zhenya Huang, Jiayu Liu, Minmao Wang, Hongke Zhao, and Xin Li. Automated creation of reusable and diverse toolsets for enhancing llm reasoning. *AAAI Conference on Artificial Intelligence*, 2025.

[734] Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. Memory-assisted prompt editing to improve gpt-3 after deployment. *Conference on Empirical Methods in Natural Language Processing*, 2022.

[735] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, S. Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, A. Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. *Neural Information Processing Systems*, 2023.

[736] Xinji Mai, Haotian Xu, W. Xing, Weinong Wang, Yingying Zhang, and Wenqiang Zhang. Agent rl scaling law: Agent rl with spontaneous code execution for mathematical problem solving, arXiv preprint arXiv:2505.07773, 2025. URL https://arxiv.org/abs/2505.07773v2.

[737] Amjad Yousef Majid, Serge Saaybi, Tomas van Rietbergen, Vincent François-Lavet, R. V. Prasad, and Chris Verhoeven. Deep reinforcement learning versus evolution strategies: A comparative survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[738] D. Maldonado, Edison Cruz, Jackeline Abad Torres, P. Cruz, and Silvana del Pilar Gamboa Benitez. Multi-agent systems: A survey about its components, framework and workflow. *IEEE Access*, 2024.

[739] Zhao Mandi, Shreeya Jain, and Shuran Song. Roco: Dialectic multi-robot collaboration with large language models. *IEEE International Conference on Robotics and Automation*, 2023.

[740] Jeremy R. Manning, Sean M. Polyn, G. Baltuch, B. Litt, and M. Kahana. Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National Academy of Sciences of the United States of America*, 2011.

[728] Qun Ma, Xiao Xue, Deyu Zhou, Xiangning Yu, Donghua Liu, Xuwen Zhang, Zihan Zhao, Yifan Shen, Peilin Ji, Juanjuan Li, Gang Wang, and Wanpeng Ma. 计算实验与基于大型语言模型的智能体：综述与展望，arXiv 预印本 arXiv:2402.00262，2024。URL https://arxiv.org/abs/2402.00262v1.

[729] Xin Ma, Yang Liu, Jingjing Liu, and Xiaoxu Ma. Mesa-extrapolation：一种用于增强大型语言模型外推能力的编织位置编码方法。*NeuralInformationProcessingSystems*, 2024。

[730] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 用于检索增强型大型语言模型的查询重写，arXiv 预印本 arXiv:2305.14283，2023。URLhttps://arxiv.org/abs/2305.14283v3.

[731] Xuezhe Ma, Xiaomeng Yang, Wenhan Xiong, Beidi Chen, Lili Yu, Hao Zhang, Jonathan May, Luke S. Zettlemoyer, Omer Levy, and Chunting Zhou. Megalodon: Efficient llm pretraining and inference with unlimited context length. *NeuralInformationProcessingSystems*, 2024.

[732] Y Ma, Z Gou, J Hao, R Xu, S Wang, and L Pan. . . . Sciagent: 基于工具增强的语言模型用于科学推理. 2024. URLhttps://arxiv.org/abs/2402.11451.

[733] 马智远，黄振亚，刘佳宇，王民茂，赵宏科，李欣。为增强大型语言模型推理能力而自动创建可重用和多样化的工具集。*AAAI*人工智能会议, 2025.

[734] Aman Madaan、Niket Tandon、Peter Clark 和 Yiming Yang。部署后使用记忆辅助提示编辑来改进 GPT-3。自然语言处理经验方法会议，2022年。

[735] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, ShrimaiPrabhumoye, Yiming Yang, S. Welleck, BodhisattwaPrasadMajumder, Shashank Gupta, A. Yazdanbakhsh, 和 Peter Clark. Self-refine: 基于自反馈的迭代优化. 神经信息处理系统, 2023.

[736] 辛吉迈, 肖浩天, Xing W., 王文农, 张颖颖, 张文强. Agent RL缩放定律：用于数学问题解决的具有自发代码执行的Agent RL, arXiv预印本arXiv:2505.07773, 2025. URLhttps://arxiv.org/abs/2505.07773v2.

阿姆贾德·尤瑟夫·马吉德、塞尔日·萨比、托马斯·范·里特伯格伦、文森特·弗朗索瓦-拉韦、R·V·普拉萨德和克里斯·弗罗因霍文。深度强化学习与进化策略：一项比较性综述。*IEEE Transactions onNeural Networks and LearningSystems*, 2021.

[738] D. Maldonado, EdisonCruz, Jackeline AbadTorres, P. Cruz, and Silvana delPilar GamboaBenitez. 多智能体系统：关于其组件、框架和工作流程的调查。*IEEE Access*, 2024。

[739] Zhao Mandi, Shreeya Jain, and Shuran Song. Roco: 基于大型语言模型的辩证多机器人协作。 *IEEE International Conference on Robotics and Automation*, 2023。

[740] Jeremy R. Manning, Sean M. Polyn, G. Baltuch, B. Litt, and M. Kahana. 海马回路的振荡模式揭示记忆搜索过程中的情境重建。 *Proceedings of the NationalAcademyof Sciences of the United States of America*, 2011。

[741] Qiheng Mao, Zemin Liu, Chenghao Liu, Zhuo Li, and Jianling Sun. Advancing graph representation learning with large language models: A comprehensive survey of techniques, arXiv preprint arXiv:2402.05952v1, 2024. URL https://arxiv.org/abs/2402.05952v1.

[742] Amin Hosseiny Marani, Ulie Schnaithmann, Youngseo Son, Akil Iyer, Manas Paldhe, and Arushi Raghuvanshi. Graph integrated language transformers for next action prediction in complex phone calls. *North American Chapter of the Association for Computational Linguistics*, 2024.

[743] Sophia Maria. Compass-v2 technical report, arXiv preprint arXiv:2504.15527, 2025. URL https://arxiv.org/abs/2504.15527v1.

[744] Viorica Marian and U. Neisser. Language-dependent recall of autobiographical memories. *Journal of experimental psychology. General*, 2000.

[745] S. Mariani and Andrea Omicini. Special issue "multi-agent systems": Editorial. *Applied Sciences*, 2019.

[746] Vasilije Markovic, Lazar Obradović, L'aszl'o Hajdu, and Jovan Pavlović. Optimizing the interface between knowledge graphs and llms for complex reasoning, arXiv preprint arXiv:2505.24478, 2025. URL https://arxiv.org/abs/2505.24478v1.

[747] Sami Marreed, Alon Oved, Avi Yaeli, Segev Shlomov, Ido Levy, Offer Akrabi, Aviad Sela, Asaf Adi, and Nir Mashkif. Towards enterprise-ready computer using generalist agent. 2025.

[748] Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey, arXiv preprint arXiv:2404.11584, 2024. URL https://arxiv.org/abs/2404.11584v1.

[749] Nicholas Matsumoto, Jay Moran, Hyunjun Choi, Miguel E. Hernandez, Mythreye Venkatesan, Paul Wang, and Jason H. Moore. Kragen: a knowledge graph-enhanced rag framework for biomedical problem solving using large language models. *Bioinformatics*, 2024.

[750] Ryoga Matsuo, Stefan Uhlich, Arun Venkitaraman, Andrea Bonetti, Chia-Yu Hsieh, Ali Momeni, Lukas Mauch, Augusto Capone, Eisaku Ohbuchi, and Lorenzo Servadei. Schemato - an llm for netlist-to-schematic conversion. arXiv preprint, 2024.

[751] Costas Mavromatis, V. Ioannidis, Shen Wang, Da Zheng, Soji Adeshina, Jun Ma, Han Zhao, C. Faloutsos, and G. Karypis. Train your own gnn teacher: Graph-aware distillation on textual graphs. *ECML/PKDD*, 2023.

[752] James L. McClelland, B. McNaughton, and R. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychology Review*, 1995.

[753] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *Annual Meeting of the Association for Computational Linguistics*, 2019.

[754] Daniel McDuff, M. Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak N. Patel, D. Webster, Ewa Dominowska,

[741] 毛启恒, 刘泽民, 刘成浩, 李卓, 和 孙建玲. 基于大型语言模型的图表示学习进展：技术综述, arXiv 预印本 arXiv:2402.05952v1, 2024. URLhttps://arxiv.org/abs/2402.05952v1.

[742] Amin Hosseiny Marani, Ulie Schnaithmann, Youngseo Son, Akil Iyer, Manas Paldhe, 和 Arushi Raghuvanshi. 用于复杂电话中下一步动作预测的图集成语言转换器. 美国计算语言学协会北美分会, 2024.

[743] Sophia Maria. Compass-v2 技术报告，arXiv 预印本 arXiv:2504.15527, 2025. URL https://arxiv.org/abs/2504.15527v1.

[744] Viorica Marian 和 U. Neisser. 语言依赖的自传体记忆召回. 实验心理学杂志. 一般, 2000.

[745] S. Mariani 和 Andrea Omicini. 特刊“多智能体系统”：编者按. *Applied Sciences*, 2019.

[746] Vasilije Markovic, Lazar Obradović, L'aszl'o Hajdu, 和 Jovan Pavlović. 优化知识图谱和 llms 之间的接口以进行复杂推理，arXiv 预印本 arXiv:2505.24478, 2025. URL https://arxiv.org/abs/2505.24478v1.

[747] Sami Marreed, Alon Oved, Avi Yaeli, Segev Shlomov, Ido Levy, Offer Akrabi, Aviad Sela, Asaf Adi, 和 Nir Mashkif. 迈向企业级计算机使用通用智能体. 2025.

[748] Tula Masterman, Sandi Besen, Mason Sawtell, 和 Alex Chao. 用于推理、规划和工具调用的新兴人工智能智能体架构的格局：一项调查，arXiv 预印本 arXiv:2404.11584, 2024. URLhttps://arxiv.org/abs/2404.11584v1.

[749] Nicholas Matsumoto、Jay Moran、Hyunjun Choi、Miguel E. Hernandez、Mythreye Venkatesan、Paul Wang 和 Jason H. Moore。Kragen：一个用于生物医学问题解决的、基于知识图谱增强的检索增强生成框架，使用大型语言模型。*Bioinformatics*, 2024.

Ryoga Matsuo、Stefan Uhlich、Arun Venkitaraman、Andrea Bonetti、Chia-Yu Hsieh、Ali Momeni、Lukas Mauch、Augusto Capone、Eisaku Ohbuchi 和 Lorenzo Servadei。Schemato - 一种用于网表到原理图转换的大型语言模型。arXiv 预印本，2024年。

[751] Costas Mavromatis、V. Ioannidis、Shen Wang、Da Zheng、Soji Adeshina、Jun Ma、Han Zhao、C. Falout-sos 和 G. Karypis。训练你自己的 GNN 教师：文本图上的图感知蒸馏。*ECML/PKDD*, 2023.

[752] 詹姆斯·L·麦克莱兰、B·麦克诺顿和R·奥里利。海马体和新皮层中存在互补性学习系统的原因：从连接主义学习和记忆模型的成功与失败中获得的启示。心理学评论, 1995。

[753] R·托马斯·麦科伊、艾莉·帕维克和塔尔·林岑。理由错误：诊断自然语言推理中的句法启发式。计算语言学协会年会, 2019。

[754] 丹尼尔·麦克杜夫、M·沙克曼、陶图、阿尼尔·帕勒普、王爱美、杰克·加里森、卡兰·辛哈尔、亚什·夏尔马、谢库菲·阿齐兹、卡维塔·库尔卡尼、刘昊、杨成、刘云、S·马赫达维、苏什安特·普拉卡什、阿努帕姆·帕塔克、克里斯托弗·塞姆图尔斯、施韦特克·N·帕特尔、D·韦伯斯特、埃娃·多明诺斯卡,

Juraj Gottweis, Joelle Barral, Katherine Chou, G. Corrado, Yossi Matias, Jacob Sunshine, A. Karthikesalingam, and Vivek Natarajan. Towards accurate differential diagnosis with large language models. *Nature*, 2023.

[755] AD McNaughton, G Ramalaxmi, A Kruel, and CR Knutson. . . . Cactus: Chemistry agent connecting tool-usage to science, arxiv, 2024.

[756] Sushant Mehta, R. Dandekar, R. Dandekar, and S. Panat. Latent multi-head attention for small language models, arXiv preprint arXiv:2506.09342, 2025. URL https://arxiv.org/abs/2506.09342v2.

[757] Kai Mei, Zelong Li, Shuyuan Xu, Ruosong Ye, Yingqiang Ge, and Yongfeng Zhang. Aios: Llm agent operating system, arXiv preprint arXiv:2403.16971, 2024. URL https://arxiv.org/abs/2403.16971v4.

[758] Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, and Xueqi Cheng. Slang: New concept comprehension of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 12558–12575. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.698. URL http://dx.doi.org/10.18653/v1/2024.emnlp-main.698.

[759] Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Ruibin Yuan, and Xueqi Cheng. Hiddenguard: Fine-grained safe generation with specialized representation router, arXiv preprint arXiv:2410.02684, 2024. URL https://arxiv.org/abs/2410.02684.

[760] Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Yuyao Ge, Jun Wan, Yurong Wu, and Xueqi Cheng. a1: Steep test-time scaling law via environment augmented generation, arXiv preprint arXiv:2504.14597, 2025. URL https://arxiv.org/abs/2504.14597.

[761] Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Jiayi Mao, and Xueqi Cheng. "not aligned" is not "malicious": Being careful about hallucinations of large language models' jailbreak, arXiv preprint arXiv:2406.11668, 2025. URL https://arxiv.org/abs/2406.11668.

[762] T. Meiser and A. Bröder. Memory for multidimensional source information. *Journal of Experimental Psychology. Learning, Memory and Cognition*, 2002.

[763] Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. *International Conference on Learning Representations*, 2017.

[764] Kevin Meng, David Bau, A. Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Neural Information Processing Systems*, 2022.

[765] Yuxian Meng, Shi Zong, Xiaoya Li, Xiaofei Sun, Tianwei Zhang, Fei Wu, and Jiwei Li. Gnn-lm: Language modeling based on global contexts via gnn. *International Conference on Learning Representations*, 2021.

[766] Agnieszka Mensfelt, Kostas Stathis, and Vince Trencsenyi. Towards logically sound natural language reasoning with logic-enhanced language model agents, arXiv preprint arXiv:2408.16081, 2024. URL https://arxiv.org/abs/2408.16081v2.

[767] G. M. Mensink and J. Raaijmakers. A model for interference and forgetting. arXiv preprint, 1988.

Juraj Gottweis, Joelle Barral, Katherine Chou, G. Corrado, Yossi Matias, Jacob Sunshine, A. Karthike-salingam, and Vivek Natarajan. 大型语言模型在准确鉴别诊断方面的进展。*Nature*, 2023。

[755] AD McNaughton, G Ramalaxmi, A Kruel, and CR Knutson. . . . Cactus: 连接工具使用与科学的化学代理, arxiv, 2024。

[756] Sushant Mehta, R. Dandekar, R. Dandekar, and S. Panat. 小型语言模型的潜在多头注意力机制，arXiv preprintarXiv:2506.09342, 2025. URL https://arxiv.org/abs/2506.09342v2.

[757] Kai Mei, Zelong Li, Shuyuan Xu, Ruosong Ye, Yingqiang Ge, and Yongfeng Zhang. Aios: Llm代理操作系统，arXiv preprint arXiv:2403.16971, 2024. URL https://arxiv.org/abs/2403.16971v4.

[758] Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, and Xueqi Cheng. Slang: 大型语言模型的新概念理解. In *2024* 年自然语言处理经验方法会议论文集, 第 12558–12575 页. 计算语言学协会, 2024. doi: 10.18653/v1/2024.emnlp-main.698.
URLhttp://dx.doi.org/10.18653/v1/2024.emnlp-main.698.

[759] Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Ruibin Yuan, and Xueqi Cheng. Hiddenguard: 基于专用表示路由器的细粒度安全生成, arXiv 预印本 arXiv:2410.02684, 2024.
URLhttps://arxiv.org/abs/2410.02684.

[760] Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Yuyao Ge, Jun Wan, Yurong Wu, and Xueqi Cheng. a1: 通过环境增强生成实现陡峭的测试时扩展规律, arXiv 预印本 arXiv:2504.14597, 2025.
URLhttps://arxiv.org/abs/2504.14597.

[761] 梅凌瑞, 刘胜华, 王依伟, 毕宝龙, 毛佳怡, 和程雪琪. "不匹配"不是"恶意": 仔细对待大语言模型的越狱幻觉, arXiv 预印本 arXiv:2406.11668, 2025. URLhttps://arxiv.org/abs/2406.11668.

[762] T. Meiser 和 A. Bröder. 多维源信息的记忆. 实验心理学杂志. 学习、记忆与认知, 2002.

[763] Gábor Melis, Chris Dyer, 和 Phil Blunsom. 神经语言模型评估的现状. 学习表示国际会议, 2017.

[764] Kevin Meng, David Bau, A. Andonian, 和 Yonatan Belinkov. 定位和编辑 GPT 中的事实关联. 神经信息处理系统, 2022.

[765] 孟宇娴, 宗石, 李晓雅, 孙晓飞, 张天伟, 吴飞, 和 李继伟. Gnn-lm: 基于全局上下文的基于gnn的语言模型. 国际学习表征会议, 2021.

[766] Agnieszka Mensfelt, Kostas Stathis, 和 Vince Trencsenyi. 迈向逻辑上可靠的天然语言推理: 基于逻辑增强语言模型代理的推理, arXiv预印本 arXiv:2408.16081, 2024. URLhttps://arxiv.org/abs/2408.16081v2.

[767] G. M. Mensink 和 J. Raaijmakers. 干扰和遗忘的模型. arXiv预印本, 1988.

[768] Thomas Merth, Qichen Fu, Mohammad Rastegari, and Mahyar Najibi. Superposition prompting: Improving and accelerating retrieval-augmented generation. *International Conference on Machine Learning*, 2024.

[769] B. Meskó. Prompt engineering as an important emerging skill for medical professionals: Tutorial. *Journal of Medical Internet Research*, 2023.

[770] Yapeng Mi, Zhi Gao, Xiaojian Ma, and Qing Li. Building llm agents by incorporating insights from computer systems, arXiv preprint arXiv:2504.04485, 2025. URL https://arxiv.org/abs/2504.04485v1.

[771] G. Mialon, Roberto Dessì, M. Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, R. Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. Augmented language models: a survey. *Trans. Mach. Learn. Res.*, 2023.

[772] G. Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants, arXiv preprint arXiv:2311.12983, 2023. URL https://arxiv.org/abs/2311.12983v1.

[773] Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Hongyi Jin, Tianqi Chen, and Zhihao Jia. Towards efficient generative large language model serving: A survey from algorithms to systems, arXiv preprint arXiv:2312.15234, 2023. URL https://arxiv.org/abs/2312.15234v1.

[774] Jacob Miller, Guillaume Rabusseau, and John Terilla. Tensor networks for language modeling. arXiv preprint, 2020.

[775] Xing ming Guo, Darioush Keivan, U. Syed, Lianhui Qin, Huan Zhang, G. Dullerud, Peter J. Seiler, and Bin Hu. Controlagent: Automating control system design via novel integration of llm agents and domain expertise, arXiv preprint arXiv:2410.19811, 2024. URL https://arxiv.org/abs/2410.19811v1.

[776] Soroush Mirjalili, Patrick S. Powell, Jonathan Strunk, Taylor A James, and Audrey Duarte. Context memory encoding and retrieval temporal dynamics are modulated by attention across the adult lifespan. *eNeuro*, 2021.

[777] Ishan Misra and L. Maaten. Self-supervised learning of pretext-invariant representations. *Computer Vision and Pattern Recognition*, 2019.

[778] Ali Modarressi, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schütze. Ret-llm: Towards a general read-write memory for large language models, arXiv preprint arXiv:2305.14322, 2024. URL https://arxiv.org/abs/2305.14322.

[779] Ali Modarressi, Abdullatif Köksal, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schutze. Memllm: Finetuning llms to use an explicit read-write memory. *Trans. Mach. Learn. Res.*, 2024.

[780] Behnam Mohammadi. Pel, a programming language for orchestrating ai agents, arXiv preprint arXiv:2505.13453, 2025. URL https://arxiv.org/abs/2505.13453v2.

[781] Amirkeivan Mohtashami and Martin Jaggi. Landmark attention: Random-access infinite context length for transformers. *Neural Information Processing Systems*, 2023.

[768] 托马斯·梅尔特、齐晨·傅、穆罕默德·拉斯特加里和马哈亚尔·纳吉比。叠加提示：改进和加速检索增强生成。国际机器学习会议, 2024。

[769] B. 梅斯科。提示工程作为医疗专业人员的一项重要新兴技能：教程。医疗互联网研究杂志, 2023。

[770] 叶鹏、高志、马晓健和黎青。通过结合计算机系统中的见解构建 llm 代理，arXiv 预印本 arXiv:2504.04485，2025。URLhttps://arxiv.org/abs/2504.04485v1。

[771] G. 米亚隆、罗伯托·德西、M. 罗梅利、克里斯托弗罗斯·纳尔马潘蒂斯、拉马卡南特·帕苏努鲁、R. 拉莱努、巴蒂斯特·罗齐埃、蒂莫·希克、简·德维维迪-尤、阿斯拉·塞利基尔马兹、爱德华·格拉夫、扬·勒库恩和托马斯·斯卡洛姆。增强语言模型：一项调查。机器学习研究传输, 2023。

[772] G. Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, 和 Thomas Scialom. Gaia: 一个通用人工智能助手的基准测试，arXiv 预印本 arXiv:2311.12983, 2023. URL https://arxiv.org/abs/2311.12983v1.

[773] Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Hongyi Jin, Tianqi Chen, and Zhihao Jia. 迈向高效生成式大型语言模型服务：从算法到系统的综述，arXiv 预印本 arXiv:2312.15234，2023。URLhttps://arxiv.org/abs/2312.15234v1.

[774] Jacob Miller、Guillaume Rabusseau 和 John Terilla. 用于语言模型的张量网络。arXiv 预印本，2020.

[775] 邢名国, Darioush Keivan, U. Syed, 秦连辉, 张欢, G. Dullerud, Peter J. Seiler, 和 胡斌. Controlagent: 通过 llm 代理和领域专业知识的创新集成自动化控制系统设计, arXiv 预印本 arXiv:2410.19811, 2024. URL https://arxiv.org/abs/2410.19811v1.

[776] Soroush Mirjalili, Patrick S. Powell, Jonathan Strunk, Taylor A James, 和 Audrey Duarte. 上下文记忆编码和检索的时序动态受成年期注意力的调节. *eNeuro*, 2021.

[777] Ishan Misra 和 L. Maaten. 前置不变表示的自监督学习. 计算机视觉和模式识别, 2019.

[778] Ali Modarressi, Ayyoob Imani, Mohsen Fayyaz, 和 Hinrich Schütze. Ret-llm: 面向大型语言模型的通用读写内存, arXiv 预印本 arXiv:2305.14322, 2024. URL https://arxiv.org/abs/2305.14322.

[779] Ali Modarressi, Abdullatif Köksal, Ayyoob Imani, Mohsen Fayyaz, 和 Hinrich Schutze. Memllm: 微调大型语言模型以使用显式读写内存。机器学习研究杂志, 2024.

[780] Behnam Mohammadi. Pel, 一种用于编排人工智能代理的编程语言, arXiv 预印本 arXiv:2505.13453, 2025. URLhttps://arxiv.org/abs/2505.13453v2.

[781] Amirkeivan Mohtashami 和 Martin Jaggi. Landmark attention: 随机访问无限上下文长度的 Transformer。*Neural InformationProcessingSystems*, 2023.

[782] Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. Skill: Structured knowledge infusion for large language models. *North American Chapter of the Association for Computational Linguistics*, 2022.

[783] Dimitri Coelho Mollo and Raphael Milliere. The vector grounding problem, arXiv preprint arXiv:2304.01481, 2023. URL https://arxiv.org/abs/2304.01481v2.

[784] Nieves Montes, N. Osman, and C. Sierra. Combining theory of mind and abduction for cooperation under imperfect information. *European Workshop on Multi-Agent Systems*, 2022.

[785] Suhong Moon, Siddharth Jha, Lutfi Eren Erdogan, Sehoon Kim, Woosang Lim, Kurt Keutzer, and A. Gholami. Efficient and scalable estimation of tool representations in vector space, arXiv preprint arXiv:2409.02141, 2024. URL https://arxiv.org/abs/2409.02141v1.

[786] Shinsuke Mori. A stochastic parser based on an slm with arboreal context trees. *International Conference on Computational Linguistics*, 2002.

[787] Meredith Ringel Morris. Prompting considered harmful. *Communications of the ACM*, 2024.

[788] Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, D. Klakow, and Yanai Elazar. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *Annual Meeting of the Association for Computational Linguistics*, 2023.

[789] Sajad Mousavi, Ricardo Luna Guti'errez, Desik Rengarajan, Vineet Gundecha, Ashwin Ramesh Babu, Avisek Naug, Antonio Guillen-Perez, and S. Sarkar. N-critics: Self-refinement of large language models with ensemble of critics. arXiv preprint, 2023.

[790] Manisha Mukherjee, Sungchul Kim, Xiang Chen, Dan Luo, Tong Yu, and Tung Mai. From documents to dialogue: Building kg-rag enhanced ai assistants, arXiv preprint arXiv:2502.15237, 2025. URL https://arxiv.org/abs/2502.15237v1.

[791] Tergel Munkhbat, Namgyu Ho, Seohyun Kim, Yongjin Yang, Yujin Kim, and Se young Yun. Self-training elicits concise reasoning in large language models, arXiv preprint arXiv:2502.20122, 2025. URL https://arxiv.org/abs/2502.20122v3.

[792] Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with infini-attention, arXiv preprint arXiv:2404.07143, 2024. URL https://arxiv.org/abs/2404.07143v2.

[793] Eliya Nachmani, Alon Levkovitch, Julián Salazar, Chulayutsh Asawaroengchai, Soroosh Mariooryad, R. Skerry-Ryan, and Michelle Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered llm. *International Conference on Learning Representations*, 2023.

[794] L. Nadel, Jessica D. Payne, and W. J. Jacobs. The relationship between episodic memory and context: clues from memory errors made while under stress. *Physiological Research*, 2002.

[795] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, arXiv preprint arXiv:2112.09332, 2022. URL https://arxiv.org/abs/2112.09332.

[782] Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. Skill: 大型语言模型的结构化知识注入. 北美学术计算语言学分会, 2022.

[783] Dimitri Coelho Mollo and Raphael Milliere. 向量接地问题, arXiv 预印本 arXiv:2304.01481, 2023. URLhttps://arxiv.org/abs/2304.01481v2.

[784] Nieves Montes, N. Osman, and C. Sierra. 结合心智理论和溯因推理以在信息不完整的情况下进行合作. 欧洲多智能体系统研讨会, 2022.

[785] Suhong Moon, Siddharth Jha, Lutfi Eren Erdogan, Sehoon Kim, Woosang Lim, Kurt Keutzer, and A. Gholami. 向量空间中工具表示的有效和可扩展估计, arXiv 预印本 arXiv:2409.02141, 2024. URLhttps://arxiv.org/abs/2409.02141v1.

[786] 森伸介. 基于具有树形上下文树的slm的随机解析器. 国际计算语言学会议,2002.

[787] Meredith Ringel Morris. 提示有害. *ACM* 通讯, 2024.

[788] Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, D. Klakow, 和 Yanai Elazar. 少样本微调与情境学习: 公平的比较和评估. 计算语言学协会年度会议, 2023.

[789] Sajad Mousavi, Ricardo Luna Guti'errez, Desik Rengarajan, Vineet Gundecha, Ashwin Ramesh Babu, Avisek Naug, Antonio Guillen-Perez, 和 S. Sarkar. N-critics: 批评家集合的自我改进大型语言模型. arXiv 预印本, 2023.

[790] Manisha Mukherjee, Sungchul Kim, Xiang Chen, Dan Luo, Tong Yu, 和 Tung Mai. 从文档到对话: 构建kg-rag增强型ai助手, arXiv 预印本 arXiv:2502.15237, 2025. URLhttps://arxiv.org/abs/2502.15237v1.

[791] Tergel Munkhbat、Namgyu Ho、Seohyun Kim、Yongjin Yang、Yujin Kim 和 Se young Yun。Self-training elicits concise reasoning in large language models，arXiv preprint arXiv:2502.20122，2025。URLhttps://arxiv.org/abs/2502.20122v3。

[792] Tsendsuren Munkhdalai、Manaal Faruqui和Siddharth Gopal。不留任何上下文：具有无限注意力机制的无限上下文高效转换器，arXiv预印本arXiv:2404.07143，2024年。URLhttps://arxiv.org/abs/2404.07143v2。

[793] Eliya Nachmani、Alon Levkovitch、Julián Salazar、Chulayutsh Asawaroengchai、Soroosh Mariooryad、R. Skerry-Ryan 和 Michelle Tadmor Ramanovich。基于频谱的 LLM 的语音问答和语音延续。国际学习表征会议，2023。

[794] L. Nadel, Jessica D. Payne, 和 W. J. Jacobs. 病例记忆与上下文的关系：压力下记忆错误提供的线索。 生理学研究, 2002。

[795] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, 和 John Schulman. Webgpt：带人类反馈的浏览器辅助问答，arXiv 预印本 arXiv:2112.09332，2022。URL https://arxiv.org/abs/2112.09332.

[796] Koichi Namekata, Amirmojtaba Sabour, Sanja Fidler, and Seung Wook Kim. Emerdiff: Emerging pixel-level semantic knowledge in diffusion models, arXiv preprint arXiv:2401.11739, 2024. URL https://arxiv.org/abs/2401.11739.

[797] Sundaraparipurnan Narayanan and Sandeep Vishwakarma. Guard-d-llm: An llm-based risk assessment engine for the downstream uses of llms. arXiv preprint, 2024.

[798] Usman Naseem, Surendrabikram Thapa, Qi Zhang, Liang Hu, Anum Masood, and Mehwish Nasim. Reducing knowledge noise for improved semantic analysis in biomedical natural language processing applications. *Clinical Natural Language Processing Workshop*, 2023.

[799] Deepak Nathani, David Wang, Liangming Pan, and W. Wang. Maf: Multi-aspect feedback for improving reasoning in large language models. *Conference on Empirical Methods in Natural Language Processing*, 2023.

[800] Aashutosh Nema, Samaksh Gulati, Evangelos Giakoumakis, and Bipana Thapaliya. Modp: Multi objective directional prompting, arXiv preprint arXiv:2504.18722, 2025. URL https://arxiv.org/abs/2504.18722v1.

[801] Christian D. Newman, Anthony S Peruma, and Reem S. Alsuhaibani. Modeling the relationship between identifier name and behavior. *IEEE International Conference on Software Maintenance and Evolution*, 2019.

[802] M. Nieznański, Michał Obidziński, Emilia Zyskowska, and Daria Niedziałkowska. Executive resources and item-context binding: Exploring the influence of concurrent inhibition, updating, and shifting tasks on context memory. *Advances in Cognitive Psychology*, 2015.

[803] M. Nieznański, Michał Obidziński, and Daria Ford. Does context recollection depend on the base-rate of contextual features? *Cognitive Processing*, 2023.

[804] C. Nourani and P. Eklund. Concepts ontology algebras and role descriptions. *Conference on Computer Science and Information Systems*, 2017.

[805] Felix Ocker, Daniel Tanneberg, Julian Eggert, and Michael Gienger. Tulip agent - enabling llm-based agents to solve tasks using large tool libraries. arXiv preprint, 2024.

[806] Felix Ocker, J. Deigmöller, Pavel Smirnov, and Julian Eggert. A grounded memory system for smart personal assistants, arXiv preprint arXiv:2505.06328, 2025. URL https://arxiv.org/abs/2505.06328v1.

[807] OpenAI. Computer-using agent, 2025. URL https://openai.com/index/computer-using-agent/. OpenAI Technical Report.

[808] OpenAI. Swarm: Educational framework exploring ergonomic, lightweight multi-agent orchestration. https://github.com/openai/swarm, 2025. [Online; accessed 17-July-2025].

[809] Jonas Oppenlaender. Dangermaps: Personalized safety advice for travel in urban environments using a retrieval-augmented language model, arXiv preprint arXiv:2503.14103, 2025. URL https://arxiv.org/abs/2503.14103v3.

[810] A. Orhan. Recognition, recall, and retention of few-shot memories in large language models, arXiv preprint arXiv:2303.17557, 2023. URL https://arxiv.org/abs/2303.17557v1.

[796] Koichi Namekata、Amirmojtaba Sabour、Sanja Fidler 和 Seung Wook Kim。Emerdiff：在扩散模型中涌现的像素级语义知识，arXiv 预印本 arXiv:2401.11739，2024年。URLhttps://arxiv.org/abs/2401.11739。

Sundaraparipurnan Narayanan 和 Sandeep Vishwakarma. Guard-d-llm: 基于LLM的下游应用风险评估引擎. arXiv预印本, 2024.

Usman Naseem、Surendrabikram Thapa、Qi Zhang、Liang Hu、Anum Masood和Mehwish Nasim。减少知识噪声以改进生物医学自然语言处理应用中的语义分析。临床自然语言处理研讨会, 2023。

[799] Deepak Nathani、David Wang、Liangming Pan和W. Wang。Maf：用于改进大型语言模型推理的多方面反馈。自然语言处理经验方法会议, 2023。

[800] Aashutosh Nema, Samaksh Gulati, Evangelos Giakoumakis, and Bipana Thapaliya. Modp: Multi objective directional prompting, arXiv preprint arXiv:2504.18722, 2025. URL https://arxiv.org/abs/2504.18722v1.

[801] Christian D. Newman, Anthony S Peruma, and Reem S. Alsuhaibani. 建模标识符名称与行为之间的关系。*IEEE International Conference on Software MaintenanceandEvolution*, 2019.

[802] M. Nieznański, Michał Obidziński, Emilia Zyskowska, 和 Daria Niedziałkowska. 执行资源和项目-上下文绑定：探索并发抑制、更新和转换任务对上下文记忆的影响。认知心理学进展, 2015.

[803] M. Nieznański, Michał Obidziński, and DariaFord. 上下文回忆是否依赖于上下文特征的基本率? *CognitiveProcessing*, 2023.

[804] C. Nouraniand P. Eklund. 概念本体代数和角色描述. 计算机科学与信息系统会议, 2017.

[805] Felix Ocker, Daniel Tanneberg, Julian Eggert, and Michael Gienger. Tulip代理 - 使基于llm的代理能够使用大型工具库完成任务. arXiv预印本, 2024.

[806] Felix Ocker, J. Deigmöller, Pavel Smirnov, and Julian Eggert. 智能个人助理的基于记忆的系统, arXiv预印本 arXiv:2505.06328, 2025. URLhttps://arxiv.org/abs/2505.06328v1.

[807]OpenAI. 使用计算机的代理, 2025. URL https://openai.com/index/computer-using-agent/. OpenAI技术报告.

[808] OpenAI. Swarm: 探索人体工程学、轻量级多代理编排的教育框架. https://github.com/openai/swarm,2025. [在线; 访问于 17-July-2025].

[809] Jonas Oppenlaender. Dangermaps: 城市环境中使用检索增强语言模型的个性化安全建议, arXiv preprint arXiv:2503.14103, 2025。URLhttps://arxiv.org/abs/2503.14103v3.

[810] A. Orhan. 大型语言模型中少样本记忆的识别、召回和保持，arXiv preprint arXiv:2303.17557，2023。URLhttps://arxiv.org/abs/2303.17557v1.

[811] Gustavo Ortiz-Hernández, Alejandro Guerra-Hernández, J. Hübner, and W. A. Luna-Ramírez. Modularization in belief-desire-intention agent programming and artifact-based environments. *PeerJ Computer Science*, 2022.

[812] Wendkûuni C. Ouédraogo, A. Kaboré, Haoye Tian, Yewei Song, Anil Koyuncu, Jacques Klein, David Lo, and Tegawend'e F. Bissyand'e. Large-scale, independent and comprehensive study of the power of llms for test case generation. arXiv preprint, 2024.

[813] Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph Gonzalez. Memgpt: Towards llms as operating systems, arXiv preprint arXiv:2310.08560, 2023. URL https://arxiv.org/abs/2310.08560v2.

[814] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems, arXiv preprint arXiv:2310.08560, 2024. URL https://arxiv.org/abs/2310.08560.

[815] Constantin-Valentin Pal, F. Leon, M. Paprzycki, and M. Ganzha. A review of platforms for the development of agent systems. *Inf.*, 2020.

[816] Qianjun Pan, Wenkai Ji, Yuyang Ding, Junsong Li, Shilian Chen, Junyi Wang, Jie Zhou, Qin Chen, Min Zhang, Yulan Wu, and Liang He. A survey of slow thinking-based reasoning llms using reinforced learning and inference-time scaling law, arXiv preprint arXiv:2505.02665, 2025. URL https://arxiv.org/abs/2505.02665v2.

[817] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[818] Xu Pan, Ely Hahami, Zechen Zhang, and H. Sompolinsky. Memorization and knowledge injection in gated llms, arXiv preprint arXiv:2504.21239, 2025. URL https://arxiv.org/abs/2504.21239v1.

[819] Bo Pang, Hanze Dong, Jiacheng Xu, Silvio Savarese, Yingbo Zhou, and Caiming Xiong. Bolt: Bootstrap long chain-of-thought in language models without distillation, arXiv preprint arXiv:2502.03860, 2025. URL https://arxiv.org/abs/2502.03860v1.

[820] Jianhui Pang, Fanghua Ye, Derek F. Wong, and Longyue Wang. Anchor-based large language models. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[821] Bhargavi Paranjape, Scott M. Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. Art: Automatic multi-step reasoning and tool-use for large language models, arXiv preprint arXiv:2303.09014, 2023. URL https://arxiv.org/abs/2303.09014v1.

[822] A Parisi, Y Zhao, and N Fiedel. Talm: Tool augmented language models. 2022. URL https://arxiv.org/abs/2205.12255.

[823] Dongju Park and Chang Wook Ahn. Self-supervised contextual data augmentation for natural language processing. *Symmetry*, 2019.

[824] J. Park, Lindsay Popowski, Carrie J. Cai, M. Morris, Percy Liang, and Michael S. Bernstein. Social simulacra: Creating populated prototypes for social computing systems. *ACM Symposium on User Interface Software and Technology*, 2022.

[811] Gustavo Ortiz-Hernández, Alejandro Guerra-Hernández, J. Hübner, 和 W. A. Luna-Ramírez. 在信念-愿望-意图代理编程和基于物品的环境中实现模块化。 *PeerJComputer Science*, 2022。

[812] Wendkûuni C. Ouédraogo, A. Kaboré, Haoye Tian, Yewei Song, Anil Koyuncu, Jacques Klein, David Lo, 和 Tegawend'e F. Bissyand'e. 对用于测试用例生成的 llms 功力的大规模、独立和综合研究。 arXiv preprint, 2024。

[813] Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, 和 Joseph Gonzalez. Memgpt: 迈向将 llms 作为操作系统，arXiv preprint arXiv:2310.08560, 2023。 URL https://arxiv.org/abs/2310.08560v2.

[814] 查尔斯·帕克，莎拉·伍德尔斯，凯文·林，菲比安·方，希希尔·G·帕蒂尔，伊昂·斯托伊卡和约瑟夫·E·冈萨雷斯。Memgpt：迈向将大型语言模型作为操作系统，arXiv预印本arXiv:2310.08560，2024年。URLhttps://arxiv.org/abs/2310.08560。

康斯坦丁-瓦伦丁·帕尔、F·莱昂、M·帕普里茨基和M·甘查. 代理系统开发平台的综述. *Inf.*, 2020.

[816] 潘千军, 李文凯, 丁宇阳, 李俊松, 陈世廉, 王俊毅, 周杰, 陈秦, 张敏, 吴雨兰, 何亮. 基于慢思考的推理大语言模型综述：强化学习和推理时扩展律的应用, arXiv 预印本 arXiv:2505.02665, 2025. URL https://arxiv.org/abs/2505.02665v2.

[817] 潘世瑞, 罗林浩, 王宇飞, 陈晨, 王嘉普, 和吴新东. 统一大语言模型和知识图谱: 一份路线图. *IEEE Transactions on Knowledge and DataEngineering*, 2023.

[818] 潘旭, 哈哈米·埃利, 张哲辰, 和 H. Sompolinsky. 门控 LLM 中的记忆和知识注入, arXiv preprint arXiv:2504.21239, 2025. URL https://arxiv.org/abs/2504.21239v1.

[819] 庞博, 董汉泽, 许家成, Savarese Silvio, 周英博, 和熊才明. Bolt: 无蒸馏的语言模型中引导长思维链, arXiv preprint arXiv:2502.03860, 2025. URLhttps://arxiv.org/abs/2502.03860v1.

[820] Jianhui Pang, Fanghua Ye, Derek F. Wong, and Longyue Wang. 基于锚点的大语言模型。计算语言学协会年会, 2024。

[821] Bhargavi Paranjape, Scott M. Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. 艺术：面向大语言模型的自动多步推理和工具使用，arXiv 预印本 arXiv:2303.09014，2023。URLhttps://arxiv.org/abs/2303.09014v1.

[822] A Parisi, Y Zhao, and N Fiedel. Talm：工具增强语言模型。2022。URL https://arxiv.org/abs/2205.12255.

[823] Dongju Park and Chang Wook Ahn. 自然语言处理的自监督上下文数据增强。对称性,2019。

[824] J. Park, Lindsay Popowski, Carrie J. Cai, M. Morris, Percy Liang, and Michael S. Bernstein. 社会拟像：为社交计算系统创建有人原型。 *ACM* 用户界面软件与技术研讨会, 2022。

[825] J. Park, Joseph C. O'Brien, Carrie J. Cai, M. Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *ACM Symposium on User Interface Software and Technology*, 2023.

[826] Soya Park, J. Zamfirescu-Pereira, and Chinmay Kulkarni. Model behavior specification by leveraging llm self-playing and self-improving, arXiv preprint arXiv:2503.03967, 2025. URL https://arxiv.org/abs/2503.03967v1.

[827] Rajvardhan Patil and Venkat Gudivada. A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences*, 2024.

[828] Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive apis, arXiv preprint arXiv:2305.15334, 2023. URL https://arxiv.org/abs/2305.15334.

[829] Shishir G. Patil, Huanzhi Mao, Charlie Cheng-Jie Ji, Fanjia Yan, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*, 2025.

[830] Shuva Paul, Farhad Alemi, and Richard Macwan. Llm-assisted proactive threat intelligence for automated reasoning, arXiv preprint arXiv:2504.00428, 2025. URL https://arxiv.org/abs/2504.00428v1.

[831] Saurav Pawar, S. Tonmoy, S. M. M. Zaman, Vinija Jain, Aman Chadha, and Amitava Das. The what, why, and how of context length extension techniques in large language models - a detailed survey. arXiv preprint, 2024.

[832] Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph retrieval-augmented generation: A survey. *ArXiv*, abs/2408.08921, 2024. URL https://api.semanticscholar.org/CorpusID:271903170.

[833] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *International Conference on Learning Representations*, 2023.

[834] Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Learning from context or names? an empirical study on neural relation extraction. *Conference on Empirical Methods in Natural Language Processing*, 2020.

[835] Ji-Lun Peng, Sijia Cheng, Egil Diau, Yung-Yu Shih, Po-Heng Chen, Yen-Ting Lin, and Yun-Nung Chen. A survey of useful llm evaluation, arXiv preprint arXiv:2406.00936, 2024. URL https://arxiv.org/abs/2406.00936v1.

[836] Bryan Perozzi, Bahare Fatemi, Dustin Zelle, Anton Tsitsulin, Mehran Kazemi, Rami Al-Rfou, and Jonathan J. Halcrow. Let your graph do the talking: Encoding structured data for llms, arXiv preprint arXiv:2402.05862, 2024. URL https://arxiv.org/abs/2402.05862v1.

[837] E. Pesce and G. Montana. Improving coordination in small-scale multi-agent deep reinforcement learning through memory-driven communication. *Machine-mediated learning*, 2019.

[825] J. Park, Joseph C. O' Brien, Carrie J. Cai, M. Morris, Percy Liang, andMichael S. Bernstein. 生成式代理：人类行为的交互式虚拟形象。 *ACM* 用户界面软件与技术研讨会, 2023。

[826] Soya Park, J. Zamfirescu-Pereira, and Chinmay Kulkarni. 通过利用 llm 自我博弈和自我改进来建模行为规范，arXiv 预印本 arXiv:2503.03967, 2025。URL https://arxiv.org/abs/2503.03967v1.

[827] Rajvardhan Patil and Venkat Gudivada. 大型语言模型（llms）当前趋势、技术和挑战的综述。 应用科学, 2024。

[828] Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla：与海量 API 连接的大型语言模型，arXiv 预印本 arXiv:2305.15334, 2023。URL https://arxiv.org/abs/2305.15334.

[829] Shishir G. Patil, Huanzhi Mao, Charlie Cheng-Jie Ji, Fanjia Yan, Vishnu Suresh, Ion Stoica, 和 Joseph E. Gonzalez. The berkeley function calling leaderboard (bfcl): 从工具使用到大型语言模型的代理评估. 在 第*42*届国际机器学习会议, 2025.

[830] 舒瓦·保罗，法哈德·阿莱米和理查德·麦克万。Llm辅助的主动威胁情报用于自动化推理，arXiv预印本arXiv:2504.00428，2025年。URLhttps://arxiv.org/abs/2504.00428v1。

[831] Saurav Pawar、S. Tonmoy、S. M. M. Zaman、Vinija Jain、Aman Chadha 和 Amitava Das。大型语言模型中上下文长度扩展技术的“是什么、为什么和如何”——一项详细调查。arXiv 预印本，2024。

[832] 彭博, 朱云, 刘永超, 博晓和, 石海舟, 霍传涛, 张岩, 唐思亮. 图检索增强生成：一项调查. *arXiv*, abs/2408.08921, 2024. URLhttps://api.semanticscholar.org/CorpusID:271903170.

[833] 彭 Bowen, Jeffrey Quesnelle, 范洪路, 和 Enrico Shippole. Yarn：大型语言模型的高效上下文窗口扩展. 学习表示国际会议, 2023.

[834] 彭 Hao, 高天宇, 韩旭, 林岩凯, 李鹏, 刘志远, 孙毛松, 周杰. 从上下文名称中学习？ 神经关系抽取的实证研究. 自然语言处理经验方法会议, 2020.

[835] 彭 Ji-Lun, 程思嘉, Egil Diau, 施永宇, 陈伯恒, 林彦婷, 和 陈云农. 有用 llm 评估的调查, arXiv 预印本 arXiv:2406.00936, 2024. URL https://arxiv.org/abs/2406.00936v1.

[836] Bryan Perozzi, Bahare Fatemi, Dustin Zelle, Anton Tsitsulin, Mehran Kazemi, Rami Al-Rfou, and Jonathan J. Halcrow. 让你的图来说话：为大型语言模型编码结构化数据，arXiv 预印本 arXiv:2402.05862，2024。URLhttps://arxiv.org/abs/2402.05862v1.

[837] E. Pesce and G. Montana. 通过记忆驱动的通信改进小规模多智能体深度强化学习中的协调。机器中介学习, 2019。

[838] F. Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. How context affects language models' factual predictions. *Conference on Automated Knowledge Base Construction*, 2020.

[839] Yue Pi, Wang Zhang, Yong Zhang, Hairong Huang, Baoquan Rao, Yulong Ding, and Shuanghua Yang. Applications of multi-agent deep reinforcement learning communication in network management: A survey, arXiv preprint arXiv:2407.17030, 2024. URL https://arxiv.org/abs/2407.17030v1.

[840] Nancirose Piazza and Vahid Behzadan. A theory of mind approach as test-time mitigation against emergent adversarial communication. *Adaptive Agents and Multi-Agent Systems*, 2023.

[841] Mathis Pink, Vy A. Vo, Qinyuan Wu, Jianing Mu, Javier S. Turek, Uri Hasson, Kenneth A. Norman, Sebastian Michelmann, Alexander Huth, and Mariya Toneva. Assessing episodic memory in llms with sequence order recall tasks, arXiv preprint arXiv:2410.08133, 2024. URL https://arxiv.org/abs/2410.08133v1.

[842] Fahmida Liza Piya and Rahmatollah Beheshti. Advancing feature extraction in healthcare through the integration of knowledge graphs and large language models. *AAAI Conference on Artificial Intelligence*, 2025.

[843] A. Plaat, M. V. Duijn, N. V. Stein, Mike Preuss, P. V. D. Putten, and K. Batenburg. Agentic large language models, a survey, arXiv preprint arXiv:2503.23037, 2025. URL https://arxiv.org/abs/2503.23037v2.

[844] Moritz Plenz and Anette Frank. Graph language models. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[845] Sean M. Polyn, K. Norman, and M. Kahana. A context maintenance and retrieval model of organizational processes in free recall. *Psychology Review*, 2009.

[846] Liam Pond and Ichiro Fujinaga. Teaching llms music theory with in-context learning and chain-of-thought prompting: Pedagogical strategies for machines. *International Conference on Computer Supported Education*, 2025.

[847] V Porcu. The role of memory in llms: Persistent context for smarter conversations. *Int. J. Sci. Res. Manag.(IJSRM)*, 12:1673–1691, 2024.

[848] Ofir Press, Noah A. Smith, and M. Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *International Conference on Learning Representations*, 2021.

[849] Xavier Puig, K. Ra, Marko Boben, Jiaman Li, Tingwu Wang, S. Fidler, and A. Torralba. Virtualhome: Simulating household activities via programs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[850] Pranav Putta, Edmund Mills, Naman Garg, S. Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent q: Advanced reasoning and learning for autonomous ai agents, arXiv preprint arXiv:2408.07199, 2024. URL https://arxiv.org/abs/2408.07199v1.

[851] S. Qasim, Hassan Mahmood, and F. Shafait. Rethinking table recognition using graph neural networks. *IEEE International Conference on Document Analysis and Recognition*, 2019.

[838] F. Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 上下文如何影响语言模型的事实性预测。 自动化知识库构建会议, 2020。

[839] Yue Pi, Wang Zhang, Yong Zhang, Hairong Huang, Baoquan Rao, Yulong Ding, and Shuanghua Yang. 多智能体深度强化学习通信在网络管理中的应用：一项调查，arXiv 预印本 arXiv:2407.17030，2024。URL https://arxiv.org/abs/2407.17030v1。

[840] Nancirose Piazza and Vahid Behzadan. 一种心智理论方法作为对抗性通信的测试时缓解方法。 自适应智能体和多智能体系统, 2023。

[841] Mathis Pink, Vy A. Vo, Qinyuan Wu, Jianing Mu, Javier S. Turek, Uri Hasson, Kenneth A. Norman, Sebastian Michelmann, Alexander Huth, and Mariya Toneva. 使用序列顺序回忆任务评估大型语言模型的情景记忆，arXiv 预印本 arXiv:2410.08133，2024。URL https://arxiv.org/abs/2410.08133v1。

[842] Fahmida Liza Piya 和 Rahmatollah Beheshti. 通过知识图谱和大型语言模型的集成推进医疗保健中的特征提取。 *AAAI* 人工智能会议, 2025。

[843] A. Plaat, M. V. Duijn, N. V. Stein, Mike Preuss, P. V. D. Putten, 和 K. Batenburg. 代理式大型语言模型，一项调查，arXiv 预印本 arXiv:2503.23037, 2025。URL https://arxiv.org/abs/2503.23037v2.

[844] Moritz Plenz and Anette Frank. Graph language models. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[845] Sean M. Polyn, K. Norman, 和 M. Kahana. 自由回忆中组织过程的一种上下文维护和检索模型。 心理学评论, 2009。

[846] Liam Pond 和 Ichiro Fujinaga. 使用上下文学习和思维链提示教授 llms 音乐理论：机器的教学策略。 国际计算机支持教育会议, 2025。

[847] V Porcu. 大型语言模型中的记忆作用：持久上下文以实现更智能的对话。 *Int. J.Sci.Res.Manag.(IJSRM)*, 12:1673–1691, 2024。

[848] Ofir Press, Noah A. Smith, 和 M. Lewis. 短期训练，长期测试：带线性偏差的注意力机制实现输入长度外推。 *InternationalConferenceonLearningRepresentations*, 2021。

[849] Xavier Puig, K. Ra, Marko Boben, Jiaman Li, Tingwu Wang, S. Fidler, 和 A. Torralba. Virtualhome：通过程序模拟家庭活动。 *2018 IEEE/CVF Conference on Computer VisionandPattern Recognition*, 2018。

[850] Pranav Putta, Edmund Mills, Naman Garg, S. Motwani, Chelsea Finn, Divyansh Garg, 和 Rafael Rafailov. Agent q：为自主 AI 代理的高级推理和学习，arXiv 预印本 arXiv:2408.07199，2024。URLhttps://arxiv.org/abs/2408.07199v1.

[851] S. Qasim, Hassan Mahmood, andF. Shafait. 使用图神经网络重新思考表格识别。*IEEE International Conference on Document Analysis and Recognition*, 2019.

[852] Peng Qi, Haejun Lee, OghenetegiriTGSido, and Christopher D. Manning. Answering open-domain questions of varying reasoning steps from text. *Conference on Empirical Methods in Natural Language Processing*, 2020.

[853] Yong Qi, Gabriel Kyebambo, Siyuan Xie, Wei Shen, Shenghui Wang, Bitao Xie, Bin He, Zhipeng Wang, and Shuo Jiang. Safety control of service robots with llms and embodied knowledge graphs, arXiv preprint arXiv:2405.17846, 2024. URL https://arxiv.org/abs/2405.17846v1.

[854] Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Xinyue Yang, Jiadai Sun, Yu Yang, Shuntian Yao, Tianjie Zhang, Wei Xu, Jie Tang, and Yuxiao Dong. Webrl: Training llm web agents via self-evolving online curriculum reinforcement learning, arXiv preprint arXiv:2411.02337, 2024. URL https://arxiv.org/abs/2411.02337v3.

[855] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. Chatdev: Communicative agents for software development, arXiv preprint arXiv:2307.07924, 2024. URL https://arxiv.org/abs/2307.07924.

[856] Cheng Qian, Chi Han, Y. Fung, Yujia Qin, Zhiyuan Liu, and Heng Ji. Creator: Tool creation for disentangling abstract and concrete reasoning of large language models. *Conference on Empirical Methods in Natural Language Processing*, 2023.

[857] Cheng Qian, Jiahao Li, Yufan Dang, Wei Liu, Yifei Wang, Zihao Xie, Weize Chen, Cheng Yang, Yingli Zhang, Zhiyuan Liu, and Maosong Sun. Iterative experience refinement of software-developing agents, arXiv preprint arXiv:2405.04219, 2024. URL https://arxiv.org/abs/2405.04219v1.

[858] Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tur, Gokhan Tur, and Heng Ji. Toolrl: Reward is all tool learning needs, arXiv preprint arXiv:2504.13958, 2025. URL https://arxiv.org/abs/2504.13958v1.

[859] Hongjin Qian, Zheng Liu, Peitian Zhang, Zhicheng Dou, and Defu Lian. Boosting long-context management via query-guided activation refilling, arXiv preprint arXiv:2412.12486, 2024. URL https://arxiv.org/abs/2412.12486v3.

[860] Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. Memorag: Boosting long context processing with global memory-enhanced retrieval augmentation, arXiv preprint arXiv:2409.05591, 2025. URL https://arxiv.org/abs/2409.05591.

[861] Kangan Qian, Sicong Jiang, Yang Zhong, Ziang Luo, Zilin Huang, Tianze Zhu, Kun Jiang, Mengmeng Yang, Zheng Fu, Jinyu Miao, Yining Shi, He Zhe Lim, Li Liu, Tianbao Zhou, Hongyi Wang, Huang Yu, Yifei Hu, Guang Li, Guangyao Chen, Hao Ye, Lijun Sun, and Diange Yang. Agentthink: A unified framework for tool-augmented chain-of-thought reasoning in vision-language models for autonomous driving, arXiv preprint arXiv:2505.15298, 2025. URL https://arxiv.org/abs/2505.15298v3.

[862] Changze Qiao and Mingming Lu. Efficiently enhancing general agents with hierarchical-categorical memory, arXiv preprint arXiv:2505.22006, 2025. URL https://arxiv.org/abs/2505.22006v1.

[852] 彭琪, 李海俊, OghenetegiriTGSido, 和 Christopher D. Manning. 从文本中回答推理步骤不同的开放域问题. 自然语言处理经验方法会议, 2020.

[853] 齐勇, Gabriel Kyebambo, 谢思源, 沈巍, 王胜辉, 谢比特, 何斌, 王志鹏, 和 姜硕. 基于LLM和具身知识图谱的服务机器人安全控制, arXiv预印本 arXiv:2405.17846, 2024. URLhttps://arxiv.org/abs/2405.17846v1.

[854] 齐哲瀚, 刘晓, Iat Long Iong, 赖汉宇, 孙学桥, 杨新悦, 孙嘉戴, 杨宇, 姚顺天, 张天杰, 许伟, 唐杰, 和 董宇晓. Webrl: 通过自进化在线课程强化学习训练LLM网络代理, arXiv预印本 arXiv:2411.02337, 2024. URL https://arxiv.org/abs/2411.02337v3.

[855] 陈倩, 刘伟, 刘洪章, 陈诺, 唐宇帆, 李嘉豪, 杨成, 陈伟泽, 苏宇升, 从丛, 许聚源, 李大海, 刘志远, 孙茂松. Chatdev: 软件开发通信代理, arXiv preprint arXiv:2307.07924, 2024. URL https://arxiv.org/abs/2307.07924.

[856] 陈倩, 韩池, 冯宇, 秦宇嘉, 刘志远, 和季恒. Creator: 用于分离大型语言模型抽象和具体推理的工具创建. 自然语言处理经验方法会议, 2023.

[857] 陈倩, 李嘉豪, 唐宇帆, 刘伟, 王一飞, 谢子豪, 陈伟泽, 杨成, 张颖丽, 刘志远, 孙茂松. 软件开发代理的迭代经验细化, arXiv preprint arXiv:2405.04219, 2024. URL https://arxiv.org/abs/2405.04219v1.

[858] 程乾，Emre Can Acikgoz，齐鹤，王红如，陈秀思，Dilek Hakkani-Tur，Gokhan Tur，以及季航。Toolrl: 奖励是所有工具学习所需的一切，arXiv 预印本 arXiv:2504.13958，2025。URLhttps://arxiv.org/abs/2504.13958v1.

[859] 钱宏进，刘政，张培田，窦志成，以及连德福。通过查询引导的激活重新填充来增强长上下文管理，arXiv 预印本 arXiv:2412.12486，2024。URLhttps://arxiv.org/abs/2412.12486v3.

[860] 钱宏进，刘政，张培田，毛克龙，连德福，窦志成，以及黄铁军。Memorag: 通过全局记忆增强检索增强来提升长上下文处理能力，arXiv 预印本 arXiv:2409.05591，2025。URLhttps://arxiv.org/abs/2409.05591.

[861] 甘甘千，蒋思聪，钟杨，罗向，黄子林，朱天泽，蒋坤，杨萌萌，付正，苗金宇，石一宁，李哲 Lim，刘丽，周天宝，王红毅，黄宇，胡一飞，李光，陈广耀，叶浩，孙立军，杨电。Agentthink：一种用于自动驾驶视觉语言模型的工具增强思维链推理的统一框架，arXiv预印本arXiv:2505.15298，2025。URLhttps://arxiv.org/abs/2505.15298v3.

[862] 长桥和露明明. 基于分层分类记忆高效增强通用智能体, arXiv 预印本 arXiv:2505.22006, 2025. URLhttps://arxiv.org/abs/2505.22006v1.

[863] S Qiao, H Gui, C Lv, Q Jia, H Chen, and N Zhang. Making language models better tool learners with execution feedback. 2023. URL https://arxiv.org/abs/2305.13068.

[864] Binjie Qin, Haohao Mao, Ruipeng Zhang, Y. Zhu, Song Ding, and Xu Chen. Working memory inspired hierarchical video decomposition with transformative representations, arXiv preprint arXiv:2204.10105, 2022. URL https://arxiv.org/abs/2204.10105v3.

[865] Bowen Qin, Binyuan Hui, Lihan Wang, Min Yang, Jinyang Li, Binhua Li, Ruiying Geng, Rongyu Cao, Jian Sun, Luo Si, Fei Huang, and Yongbin Li. A survey on text-to-sql parsing: Concepts, methods, and future directions, arXiv preprint arXiv:2208.13629, 2022. URL https://arxiv.org/abs/2208.13629v1.

[866] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Y. Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shi Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bo Li, Ziwei Tang, Jing Yi, Yu Zhu, Zhenning Dai, Lan Yan, Xin Cong, Ya-Ting Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. Tool learning with foundation models. *ACM Computing Surveys*, 2023.

[867] Yujia Qin, Shi Liang, Yining Ye, Kunlun Zhu, Lan Yan, Ya-Ting Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Marc H. Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. Toolllm: Facilitating large language models to master 16000+ real-world apis. *International Conference on Learning Representations*, 2023.

[868] Zhen Qin and Yiran Zhong. Accelerating toeplitz neural network with constant-time inference complexity. *Conference on Empirical Methods in Natural Language Processing*, 2023.

[869] Zhen Qin, Xiaodong Han, Weixuan Sun, Bowen He, Dong Li, Dongxu Li, Yuchao Dai, Lingpeng Kong, and Yiran Zhong. Toeplitz neural network for sequence modeling. *International Conference on Learning Representations*, 2023.

[870] Zhen Qin, Weigao Sun, Dong Li, Xuyang Shen, Weixuan Sun, and Yiran Zhong. Lightning attention-2: A free lunch for handling unlimited sequence lengths in large language models. arXiv preprint, 2024.

[871] Jiahao Qiu, Xinzhe Juan, Yiming Wang, Ling Yang, Xuan Qi, Tongcheng Zhang, Jiacheng Guo, Yifu Lu, Zixin Yao, Hongru Wang, Shilong Liu, Xun Jiang, Liu Leqi, and Mengdi Wang. Agentdistill: Training-free agent distillation with generalizable mcp boxes, arXiv preprint arXiv:2506.14728, 2025. URL https://arxiv.org/abs/2506.14728v1.

[872] Jiahao Qiu, Fulian Xiao, Yiming Wang, Yuchen Mao, Yijia Chen, Xinzhe Juan, Siran Wang, Xuan Qi, Tongcheng Zhang, Zixin Yao, Jiacheng Guo, Yifu Lu, Charles Argon, Jundi Cui, Daixin Chen, Junran Zhou, Shuyao Zhou, Zhanpeng Zhou, Ling Yang, Shilong Liu, Hongru Wang, Kaixuan Huang, Xun Jiang, Yuming Cao, Yue Chen, Yunfei Chen, Zhengyi Chen, Ruowei Dai, Mengqiu Deng, Jiye Fu, Yu Gu, Zijie Guan, Zirui Huang, Xiaoyan Ji, Yumeng Jiang, Delong Kong, Haolong Li, Jiaqi Li, Ruipeng Li, Tianze Li, Zhuo-Yang Li, Haixia Lian, Meng Lin, Xudong Liu, Jiayi Lu, Jinghan Lu, Wanyu Luo, Ziyue Luo, Zihao Pu, Zhi Qiao, Rui-Fang Ren, Liang Wan, Ruixiang Wang, Tianhui Wang, Yang Wang, Zeyu Wang, Zihua Wang, Yujia Wu, Zhaoyi Wu, Hao Xin, Weiao Xing, Ruojun Xiong, Weijie Xu, Yao Shu, Xiao Yao, Xiaorui Yang, Yuchen Yang, Nan Yi, Jiadong Yu, Yang Yu, Huiting Zeng, Danni Zhang, Yunjie Zhang, Zhaoyu Zhang, Zhiheng Zhang, Xiaofeng Zheng, Peirong Zhou, Li-Ying Zhong, Xiaoyin Zong, Ying Zhao, Zhen Chen, Lin Ding, Xiaoyu Gao, Bingbing

[863] S Qiao, H Gui, C Lv, Q Jia, H Chen, and N Zhang. 使语言模型成为更好的工具学习者，通过执行反馈. 2023. URLhttps://arxiv.org/abs/2305.13068.

[864] 秦斌杰, 毛浩浩, 张瑞鹏, Y. 朱, 丁松, 和 陈旭. 受工作记忆启发的分层视频分解与转换表示, arXiv 预印本 arXiv:2204.10105, 2022. URLhttps://arxiv.org/abs/2204.10105v3.

[865] 秦 Bowen, 贺彬源, 王丽涵, 杨敏, 李金阳, 李丙华, 耿瑞莹, 曹荣宇, 孙健, 梁洛, 黄飞, 李永斌. 文本到SQL解析综述：概念、方法与未来方向，arXiv预印本arXiv:2208.13629, 2022. URLhttps://arxiv.org/abs/2208.13629v1.

[866] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Y. Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shi Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bo Li, Ziwei Tang, Jing Yi, Yu Zhu, Zhenning Dai, Lan Yan, Xin Cong, Ya-Ting Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 工具学习与基础模型. *ACM ComputingSurveys*, 2023.

[867] Yujia Qin, Shi Liang, Yining Ye, Kunlun Zhu, Lan Yan, Ya-Ting Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Marc H. Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. Toolllm：助力大型语言模型掌握 16000+ 现实世界 API.国际学习表示会议, 2023.

[868] Zhen Qin和Zhong Yiran. 常数时间推理复杂度加速的托普利茨神经网络. 自然语言处理经验方法会议,2023.

[869] Zhen Qin, Han Xiaodong, Sun Weixuan, He Bowen, Li Dong, Li Dongxu, Dai Yuchao, Kong Lingpeng, 和 Zhong Yiran. 用于序列建模的托普利茨神经网络. 学习表示国际会议, 2023.

[870] Zhen Qin, Sun Weigao, Li Dong, Shen Xuyang, Sun Weixuan, 和 Zhong Yiran. Lightningattention-2: 大型语言模型中处理无限序列长度的免费午餐. arXiv预印本, 2024.

[871] 邱嘉豪, Juan Xinzhe, Wang Yiming, Yang Ling, Qi Xuan, 张通成, Guo Jiacheng, Lu Yifu, Yao Zixin, 王红如, 刘世龙, Jiang Xun, 刘乐奇, 和 Wang Mengdi. Agentdistill: 无需训练的通用mcp框的智能体蒸馏, arXiv预印本 arXiv:2506.14728, 2025. URLhttps://arxiv.org/abs/2506.14728v1.

[872] 邱嘉豪,肖fulian,王Yiming,毛Yuchen,陈Yijia,Juan Xinzhe,王Siran,齐Xuan,张Tongcheng,姚Zixin,郭Jiacheng,卢Yifu,Argon Charles,崔Jundi,陈Daixin,周Junran,周Shuyao,周Zhanpeng,杨Ling,刘Shilong,王Hongru,黄Kaixuan,蒋Xun,曹Yuming,陈Yue,陈Yunfei,陈Zhengyi,戴Ruowei,邓Mengqiu,付Jiye,顾Yu,关Zijie,黄Zirui,季Xiaoyan,蒋Yumeng,孔Delong,李Haolong,李Jiaqi,李Ruipeng,李Tianze,李Zhuo-Yang,连Haixia,林Meng,刘Xudong,卢Jiayi,卢Jinghan,罗Wanyu,罗Ziyue,浦Zihao,乔Zhi,任Rui-Fang,万Liang,王Ruixiang,王Tianhui,王Yang,王Zeyu,王Zihua,吴Yujia,吴Zhaoyi,辛Hao,刑Weiao,熊Ruojun,徐Weijie,舒Yao,姚Xiao,杨Xiaorui,杨Yuchen,伊Nan,于Jiadong,于Yang,曾Huiting,张Danni,张Yunjie,张Zhaoyu,张Zhiheng,郑Xiaofeng,周Peirong,钟Li-Ying,宗Xiaoyin,赵Ying,陈Zhen,丁Lin,高Xiaoyu,冰冰

Gong, Yichao Li, Yang Liao, Guang Ma, Tianyuan Ma, Xinrui Sun, Tianyi Wang, Han Xia, Ruobing Xian, Gen Ye, Tengfei Yu, Wentao Zhang, Yuxi Wang, Xi Gao, and Mengdi Wang. On path to multimodal historical reasoning: Histbench and histagent, arXiv preprint arXiv:2505.20246, 2025. URL https://arxiv.org/abs/2505.20246v3.

[873] Ruidi Qiu, Grace Li Zhang, Rolf Drechsler, Ulf Schlichtmann, and Bing Li. Autobench: Automatic testbench generation and evaluation using llms for hdl design. *Workshop on Machine Learning for CAD*, 2024.

[874] Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Jirong Wen. Tool learning with large language models: A survey. *Frontiers Comput. Sci.*, 2024.

[875] Xiaoye Qu, Yafu Li, Zhao yu Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, Peng Li, Wei Wei, Jing Shao, Chaochao Lu, Yue Zhang, Xian-Sheng Hua, Bowen Zhou, and Yu Cheng. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond, arXiv preprint arXiv:2503.21614, 2025. URL https://arxiv.org/abs/2503.21614v1.

[876] Victor Quintanar-Zilinskas. A neuromimetic approach to the serial acquisition, long-term storage, and selective utilization of overlapping memory engrams. *bioRxiv*, 2019.

[877] Stephan Raaijmakers, Roos Bakker, Anita Cremers, Roy de Kleijn, Tom Kouwenhoven, and Tessa Verhoef. Memory-augmented generative adversarial transformers, arXiv preprint arXiv:2402.19218, 2024. URL https://arxiv.org/abs/2402.19218.

[878] Ella Rabinovich and Ateret Anaby-Tavor. On the robustness of agentic function calling. *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, 2025.

[879] Neil C. Rabinowitz, Frank Perbet, H. F. Song, Chiyuan Zhang, S. Eslami, and M. Botvinick. Machine theory of mind. *International Conference on Machine Learning*, 2018.

[880] Zackary Rackauckas. Rag-fusion: a new take on retrieval-augmented generation. *International Journal on Natural Language Computing*, 2024.

[881] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 2021.

[882] Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 2019.

[883] Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, S. Sakib, Most. Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 2024.

[884] Keshav Ramji, Young-Suk Lee, R. Astudillo, M. Sultan, Tahira Naseem, Asim Munawar, Radu Florian, and S. Roukos. Self-refinement of language models from external proxy metrics feedback, arXiv preprint arXiv:2403.00827, 2024. URL https://arxiv.org/abs/2403.00827v1.

Gong, Yichao Li, Yang Liao, Guang Ma, Tianyuan Ma, Xinrui Sun, Tianyi Wang, Han Xia, Ruobing Xian, Gen Ye, Tengfei Yu, Wentao Zhang, Yuxi Wang, Xi Gao, and Mengdi Wang. On path to multimodal historical reasoning: Histbench and histagent, arXiv preprint arXiv:2505.20246, 2025. URL https://arxiv.org/abs/2505.20246v3.

[873] 邱瑞迪, 张格蕾丝, 德雷克斯勒·罗尔夫, 施利茨曼·乌尔夫, 李冰. Autobench: 使用LLM自动生成和评估hdl设计的测试平台. *CAD*机器学习研讨会, 2024.

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Jirong Wen. 基于大型语言模型的工具学习：一项调查。*Frontiers Comput.Sci.*, 2024.

[875] Xiaoye Qu, Yafu Li, Zhao yu Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, Peng Li, Wei Wei, Jing Shao, Chaochao Lu, Yue Zhang, Xian- Sheng Hua, Bowen Zhou, and Yu Cheng. 大型推理模型的推理研究：语言、多模态和超越，arXiv 预印本 arXiv:2503.21614，2025。URLhttps://arxiv.org/abs/2503.21614v1.

[876] Victor Quintanar-Zilinskas. 一种神经模拟方法，用于串行获取、长期存储和选择性利用重叠记忆突触。*bioRxiv*, 2019。

[877] Stephan Raaijmakers, Roos Bakker, Anita Cremers, Roy de Kleijn, Tom Kouwenhoven, and Tessa Verhoef. 增强记忆的生成对抗变换器，arXiv 预印本 arXiv:2402.19218，2024。
URLhttps://arxiv.org/abs/2402.19218.

[878] Ella Rabinovich and Ateret Anaby-Tavor. On the robustness of agentic function calling. *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, 2025.

[879] Neil C. Rabinowitz, Frank Perbet, H. F. Song, Chiyuan Zhang, S. Eslami, and M. Botvinick. 心理机器理论. 机器学习国际会议, 2018.

[880] Zackary Rackauckas. Rag-fusion：一种关于检索增强生成的新方法. 自然语言计算国际期刊, 2024.

[881] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and I. Sutskever. 从自然语言监督中学习可迁移的视觉模型. 机器学习国际会议, 2021.

[882] Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 使用统一的文本到文本转换器探索迁移学习的极限. 机器学习研究杂志, 2019.

[883] Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, S. Sakib, Most. Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 大型语言模型综述：架构、应用、分类、开放问题和挑战. *IEEEAccess*, 2024.

[884] Keshav Ramji, Young-Suk Lee, R. Astudillo, M. Sultan, Tahira Naseem, Asim Munawar, Radu Florian, and S. Roukos. 基于外部代理指标反馈的语言模型自完善, arXiv 预印本 arXiv:2403.00827, 2024.
URLhttps://arxiv.org/abs/2403.00827v1.

[885] Sumedh Rasal. An artificial neuron for enhanced problem solving in large language models, arXiv preprint arXiv:2404.14222, 2024. URL https://arxiv.org/abs/2404.14222v1.

[886] Shaina Raza, Ranjan Sapkota, Manoj Karkee, and Christos Emmanouilidis. Trism for agentic ai: A review of trust, risk, and security management in llm-based agentic multi-agent systems, arXiv preprint arXiv:2506.04133, 2025. URL https://arxiv.org/abs/2506.04133v2.

[887] Jing Ren and Feng Xia. Brain-inspired artificial intelligence: A comprehensive review, arXiv preprint arXiv:2408.14811, 2024. URL https://arxiv.org/abs/2408.14811v1.

[888] Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. Towards scientific intelligence: A survey of llm-based scientific agents, arXiv preprint arXiv:2503.24047, 2025. URL https://arxiv.org/abs/2503.24047v2.

[889] Xubin Ren, Jiabin Tang, Dawei Yin, Nitesh V. Chawla, and Chao Huang. A survey of large language models for graphs. *Knowledge Discovery and Data Mining*, 2024.

[890] Yi Ren, Shangmin Guo, Linlu Qiu, Bailin Wang, and Danica J. Sutherland. Bias amplification in language model evolution: An iterated learning perspective. *Neural Information Processing Systems*, 2024.

[891] Alireza Rezazadeh, Zichao Li, Wei Wei, and Yujia Bao. From isolated conversations to hierarchical schemas: Dynamic tree memory representation for llms. *International Conference on Learning Representations*, 2024.

[892] Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. Are red roses red? evaluating consistency of question-answering models. *Annual Meeting of the Association for Computational Linguistics*, 2019.

[893] Yara Rizk, Abhishek Bhandwalder, S. Boag, Tathagata Chakraborti, Vatche Isahagian, Y. Khazaeni, Falk Pollock, and Merve Unuvar. A unified conversational assistant framework for business process automation, arXiv preprint arXiv:2001.03543, 2020. URL https://arxiv.org/abs/2001.03543v1.

[894] Yara Rizk, Vatche Isahagian, S. Boag, Y. Khazaeni, Merve Unuvar, Vinod Muthusamy, and Rania Y. Khalaf. A conversational digital assistant for intelligent process automation. *International Conference on Business Process Management*, 2020.

[895] S. Rizvi, Nazreen Pallikkavaliyaveetil, David Zhang, Zhuoyang Lyu, Nhi Nguyen, Haoran Lyu, B. Christensen, J. O. Caro, Antonio H. O. Fonseca, E. Zappala, Maryam Bagherian, Christopher Averill, C. Abdallah, Amin Karbasi, Rex Ying, M. Brbic, R. M. Dhodapkar, and David van Dijk. Fimp: Foundation model-informed message passing for graph neural networks, arXiv preprint arXiv:2210.09475v5, 2022. URL https://arxiv.org/abs/2210.09475v5.

[896] Joshua Robinson, Christopher Rytting, and D. Wingate. Leveraging large language models for multiple choice question answering. *International Conference on Learning Representations*, 2022.

[897] Juri Di Rocco, D. D. Ruscio, Claudio Di Sipio, P. T. Nguyen, and Riccardo Rubei. On the use of large language models in model-driven engineering. *Journal of Software and Systems Modeling*, 2024.

[898] Juan David Salazar Rodriguez, Sam Conrad Joyce, and Julfendi Julfendi. Using customized gpt to develop prompting proficiency in architectural ai-generated images, arXiv preprint arXiv:2504.13948, 2025. URL https://arxiv.org/abs/2504.13948v2.

[885] Sumedh Rasal. 大型语言模型中增强问题解决能力的人工神经元，arXiv 预印本 arXiv:2404.14222，2024。URLhttps://arxiv.org/abs/2404.14222v1.

[886] Shaina Raza, Ranjan Sapkota, Manoj Karkee, 和 Christos Emmanouilidis. Trism for agentic ai：基于 LLM 的自主多智能体系统中的信任、风险和安全管理综述，arXiv 预印本 arXiv:2506.04133，2025。URLhttps://arxiv.org/abs/2506.04133v2.

[887] Jing Ren 和 Feng Xia. 受大脑启发的人工智能：综合综述，arXiv 预印本 arXiv:2408.14811，2024。URLhttps://arxiv.org/abs/2408.14811v1.

[888] 沈仁，蒲建，镇江仁，冷春林，谢灿，张嘉俊. 迈向科学智能：基于LLM的科学代理调查，arXiv预印本arXiv:2503.24047，2025. URLhttps://arxiv.org/abs/2503.24047v2.

[889] 任旭斌，唐嘉斌，尹道伟，Nitesh V. Chawla，黄超. 图形大型语言模型调查. 知识发现与数据挖掘, 2024.

[890] 任毅，郭尚民，邱林路，王百林，Danica J. Sutherland. 语言模型进化中的偏差放大：迭代学习视角. 神经信息处理系统, 2024.

[891] Alireza Rezazadeh，李志超，魏伟，包宇嘉. 从孤立对话到层次化模式：动态树内存表示为 LLM. 国际学习表征会议, 2024.

[892] Marco Tulio Ribeiro, Carlos Guestrin, 和 Sameer Singh. Are red roses red? 评估问答模型的连贯性。 计算语言学协会年会, 2019。

[893] Yara Rizk, Abhishek Bhandwalder, S. Boag, Tathagata Chakraborti, Vatche Isahagian, Y. Khazaeni, Falk Pollock, 和 Merve Unuvar. 用于业务流程自动化的统一对话助手框架，arXiv 预印本 arXiv:2001.03543，2020。URLhttps://arxiv.org/abs/2001.03543v1.

[894] Yara Rizk, Vatche Isahagian, S. Boag, Y. Khazaeni, Merve Unuvar, Vinod Muthusamy, 和 Rania Y. Khalaf. 用于智能流程自动化的对话式数字助手。 业务流程管理国际会议 , 2020。

[895] S. Rizvi, Nazreen Pallikkavaliyaveetil, David Zhang, Zhuoyang Lyu, Nhi Nguyen, Haoran Lyu, B. Christensen, J. O. Caro, Antonio H. O. Fonseca, E. Zappala, Maryam Bagherian, Christopher Averill, C. Abdallah, Amin Karbasi, Rex Ying, M. Brbic, R. M. Dhodapkar, and David van Dijk. Fimp: 基于大型语言模型的消息传递用于图神经网络, arXiv preprint arXiv:2210.09475v5, 2022. URLhttps://arxiv.org/abs/2210.09475v5.

[896] Joshua Robinson, Christopher Rytting, and D. Wingate. 利用大型语言模型进行选择题回答. 国际学习表示会议, 2022.

[897] Juri Di Rocco, D. D. Ruscio, Claudio Di Sipio, P. T. Nguyen, and Riccardo Rubei. 大型语言模型在模型驱动工程中的应用. 软件与系统建模杂志, 2024.

[898] Juan David Salazar Rodriguez, Sam Conrad Joyce, and Julfendi Julfendi. 使用定制 gpt 开发建筑 ai 生成的图像提示能力, arXiv preprint arXiv:2504.13948, 2025. URLhttps://arxiv.org/abs/2504.13948v2.

[899] Albert Roethel, M. Ganzha, and Anna Wr'oblewska. Enriching language models with graph-based context information to better understand textual data. *Electronics*, 2023.

[900] Reudismam Rolim, Gustavo Soares, Rohit Gheyi, and Loris D'antoni. Learning quick fixes from code repositories. *Brazilian Symposium on Software Engineering*, 2018.

[901] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. *Computer Vision and Pattern Recognition*, 2021.

[902] Hayley Ross, A. Mahabaleshwarkar, and Yoshi Suhara. When2call: When (not) to call tools. *North American Chapter of the Association for Computational Linguistics*, 2025.

[903] J. Rosser and Jakob N. Foerster. Agentbreeder: Mitigating the ai safety impact of multi-agent scaffolds, arXiv preprint arXiv:2502.00757, 2025. URL https://arxiv.org/abs/2502.00757v3.

[904] Federico Rossi, Saptarshi Bandyopadhyay, Michael T. Wolf, and M. Pavone. Review of multi-agent algorithms for collective behavior: a structural taxonomy, arXiv preprint arXiv:1803.05464, 2018. URL https://arxiv.org/abs/1803.05464v1.

[905] Federico Rossi, Saptarshi Bandyopadhyay, Michael T. Wolf, and M. Pavone. Multi-agent algorithms for collective behavior: A structural and application-focused atlas, arXiv preprint arXiv:2103.11067, 2021. URL https://arxiv.org/abs/2103.11067v1.

[906] Alex Roxin and Stefano Fusi. Efficient partitioning of memory systems and its importance for memory consolidation. *PLoS Comput. Biol.*, 2013.

[907] Kaushik Roy, Yuxin Zi, Vignesh Narayanan, Manas Gaur, and Amit P. Sheth. Knowledge-infused self attention transformers, arXiv preprint arXiv:2306.13501, 2023. URL https://arxiv.org/abs/2306.13501v1.

[908] Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Jiayang Cheng, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Neural Information Processing Systems*, 2024.

[909] Jingqing Ruan, Yihong Chen, Bin Zhang, Zhiwei Xu, Tianpeng Bao, Guoqing Du, Shiwei Shi, Hangyu Mao, Ziyue Li, Xingyu Zeng, and Rui Zhao. Tptu: Large language model-based ai agents for task planning and tool usage, arXiv preprint arXiv:2308.03427, 2023. URL https://arxiv.org/abs/2308.03427.

[910] M. Russak, Umar Jamil, Christopher Bryant, Kiran Kamble, Axel Magnuson, Mateusz Russak, and Waseem Alshikh. Writing in the margins: Better inference pattern for long context retrieval, arXiv preprint arXiv:2408.14906, 2024. URL https://arxiv.org/abs/2408.14906v1.

[911] Hyun Ryu and Eric Kim. Closer look at efficient inference methods: A survey of speculative decoding, arXiv preprint arXiv:2411.13157, 2024. URL https://arxiv.org/abs/2411.13157v2.

[912] Iman Saberi and Fatemeh Fard. Context-augmented code generation using programming knowledge graphs, arXiv preprint arXiv:2410.18251, 2024. URL https://arxiv.org/abs/2410.18251v2.

[899] Albert Roethel, M. Ganzha, 和 Anna Wr'oblewska. 使用基于图的上下文信息丰富语言模型以更好地理解文本数据。 *Electronics*, 2023。

[900] Reudismam Rolim, Gustavo Soares, Rohit Gheyi, 和 Loris D'antoni. 从代码仓库中学习快速修复。 *Brazilian Symposium onSoftware Engineering*, 2018。

[901] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, 和 B. Ommer. 使用潜在扩散模型进行高分辨率图像合成。 *ComputerVision and Pattern Recognition*, 2021。

[902] Hayley Ross, A. Mahabaleshwarkar, 和 Yoshi Suhara. When2call: 何时（不）调用工具。 *NorthAmerican Chapter of the Association forComputationalLinguistics*, 2025。

[903] J. Rosser 和 Jakob N. Foerster. Agentbreeder: 减轻多智能体脚手架的 AI 安全影响，arXiv 预印本 arXiv:2502.00757，2025。URLhttps://arxiv.org/abs/2502.00757v3.

[904] Federico Rossi, Saptarshi Bandyopadhyay, Michael T. Wolf, and M. Pavone. 集体行为的多智能体算法综述：一种结构化分类法，arXiv 预印本 arXiv:1803.05464，2018。URL https://arxiv.org/abs/1803.05464v1.

[905] Federico Rossi, Saptarshi Bandyopadhyay, Michael T. Wolf, and M. Pavone. 集体行为的多智能体算法：一种结构化和应用导向的图谱，arXiv 预印本 arXiv:2103.11067，2021。
URLhttps://arxiv.org/abs/2103.11067v1.

[906] Alex Roxin and Stefano Fusi. 内存系统的有效划分及其对记忆巩固的重要性。 *PLoSComput.Biol.*, 2013。 y

[907] Kaushik Roy, Yuxin Zi, Vignesh Narayanan, Manas Gaur, and Amit P. Sheth. 知识注入的自注意力Transformer，arXiv 预印本 arXiv:2306.13501，2023。URL https://arxiv.org/abs/2306.13501v1.

[908] 董宇如, 邱林, 胡祥坤, 张天航, 石鹏, 常树晨, 成嘉阳, 王存祥, 孙时超, 李环宇, 张子昭, 王斌杰, 蒋嘉荣, 何通, 王志国, 刘鹏飞, 张越, 张铮. Ragchecker: 一种用于诊断检索增强生成的细粒度框架. 神经信息处理系统, 2024.

[909] 阮静清, 陈一红, 张斌, 许志伟, 包天鹏, 杜国庆, 石世伟, 毛杭宇, 李子越, 曾行宇, 赵瑞. Tptu: 基于大型语言模型的任务规划和工具使用的人工智能代理, arXiv 预印本 arXiv:2308.03427, 2023.
URLhttps://arxiv.org/abs/2308.03427.

[910] M. Russak, Umar Jamil, Christopher Bryant, Kiran Kamble, Axel Magnuson, Mateusz Russak, and Waseem Alshikh. 边缘写作：长上下文检索的更好推理模式，arXiv 预印本 arXiv:2408.14906，2024。
URLhttps://arxiv.org/abs/2408.14906v1.

[911] Hyun Ryu and Eric Kim. 更仔细地查看高效推理方法：推测解码调查，arXiv 预印本 arXiv:2411.13157，2024。
URLhttps://arxiv.org/abs/2411.13157v2.

[912] Iman Saberi and Fatemeh Fard. 使用编程知识图进行上下文增强代码生成，arXiv 预印本 arXiv:2410.18251，2024。URLhttps://arxiv.org/abs/2410.18251v2.

[913] Abdulfattah Safa and Gözde Gül Sahin. A zero-shot open-vocabulary pipeline for dialogue understanding. *North American Chapter of the Association for Computational Linguistics*, 2024.

[914] Alireza Akhavan Safaei, Pegah Saboori, Reza Ramezani, and Mohammadali Nematbakhsh. Kglm-qa: A novel approach for knowledge graph-enhanced large language models for question answering. *Conference on Information and Knowledge Technology*, 2024.

[915] Avirup Saha, Lakshmi Mandal, Balaji Ganesan, Sambit Ghosh, Renuka Sindhgatta, Carlos Eberhardt, Dan Debrunner, and Sameep Mehta. Sequential api function calling using graphql schema. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[916] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, S. Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications, arXiv preprint arXiv:2402.07927, 2024. URL https://arxiv.org/abs/2402.07927v2.

[917] Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yassine Benajiba. Meminsight: Autonomous memory augmentation for llm agents, arXiv preprint arXiv:2503.21760, 2025. URL https://arxiv.org/abs/2503.21760.

[918] Jefferson Salan, Devyn E Smith, Erica S Shafer, and Rachel A Diana. Variation in encoding context benefits item recognition. *Memory & Cognition*, 2024.

[919] Alaa Saleh, Sasu Tarkoma, Praveen Kumar Donta, Naser Hossein Motlagh, S. Dustdar, Susanna Pirttikangas, and Lauri Lov'en. Usercentrix: An agentic memory-augmented ai framework for smart spaces, arXiv preprint arXiv:2505.00472, 2025. URL https://arxiv.org/abs/2505.00472v1.

[920] Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. In-context impersonation reveals large language models' strengths and biases. *Neural Information Processing Systems*, 2023.

[921] A. Samsonovich. Toward a unified catalog of implemented cognitive architectures. *Biologically Inspired Cognitive Architectures*, 2010.

[922] Narendra Reddy Sanikommu. Model context protocol: Enhancing llm performance for observability and analytics. *European journal of computer science and information technology*, 2025.

[923] S. Santhanam. Context based text-generation using lstm networks, arXiv preprint arXiv:2005.00048, 2020. URL https://arxiv.org/abs/2005.00048v1.

[924] G. Santos, Rita Maria Silva Julia, and Marcelo Zanchetta do Nascimento. Diverse prompts: Illuminating the prompt space of large language models with map-elites. *IEEE Congress on Evolutionary Computation*, 2025.

[925] Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges, arXiv preprint arXiv:2505.10468, 2025. URL https://arxiv.org/abs/2505.10468v4.

[926] Anjana Sarkar and Soumyendu Sarkar. Survey of llm agent communication with mcp: A software design pattern centric review, arXiv preprint arXiv:2506.05364, 2025. URL https://arxiv.org/abs/2506.05364v1.

[913] Abdulfattah Safa 和 Gözde Gül Sahin. 一种零样本开放词汇的对话理解流程. 美国计算语言学协会北美分会, 2024.

[914] Alireza Akhavan Safaei、Pegah Saboori、Reza Ramezani 和 Mohammadali Nematbakhsh. Kglm-qa：一种用于知识图谱增强的大型语言模型的新方法，用于问答。信息与知识技术会议, 2024.

[915] Avirup Saha、Lakshmi Mandal、Balaji Ganesan、Sambit Ghosh、Renuka Sindhgatta、Carlos Eberhardt、Dan Debrunner 和 Sameep Mehta. 使用 graphql schema 的 Sequentialapi 函数调用. 自然语言处理经验方法会议, 2024.

[916] Pranab Sahoo、Ayush Kumar Singh、Sriparna Saha、Vinija Jain、S. Mondal 和 Aman Chadha. 大型语言模型提示工程的系统调查：技术和应用，arXiv 预印本 arXiv:2402.07927，2024. URLhttps://arxiv.org/abs/2402.07927v2.

[917] Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yas- sine Benajiba. Meminsight: 自主记忆增强用于 llm 代理，arXiv 预印本 arXiv:2503.21760, 2025. URLhttps://arxiv.org/abs/2503.21760.

[918] Jefferson Salan, Devyn E Smith, Erica S Shafer, and Rachel A Diana. 编码上下文的变化有利于物品识别。*Memory&Cognition*, 2024。

[919] Alaa Saleh, Sasu Tarkoma, Praveen Kumar Donta, Naser Hossein Motlagh, S. Dustdar, Susanna Pirttikangas, and Lauri Lov'en. Usercentrix: 一个用于智能空间的代理记忆增强人工智能框架，arXiv 预印本 arXiv:2505.00472, 2025. URLhttps://arxiv.org/abs/2505.00472v1.

[920] Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, 和 Zeynep Akata. 上下文模仿揭示了大型语言模型的优势和偏见。*NeuralInformationProcessingSystems*, 2023。

[921] A. Samsonovich. 迈向统一的认知架构实现目录。*BiologicallyInspired CognitiveArchitectures*, 2010。

[922] Narendra Reddy Sanikommu. 模型上下文协议：增强 llm 性能以用于可观察性和分析。*European journalof computer science and information technology*, 2025。

[923] S. Santhanam. 基于上下文的文本生成使用 lstm 网络，arXiv 预印本 arXiv:2005.00048，2020。URLhttps://arxiv.org/abs/2005.00048v1。

[924] G. Santos, Rita Maria Silva Julia, 和 Marcelo Zanchetta do Nascimento. 多样化的提示：使用 map-elites 照亮大型语言模型的提示空间。*IEEE Congress on EvolutionaryComputation*, 2025。

[925] Ranjan Sapkota、Konstantinos I. Roumeliotis 和 Manoj Karkee。AI 代理 vs. 自主 AI：概念分类、应用与挑战，arXiv 预印本 arXiv:2505.10468，2025。URLhttps://arxiv.org/abs/2505.10468v4。

[926] Anjana Sarkar和Soumyendu Sarkar.关于llm代理与mcp通信的调查：以软件设计模式为中心的综述, arXiv预印本arXiv:2506.05364，2025年。URLhttps://arxiv.org/abs/2506.05364v1.

[927] Soumajyoti Sarkar and Leonard Lausen. Testing the limits of unified sequence to sequence llm pretraining on diverse table data tasks, arXiv preprint arXiv:2310.00789, 2023. URL https://arxiv.org/abs/2310.00789v1.

[928] Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. Raptor: Recursive abstractive processing for tree-organized retrieval. *International Conference on Learning Representations*, 2024.

[929] Gabriele Sarti. Umberto-mtsa @ accompl-it: Improving complexity and acceptability prediction with multi-task learning on self-supervised annotations (short paper). *International Workshop on Evaluation of Natural Language and Speech Tools for Italian*, 2020.

[930] Apoorv Saxena, Adrian Kochsiek, and Rainer Gemulla. Sequence-to-sequence knowledge graph completion and question answering. *Annual Meeting of the Association for Computational Linguistics*, 2022.

[931] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, R. Raileanu, M. Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Neural Information Processing Systems*, 2023.

[932] Guido Schillaci, Uwe Schmidt, and Luis Miranda. Prediction error-driven memory consolidation for continual learning: On the case of adaptive greenhouse models. *KI - Künstliche Intelligenz*, 35(1): 71–80, 2021. ISSN 1610-1987. doi: 10.1007/s13218-020-00700-8. URL http://dx.doi.org/10.1007/s13218-020-00700-8.

[933] Florian Schneider, Narges Baba Ahmadi, Niloufar Baba Ahmadi, Iris Vogel, Martin Semmann, and Christian Biemann. Collex - a multimodal agentic rag system enabling interactive exploration of scientific collections. arXiv preprint, 2025.

[934] Sheila Schoepp, Masoud Jafaripour, Yingyue Cao, Tianpei Yang, Fatemeh Abdollahi, Shadan Golestan, Zahin Sufiyan, Osmar R. Zaiane, and Matthew E. Taylor. The evolving landscape of llm- and vlm-integrated reinforcement learning. arXiv preprint, 2025.

[935] Lianlei Shan, Shixian Luo, Zezhou Zhu, Yu Yuan, and Yong Wu. Cognitive memory in large language models, arXiv preprint arXiv:2504.02441, 2025. URL https://arxiv.org/abs/2504.02441v2.

[936] Wenbo Shang and Xin Huang. A survey of large language models on generative graph analytics: Query, learning, and applications, arXiv preprint arXiv:2404.14809v2, 2024. URL https://arxiv.org/abs/2404.14809v2.

[937] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing, arXiv preprint arXiv:2310.10158, 2023. URL https://arxiv.org/abs/2310.10158.

[938] Yutong Shao and N. Nakashole. On linearizing structured data in encoder-decoder language models: Insights from text-to-sql. *North American Chapter of the Association for Computational Linguistics*, 2024.

[939] Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Synthetic prompting: Generating chain-of-thought demonstrations for large language models. *International Conference on Machine Learning*, 2023.

[927] Soumajyoti Sarkar and Leonard Lausen. 测试统一序列到序列 llm 预训练在多样化表格数据任务上的极限, arXiv 预印本 arXiv:2310.00789, 2023. URL https://arxiv.org/abs/2310.00789v1.

[928] Parth Sarthi, SalmanAbdullah, AditiTuli, ShubhKhanna, AnnaGoldie, andChristopherD. Manning. Raptor: 递归抽象处理用于树形组织检索. 学习表示国际会议, 2024.

[929] Gabriele Sarti. Umberto-mtsa @ accompl-it: 使用多任务学习在自监督标注上提高复杂性和可接受性预测 (短篇论文). 意大利自然语言和语音工具评估国际研讨会, 2020.

[930] Apoorv Saxena, Adrian Kochsiek, and Rainer Gemulla. 序列到序列知识图谱补全和问答. 计算语言学协会年度会议, 2022.

[931] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, R. Raileanu, M. Lomeli, Luke Zettlemoyer, Nicola Cancedda, 和 Thomas Scialom. Toolformer: 语言模型可以教自己使用工具。*Neural Information Processing Systems*, 2023。

[932] Guido Schillaci, Uwe Schmidt, 和 Luis Miranda. 预测误差驱动的记忆巩固用于持续学习：以自适应温室模型为例。 *KI* - 人工智能, 35(1):71–80,2021. ISSN 1610-1987. doi: 10.1007/s13218-020-00700-8. URL http://dx.doi.org/10.1007/s13218-020-00700-8.

[933] Florian Schneider, Narges Baba Ahmadi, Niloufar Baba Ahmadi, Iris Vogel, Martin Semmann, 和 Christian Biemann. Collex - 一个多模态智能问答系统，支持交互式探索科学收藏。arXiv 预印本, 2025。

[934] Sheila Schoepp, Masoud Jafaripour, Yingyue Cao, Tianpei Yang, Fatemeh Abdollahi, Shadan Golestan, Zahin Sufiyan, Osmar R. Zaiane, and Matthew E. Taylor. 大型语言模型和视觉语言模型集成强化学习的演变格局。arXiv 预印本，2025。

[935] Lianlei Shan, Shixian Luo, Zezhou Zhu, Yu Yuan, and Yong Wu. 大型语言模型中的认知记忆，arXiv preprintarXiv:2504.02441, 2025. URL https://arxiv.org/abs/2504.02441v2.

[936] Wenbo Shang and Xin Huang. 关于生成式图分析的综述：查询、学习和应用，arXiv preprint arXiv:2404.14809v2, 2024. URLhttps://arxiv.org/abs/2404.14809v2.

[937] 邵云帆, 李林阳, 戴俊奇, 和 邱锡鹏. Character-llm: 一种可训练的角色扮演代理, arXiv 预印本 arXiv:2310.10158, 2023. URLhttps://arxiv.org/abs/2310.10158.

[938] 邵宇彤和N. Nakashole. 在编码器-解码器语言模型中对结构化数据进行线性化：来自文本到SQL的见解。北美学术计算语言学协会分会, 2024.

[939] 邵志红，宫叶云，沈永隆，黄敏莉，段南，陈伟珠。合成提示：为大型语言模型生成思维链演示。机器学习国际会议，2023。

[940] Peter Shaw, Mandar Joshi, James Cohan, Jonathan Berant, Panupong Pasupat, Hexiang Hu, Urvashi Khandelwal, Kenton Lee, and Kristina Toutanova. From pixels to ui actions: Learning to follow instructions via graphical user interfaces. *Neural Information Processing Systems*, 2023.

[941] Jonathan Shen, Ruoming Pang, Ron J. Weiss, M. Schuster, N. Jaitly, Zongheng Yang, Z. Chen, Yu Zhang, Yuxuan Wang, R. Skerry-Ryan, R. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2017.

[942] Junhong Shen, Atishay Jain, Zedian Xiao, Ishan Amlekar, Mouad Hadji, Aaron Podolny, and Ameet Talwalkar. Scribeagent: Towards specialized web agents using production-scale workflow data. 2024.

[943] Junhong Shen, Hao Bai, Lunjun Zhang, Yifei Zhou, Amrith Setlur, Shengbang Tong, Diego Caples, Nan Jiang, Tong Zhang, Ameet Talwalkar, and Aviral Kumar. Thinking vs. doing: Agents that reason by scaling test-time interaction, arXiv preprint arXiv:2506.07976, 2025. URL https://arxiv.org/abs/2506.07976.

[944] Weizhou Shen, Chenliang Li, Fanqi Wan, Shengyi Liao, Shaopeng Lai, Bo Zhang, Yingcheng Shi, Yuning Wu, Gang Fu, Zhansheng Li, Bin Yang, Ji Zhang, Fei Huang, Jingren Zhou, and Ming Yan. Qwenlong-cprs: Towards ∞-llms with dynamic context optimization. arXiv preprint, 2025.

[945] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Y. Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Neural Information Processing Systems*, 2023.

[946] Zhuocheng Shen. Llm with tools: A survey, arXiv preprint arXiv:2409.18807, 2024. URL https://arxiv.org/abs/2409.18807v1.

[947] Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Daniel Y. Fu, Zhiqiang Xie, Beidi Chen, Clark W. Barrett, Joseph Gonzalez, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. High-throughput generative inference of large language models with a single gpu. *International Conference on Machine Learning*, 2023.

[948] Dingfeng Shi, Jingyi Cao, Qianben Chen, Weichen Sun, Weizhen Li, Hongxuan Lu, Fangchen Dong, Tianrui Qin, King Zhu, Minghao Liu, Jian Yang, Ge Zhang, Jiaheng Liu, Changwang Zhang, Jun Wang, Y. Jiang, and Wangchunshu Zhou. Taskcraft: Automated generation of agentic tasks, arXiv preprint arXiv:2506.10055, 2025. URL https://arxiv.org/abs/2506.10055v2.

[949] Han Shi, Jiahui Gao, Xiaozhe Ren, Hang Xu, Xiaodan Liang, Zhenguo Li, and J. Kwok. Sparsebert: Rethinking the importance analysis in self-attention. *International Conference on Machine Learning*, 2021.

[950] Peng Shi, Patrick Ng, Zhiguo Wang, Henghui Zhu, Alexander Hanbo Li, Jun Wang, C. D. Santos, and Bing Xiang. Learning contextual representations for semantic parsing with generation-augmented pre-training. *AAAI Conference on Artificial Intelligence*, 2020.

[951] Weijia Shi, Xiaochuang Han, M. Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and S. Yih. Trusting your evidence: Hallucinate less with context-aware decoding. *North American Chapter of the Association for Computational Linguistics*, 2023.

[940] 彼得·肖，曼达尔·乔希，詹姆斯·科汉，乔纳森·贝尔安特，帕努邦·帕苏帕特，胡鹤祥，乌尔瓦希·卡纳德尔瓦尔，李肯顿，克里斯蒂娜·图塔诺娃。从像素到 UI 操作：通过图形用户界面学习遵循指令。 神经信息处理系统, 2023。

[941] 乔纳森·沈，庞若鸣，罗恩·J·韦斯，M·舒斯特，N·贾蒂利，杨宗恒，陈泽，张宇，王宇轩，R·斯凯里-莱安，R·索罗斯，亚尼斯·阿吉奥米格亚纳基斯，和吴永辉。基于梅尔频谱图预测条件化 WaveNet 的自然语音合成。 *IEEE* 国际声学、语音与信号处理会议, 2017。

[942] 沈俊宏，阿蒂沙伊·贾因，肖泽丹，伊尚·阿姆莱卡尔，穆阿德·哈吉，亚伦·波多尔尼，和阿米特·塔尔沃卡尔。Scribeagent：使用生产规模工作流数据创建专用网络代理。 2024。

[943] 沈俊红, 白浩, 张伦军, 周逸飞, Amrith Setlur, 唐胜榜, Diego Caples, 姜楠, 张通, Ameet Talwalkar, 和 Aviral Kumar. 思考与行动：通过扩展测试时交互进行推理的智能体, arXiv preprint arXiv:2506.07976, 2025. URLhttps://arxiv.org/abs/2506.07976.

[944] 沈伟周, 李晨亮, 万帆奇, 廖胜怡, 赖少鹏, 张博, 石英成, 吴云宁, 傅刚, 李占胜, 杨斌, 张继, 黄飞, 周景仁, 和 阎明. Qwenlong-cprs：迈向 ∞-llms 的动态上下文优化. arXiv preprint, 2025.

[945] 沈永亮, 宋凯涛, 谭旭, 李东升, 陆明伟, 和 Y. Zhuang. Hugginggpt：使用 ChatGPT 及其在 Hugging Face 中的朋友解决 AI 任务. 神经信息处理系统, 2023.

[946] 沈卓成. 带工具的大型语言模型：一项调查，arXiv 预印本 arXiv:2409.18807，2024。URL https://arxiv.org/abs/2409.18807v1.

[947] Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Daniel Y. Fu, Zhiqiang Xie, Beidi Chen, Clark W. Barrett, Joseph Gonzalez, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. 基于单GPU的高吞吐量大型语言模型生成推理。 机器学习国际会议, 2023.

[948] 丁风石, 曹静怡, 陈千本, 孙伟辰, 李伟臻, 陆红轩, 董方辰, 秦天瑞, 朱King, 刘明浩, 杨健, 张格, 刘家恒, 张长旺, 王军, 蒋Y, 周王春树. Taskcraft: 自动生成代理任务, arXiv预印本 arXiv:2506.10055, 2025. URLhttps://arxiv.org/abs/2506.10055v2.

[949] 韩世, 佳人高, 小哲任, 汉旭, 梁晓丹, 李正国, 和 J. Kwok. Sparsebert: 重新思考自注意力中的重要性分析. 国际机器学习会议, 2021.

[950] 彭世, 乔治·宁, 王志国, 朱恒辉, 李汉波, 王军, C. D. 圣托斯, 和 邦翔. 学习上下文表示以生成式预训练进行语义解析. *AAAI* 人工智能会议, 2020.

[951] 石伟嘉, 韩晓川, M. 刘易斯, 叶莉娅·茨维特科娃, 卢克·泽特莫耶, 和 S. 伊. 信任你的证据：通过上下文感知解码减少幻觉. 北美计算语言学协会分会, 2023.

[952] Zhengliang Shi, Shen Gao, Xiuyi Chen, Yue Feng, Lingyong Yan, Haibo Shi, Dawei Yin, Zhumin Chen, Suzan Verberne, and Zhaochun Ren. Tool learning in the wild: Empowering language models as automatic tool agents. *The Web Conference*, 2024.

[953] Jay Shim, Grant Kruttschnitt, Alyssa Ma, Daniel Kim, Benjamin Chek, Athul Anand, Kevin Zhu, and Sean O'Brien. Chain-of-thought augmentation with logit contrast for enhanced reasoning in language models, arXiv preprint arXiv:2407.03600, 2024. URL https://arxiv.org/abs/2407.03600v2.

[954] Jiho Shin, Reem Aleithan, Hadi Hemmati, and Song Wang. Retrieval-augmented test generation: How far are we?, arXiv preprint arXiv:2409.12682, 2024. URL https://arxiv.org/abs/2409.12682v1.

[955] Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, Hyoungseok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, W. Park, Jung-Woo Ha, and Nako Sung. On the effect of pre-training corpora on in-context learning by a large-scale language model. *North American Chapter of the Association for Computational Linguistics*, 2022.

[956] Noah Shinn, Federico Cassano, Beck Labash, A. Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. *Neural Information Processing Systems*, 2023.

[957] Fatemeh Shiri, Xiao-Yu Guo, Mona Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[958] Masoud Shokrnezhad, Hao Yu, T. Taleb, Renwei Li, Kyunghan Lee, Jaeseung Song, and Cedric Westphal. Toward a dynamic future with adaptable computing and network convergence (acnc). *IEEE Network*, 2024.

[959] Connor Shorten, T. Khoshgoftaar, and B. Furht. Text data augmentation for deep learning. *Journal of Big Data*, 2021.

[960] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and D. Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. *Computer Vision and Pattern Recognition*, 2019.

[961] Altun Shukurlu. Improving deep knowledge tracing via gated architectures and adaptive optimization, arXiv preprint arXiv:2504.20070, 2025. URL https://arxiv.org/abs/2504.20070v1.

[962] Lynn L Siegel and M. Kahana. A retrieved context account of spacing and repetition effects in free recall. *Journal of Experimental Psychology. Learning, Memory and Cognition*, 2014.

[963] Aditi Singh, Abul Ehtesham, Gaurav Kumar Gupta, Nikhil Kumar Chatta, Saket Kumar, and T. T. Khoei. Exploring prompt engineering: A systematic review with swot analysis, arXiv preprint arXiv:2410.12843, 2024. URL https://arxiv.org/abs/2410.12843v1.

[964] Anmolika Singh and Yuhang Diao. Leveraging large language models for optimized item categorization using unspsc taxonomy. *International Journal on Cybernetics & Informatics*, 2024.

[952] 郑亮, 高沈, 陈修怡, 冯越, 严凌勇, 石海波, 尹大伟, 陈志民, Verberne Suzan, 和 任赵春. 野外工具学习：使语言模型成为自动工具代理. *The WebConference*, 2024.

[953] Jay Shim, Grant Kruttschnitt, Alyssa Ma, Daniel Kim, Benjamin Chek, Athul Anand, Kevin Zhu, 和 Sean O' Brien. 基于逻辑对比的思维链增强，用于语言模型中的推理增强, arXiv preprint arXiv:2407.03600, 2024. URL https://arxiv.org/abs/2407.03600v2.

[954] Jiho Shin, Reem Aleithan, Hadi Hemmati, 和 王宋. 检索增强测试生成：我们还有多远?, arXiv preprint arXiv:2409.12682, 2024. URL https://arxiv.org/abs/2409.12682v1.

[955] Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, Hyoungseok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, W. Park, Jung-Woo Ha, and Nako Sung. 关于大规模语言模型的预训练语料库对情境学习的影响. 北美学术计算语言学协会, 2022.

[956] Noah Shinn, Federico Cassano, Beck Labash, A. Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: 具有语言强化学习功能的语言代理. 神经信息处理系统, 2023.

[957] Fatemeh Shiri, Xiao-Yu Guo, Mona Far, Xin Yu, Reza Haf, and Yuan-Fang Li. 大规模多模态模型的空间推理能力实证分析. 自然语言处理经验方法会议, 2024.

[958] Masoud Shokrnezhad、Hao Yu、T. Taleb、Renwei Li、Kyunghan Lee、Jaeseung Song 和 Cedric Westphal。迈向动态未来：适应计算和网络融合（ACNC）。*IEEE Network*, 2024.

[959] Connor Shorten、T. Khoshgoftaar 和 B. Furht。深度学习的文本数据增强。《大数据杂志》，2021年。

[960] Mohit Shridhar、Jesse Thomason、Daniel Gordon、Yonatan Bisk、Winson Han、Roozbeh Mottaghi、Luke Zettlemoyer 和 D. Fox。Alfred：一个用于解释日常任务中基于指令的基准。计算机视觉与模式识别，2019。

[961] Altun Shukurlu. 通过门控架构和自适应优化改进深度知识追踪，arXiv 预印本 arXiv:2504.20070，2025。URLhttps://arxiv.org/abs/2504.20070v1.

[962] Lynn L Siegel 和 M. Kahana. 关于自由回忆中间距和重复效应的检索上下文解释. 《实验心理学杂志. 学习、记忆与认知》, 2014.

[963] Aditi Singh, Abul Ehtesham, Gaurav Kumar Gupta, Nikhil Kumar Chatta, Saket Kumar, 和 T. T. Khoei. 探索提示工程：具有SWOT分析的系统性综述, arXiv 预印本 arXiv:2410.12843, 2024. URLhttps://arxiv.org/abs/2410.12843v1.

[964] Anmolika Singh 和 Yuhang Diao. 利用大型语言模型进行基于 UNSPSC 分类体系的优化项目分类. 《国际网络与信息杂志》, 2024.

[965] Joykirat Singh, Raghav Magazine, Yash Pandya, and A. Nambi. Agentic reasoning and tool integration for llms via reinforcement learning, arXiv preprint arXiv:2505.01441, 2025. URL https://arxiv.org/abs/2505.01441v1.

[966] Krishnakant Singh, Thanush Navaratnam, Jannik Holmer, Simone Schaub-Meyer, and Stefan Roth. Is synthetic data all we need? benchmarking the robustness of models trained with synthetic images. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

[967] Ramneet Singh, Sathvik Joel, Abhav Mehrotra, Nalin Wadhwa, Ramakrishna Bairi, Aditya Kanade, and Nagarajan Natarajan. Code researcher: Deep research agent for large systems code and commit history, arXiv preprint arXiv:2506.11060, 2025. URL https://arxiv.org/abs/2506.11060v1.

[968] Aarush Sinha and CU Omkumar. Gmlm: Bridging graph neural networks and language models for heterophilic node classification, arXiv preprint arXiv:2503.05763, 2025. URL https://arxiv.org/abs/2503.05763v3.

[969] Sanchit Sinha, Yuguang Yue, Victor Soto, Mayank Kulkarni, Jianhua Lu, and Aidong Zhang. Maml-en-llm: Model agnostic meta-training of llms for improved in-context learning. *Knowledge Discovery and Data Mining*, 2024.

[970] Colin Sisate, Alistair Goldfinch, Vincent Waterstone, Sebastian Kingsley, and Mariana Blackthorn. Contextually entangled gradient mapping for optimized llm comprehension, arXiv preprint arXiv:2502.00048, 2025. URL https://arxiv.org/abs/2502.00048v1.

[971] Paloma Sodhi, S. R. K. Branavan, Yoav Artzi, and Ryan McDonald. Step: Stacked llm policies for web actions, arXiv preprint arXiv:2310.03720, 2024. URL https://arxiv.org/abs/2310.03720.

[972] Manthankumar Solanki. Efficient document retrieval with g-retriever. arXiv preprint, 2025.

[973] Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, G. Cerono, Yongmei Shi, Angela Rizk-Jackson, Sharat Israni, Charlotte A. Nelson, Sui Huang, and Sergio Baranzini. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 2023.

[974] Lilian Some, Wenli Yang, Michael Bain, and Byeong Kang. A comprehensive survey on integrating large language models with knowledge-based methods. *Knowledge-Based Systems*, 2025.

[975] Chan Hee Song, Jiaman Wu, Clay Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. *IEEE International Conference on Computer Vision*, 2022.

[976] Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. arXiv preprint, 2025.

[977] Woomin Song, Seunghyuk Oh, Sangwoo Mo, Jaehyung Kim, Sukmin Yun, Jung-Woo Ha, and Jinwoo Shin. Hierarchical context merging: Better long context understanding for pre-trained llms. *International Conference on Learning Representations*, 2024.

[965] Joykirat Singh, Raghav Magazine, Yash Pandya, and A. Nambi. 基于强化学习的智能推理和工具集成用于大型语言模型, arXiv 预印本 arXiv:2505.01441, 2025. URLhttps://arxiv.org/abs/2505.01441v1.

[966] Krishnakant Singh, Thanush Navaratnam, Jannik Holmer, Simone Schaub-Meyer, 和 Stefan Roth. 我们真的只需要合成数据吗？评估使用合成图像训练的模型的鲁棒性。*2024 IEEE/CVF*计算机视觉与模式识别研讨会 *(CVPRW)*, 2024.

[967] Ramneet Singh, Sathvik Joel, Abhav Mehrotra, Nalin Wadhwa, Ramakrishna Bairi, Aditya Kanade, and Nagarajan Natarajan. 代码研究员：大型系统代码和提交历史深度研究代理, arXiv 预印本 arXiv:2506.11060，2025年。URL https://arxiv.org/abs/2506.11060v1.

[968] Aarush Sinha和CU Omkumar. Gmlm: 桥接图神经网络和语言模型用于异质节点分类，arXiv预印本 arXiv:2503.05763，2025年。URLhttps://arxiv.org/abs/2503.05763v3.

[969] Sanchit Sinha、Yuguang Yue、Victor Soto、Mayank Kulkarni、Jianhua Lu和Aidong Zhang。Maml- en-llm: 为改进上下文学习进行模型无关元训练的llms。 知识发现与数据挖掘，2024年。

[970] Colin Sisate、Alistair Goldfinch、Vincent Waterstone、Sebastian Kingsley和Mariana Black- thorn。上下文纠缠梯度映射用于优化llm理解，arXiv预印本 arXiv:2502.00048，2025年。
URLhttps://arxiv.org/abs/2502.00048v1.

[971] Paloma Sodhi, S. R. K. Branavan, Yoav Artzi, and Ryan McDonald. Step: Stacked llm policies for web actions, arXiv preprint arXiv:2310.03720, 2024. URL https://arxiv.org/abs/2310.03720.

[972] Manthankumar Solanki. 基于g-retriever的高效文档检索. arXiv预印本, 202 5.

[973] Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, G. Cerono, Yongmei Shi, Angela Rizk-Jackson, Sharat Israni, Charlotte A. Nelson, Sui Huang, and Sergio Baranzini. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 2023.

[974] Lilian Some, Wenli Yang, Michael Bain, 和 Byeong Kang. 关于将大型语言模型与基于知识的方法集成的综合调查. *Knowledge-Based Systems*, 2025.

[975] Chan Hee Song, Jiaman Wu, Clay Washington, Brian M. Sadler, Wei-Lun Chao, 和 Yu Su. Llm-planner: 为具身智能体使用大型语言模型的少样本接地规划. *IEEEInternational Conferenceon Computer Vision*, 2022.

[976] Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, 和 Ji-Rong Wen. R1-searcher：通过强化学习激励大型语言模型中的搜索能力. arXiv预印本, 2025.

[977] Woomin Song, Seunghyuk Oh, Sangwoo Mo, Jaehyung Kim, Sukmin Yun, Jung-Woo Ha, 和 Jinwoo Shin. 层级上下文合并：为预训练的大型语言模型提供更好的长上下文理解.*International Conferenceon LearningRepresentations*, 2024.

[978] Woomin Song, Sai Muralidhar Jayanthi, S. Ronanki, Kanthashree Mysore Sathyendra, Jinwoo Shin, A. Galstyan, Shubham Katiyar, and S. Bodapati. Compress, gather, and recompute: Reforming long-context processing in transformers, arXiv preprint arXiv:2506.01215, 2025. URL https://arxiv.org/abs/2506.01215v1.

[979] Yewei Song, Xunzhu Tang, Cedric Lothritz, Saad Ezzini, Jacques Klein, Tegawend'e F. Bissyand'e, A. Boytsov, Ulrick Ble, and Anne Goujon. Callnavi, a challenge and empirical study on llm function calling and routing, arXiv preprint arXiv:2501.05255, 2025. URL https://arxiv.org/abs/2501.05255v2.

[980] Yueqi Song, Frank Xu, Shuyan Zhou, and Graham Neubig. Beyond browsing: Api-based web agents, arXiv preprint arXiv:2410.16464, 2025. URL https://arxiv.org/abs/2410.16464.

[981] S. Srinivasa and Jayati Deshmukh. Paradigms of computational agency. *Novel Approaches to Information Systems Design*, 2021.

[982] B. Staresina, R. Henson, N. Kriegeskorte, and Arjen Alink. Episodic reinstatement in the medial temporal lobe. *Journal of Neuroscience*, 2012.

[983] T. Staudigl, C. Vollmar, S. Noachtar, and S. Hanslmayr. Temporal-pattern similarity analysis reveals the beneficial and detrimental effects of context reinstatement on human memory. *Journal of Neuroscience*, 2015.

[984] Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. Chain of thoughtlessness? an analysis of cot in planning. *Neural Information Processing Systems*, 2024.

[985] R. Sterken and James Ravi Kirkpatrick. Conversational alignment with artificial intelligence in context. *Philosophical Perspectives*, 2025.

[986] Paul Stoewer, Achim Schilling, Andreas K. Maier, and Patrick Krauss. Multi-modal cognitive maps based on neural networks trained on successor representations, arXiv preprint arXiv:2401.01364, 2023. URL https://arxiv.org/abs/2401.01364v1.

[987] Olly Styles, Sam Miller, Patricio Cerda-Mardini, T. Guha, Victor Sanchez, and Bertie Vidgen. Workbench: a benchmark dataset for agents in a realistic workplace setting, arXiv preprint arXiv:2405.00823, 2024. URL https://arxiv.org/abs/2405.00823v2.

[988] Guangxin Su, Yifan Zhu, Wenjie Zhang, Hanchen Wang, and Ying Zhang. Bridging large language models and graph structure learning models for robust representation learning, arXiv preprint arXiv:2410.12096, 2024. URL https://arxiv.org/abs/2410.12096v1.

[989] Hong Su, Elke A. Rundensteiner, and Murali Mani. Automaton in or out: run-time plan optimization for xml stream processing. *International Symposium on Signal Processing Systems*, 2008.

[990] Hongjin Su, Ruoxi Sun, Jinsung Yoon, Pengcheng Yin, Tao Yu, and Sercan Ö. Arık. Learn-by-interact: A data-centric framework for self-adaptive agents in realistic environments. 2025.

[991] Jinyan Su, Jennifer Healey, Preslav Nakov, and Claire Cardie. Between underthinking and overthinking: An empirical study of reasoning length and correctness in llms, arXiv preprint arXiv:2505.00127, 2025. URL https://arxiv.org/abs/2505.00127v1.

[978] 宋宇民, Sai Muralidhar Jayanthi, S. Ronanki, Kanthashree Mysore Sathyendra, Jinwoo Shin, A. Galstyan, Shubham Katiyar, 和 S. Bodapati. 压缩、收集和重新计算：改进 Transformer 中的长上下文处理, arXiv 预印本 arXiv:2506.01215, 2025. URLhttps://arxiv.org/abs/2506.01215v1.

[979] 宋越, Xunzhu Tang, Cedric Lothritz, Saad Ezzini, Jacques Klein, Tegawend' e F. Bissyand' e, A. Boytsov, Ulrick Ble, 和 Anne Goujon. Callnavi, 关于 LLM 函数调用和路由的挑战和实证研究, arXiv 预印本 arXiv:2501.05255, 2025. URLhttps://arxiv.org/abs/2501.05255v2.

[980] 宋越琪, 徐帆, 周舒岩, 和 Graham Neubig. 超越浏览: 基于API的网页代理, arXiv 预印本 arXiv:2410.16464, 2025. URLhttps://arxiv.org/abs/2410.16464.

[981] S. Srinivasa 和 Jayati Deshmukh. 计算代理的范式. 信息系统设计的新方法, 2021.

[982] B. Staresina, R. Henson, N. Kriegeskorte, 和 Arjen Alink. 海马体中的情景再现. 神经科学杂志, 2012.

[983] T. Staudigl, C. Vollmar, S. Noachtar, 和 S. Hanslmayr. 时间模式相似性分析揭示了情景再现对人类记忆的利弊. 神经科学杂志, 2015.

[984] Kaya Stechly, Karthik Valmeekam, 和 Subbarao Kambhampati. 思维链的无意识性? 对规划中cot的分析。 *Neural InformationProcessingSystems*, 2024。

[985] R. Sterken 和 James Ravi Kirkpatrick. 人工智能在语境中的对话对齐。 *PhilosophicalPerspectives*,2025。

[986] Paul Stoewer, Achim Schilling, Andreas K. Maier, 和 Patrick Krauss. 基于神经网络的、基于后继表示的多模态认知地图, arXiv 预印本 arXiv:2401.01364, 2023。URLhttps://arxiv.org/abs/2401.01364v1.

[987] Olly Styles, Sam Miller, Patricio Cerda-Mardini, T. Guha, Victor Sanchez, 和 Bertie Vidgen. 工作台：一个用于现实工作场所环境中代理的基准数据集, arXiv 预印本 arXiv:2405.00823, 2024。URLhttps://arxiv.org/abs/2405.00823v2.

[988] 苏广新, 朱一帆, 张文杰, 王汉辰, 张颖. 桥接大型语言模型和图结构学习模型以实现鲁棒表示学习, arXiv 预印本 arXiv:2410.12096, 2024. URLhttps://arxiv.org/abs/2410.12096v1.

[989] 苏红, Elke A. Rundensteiner, 和 Murali Mani. 自动机在或不在：用于 XML 流处理的运行时计划优化. 信号处理系统国际会议, 2008.

[990] 苏红金, 孙若曦, Yoon Jinsung, Yin Pengcheng, 余涛, 和 Sercan Ö. Arık. 交互式学习：面向真实环境中自适应智能体的数据驱动框架. 2025.

[991] 苏金岩, Healey Jennifer, Nakov Preslav, 和 Claire Cardie. 在欠思考与过度思考之间：对大型语言模型推理长度和正确性的实证研究, arXiv 预印本 arXiv:2505.00127, 2025. URLhttps://arxiv.org/abs/2505.00127v1.

[992] Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[993] Xin Su, Man Luo, Kris W Pan, Tien Pei Chou, Vasudev Lal, and Phillip Howard. Sk-vqa: Synthetic knowledge generation at scale for training context-augmented multimodal llms. arXiv preprint, 2024.

[994] Budhitama Subagdja and A. Tan. Neural modeling of sequential inferences and learning over episodic memory. *Neurocomputing*, 2015.

[995] Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, and Xia Hu. Stop overthinking: A survey on efficient reasoning for large language models, arXiv preprint arXiv:2503.16419, 2025. URL https://arxiv.org/abs/2503.16419v3.

[996] Chuanneng Sun, Songjun Huang, and D. Pompili. Llm-based multi-agent reinforcement learning: Current and future directions, arXiv preprint arXiv:2405.11106, 2024. URL https://arxiv.org/abs/2405.11106v1.

[997] Hanshi Sun, Zhuoming Chen, Xinyu Yang, Yuandong Tian, and Beidi Chen. Triforce: Lossless acceleration of long sequence generation with hierarchical speculative decoding, arXiv preprint arXiv:2404.11912, 2024. URL https://arxiv.org/abs/2404.11912v3.

[998] Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. Adaplanner: Adaptive planning from feedback with language models. *Neural Information Processing Systems*, 2023.

[999] Jiankai Sun, Chuanyang Zheng, E. Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Wu Yuan, Qi Liu, Xihui Liu, Yu Li, Hao Dong, Yu Cheng, Ming Zhang, P. Heng, Jifeng Dai, Ping Luo, Jingdong Wang, Jingwei Wen, Xipeng Qiu, Yi-Chen Guo, Hui Xiong, Qun Liu, and Zhenguo Li. A survey of reasoning with foundation models: Concepts, methodologies, and outlook. *ACM Computing Surveys*, 2023.

[1000] Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Sai Wang, Chen Lin, Yeyun Gong, H. Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph. arXiv preprint, 2023.

[1001] Lei Sun, Zhengwei Tao, Youdi Li, and Hiroshi Arakawa. Oda: Observation-driven agent for integrating llms and knowledge graphs, arXiv preprint arXiv:2404.07677, 2024. URL https://arxiv.org/abs/2404.07677.

[1002] Lei Sun, Xinchen Wang, and Youdi Li. Pyramid-driven alignment: Pyramid principle guided integration of large language models and knowledge graphs, arXiv preprint arXiv:2410.12298, 2024. URL https://arxiv.org/abs/2410.12298v2.

[1003] Liangtai Sun, Xingyu Chen, Lu Chen, Tianle Dai, Zichen Zhu, and Kai Yu. Meta-gui: Towards multi-modal conversational agents on mobile gui. *Conference on Empirical Methods in Natural Language Processing*, 2022.

[992] 苏伟航, 唐奕晨, 艾庆瑶, 吴志静, 和 刘一群. Dragin: 基于大型语言模型实时信息需求的动态检索增强生成. 计算语言学协会年度会议, 2024.

[993] 苏欣, 罗曼, 潘建伟, 蔡天培, Lal Vasudev, 和 Howard Phillip. Sk-vqa: 用于训练上下文增强多模态llm的大规模合成知识生成. arXiv预印本, 2024.

[994] Subagdja Budhitama 和 A. Tan. 序列推理和情景记忆学习的神经建模. 神经计算, 2015.

[995] 苏杨, 崔宇能, 王观初, 张嘉木, 张天毅, 袁佳怡, 刘鸿毅, 文安, 中少宸, 陈汉杰, 和 胡夏. 停止过度思考：大型语言模型高效推理的调查, arXiv预印本 arXiv:2503.16419, 2025. URL https://arxiv.org/abs/2503.16419v3.

[996] 孙传能, 黄松军, 和 D. Pompili. 基于大型语言模型的多智能体强化学习：现状与未来方向, arXiv 预印本 arXiv:2405.11106，2024。URL https://arxiv.org/abs/2405.11106v1.

[997] 孙汉石, 陈Zooming, 杨新宇, 田元东, 和 陈倍迪. Triforce：使用分层推测解码的无损加速长序列生成，arXiv 预印本 arXiv:2404.11912，2024。URLhttps://arxiv.org/abs/2404.11912v3.

[998] 孙浩天, 庄宇晨, 孔令凯, 戴博, 和 张超. Adaplanner：基于语言模型的反馈自适应规划。神经信息处理系统, 2023。

[999] 孙建凯, 郑传阳, E. Xie, 刘正英, 崔瑞航, 邱建宁, 许嘉琪, 丁明宇, 李红阳, 耿梦哲, 吴越, 王文海, 陈俊松, 尹张悦, 任晓哲, 傅杰, 何军献, 刘武, 刘奇, 刘希惠, 李宇, 董浩, 陈宇, 张明, P. Heng, 戴继峰, 罗平, 王景东, 温景伟, 邱锡鹏, 郭一晨, 邢辉, 刘群, 和 李正国. 基础模型推理的调查：概念、方法与展望。*ACM* 计算机调查, 2023。

[1000] Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Sai Wang, Chen Lin, Yeyun Gong, H. Shum, and Jian Guo. Think-on-graph: 基于知识图谱的大语言模型的深度与负责任推理. arXiv preprint, 2023.

[1001] Lei Sun, Zhengwei Tao, Youdi Li, and Hiroshi Arakawa. Oda: 驱动观察的智能体，用于整合大语言模型和知识图谱，arXiv preprint arXiv:2404.07677, 2024. URL https://arxiv.org/abs/2404.07677.

[1002] Lei Sun, Xinchen Wang, and Youdi Li. Pyramid-driven alignment: 基于金字塔原理的大语言模型与知识图谱的整合，arXiv preprint arXiv:2410.12298, 2024. URLhttps://arxiv.org/abs/2410.12298v2.

[1003] 梁太 Sun, 熙宇 Chen, 伦 Chen, 天乐 Dai, 智宸 Zhu, 和 凯 Yu. Meta-gui: Towards multi- modal conversational agents on mobile gui. *Conference on Empirical Methods in Natural LanguageProcessing*, 2022.

[1004] Lijun Sun, Yijun Yang, Qiqi Duan, Yuhui Shi, Chao Lyu, Yu-Cheng Chang, Chin-Teng Lin, and Yang Shen. Multi-agent coordination across diverse applications: A survey, arXiv preprint arXiv:2502.14743, 2025. URL https://arxiv.org/abs/2502.14743v2.

[1005] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and B. Schiele. Meta-transfer learning for few-shot learning. *Computer Vision and Pattern Recognition*, 2018.

[1006] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 2.0: A continual pre-training framework for language understanding. *AAAI Conference on Artificial Intelligence*, 2019.

[1007] Rao Surapaneni, Miku Jha, Michael Vakoc, and Todd Segal. Announcing the agent2agent protocol (a2a). https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/, April 2025. [Online; accessed 17-July-2025].

[1008] Stefan Szeider. Mcp-solver: Integrating language models with constraint programming systems. arXiv preprint, 2024.

[1009] Daniel Szelogowski. Engram memory encoding and retrieval: A neurocomputational perspective, arXiv preprint arXiv:2506.01659, 2025. URL https://arxiv.org/abs/2506.01659v1.

[1010] N. Taatgen, David Huss, D. Dickison, and John R. Anderson. The acquisition of robust and flexible cognitive skills. *Journal of experimental psychology. General*, 2008.

[1011] Jihoon Tack, Jaehyung Kim, Eric Mitchell, Jinwoo Shin, Yee Whye Teh, and Jonathan Richard Schwarz. Online adaptation of language models with a memory of amortized contexts. *Neural Information Processing Systems*, 2024.

[1012] Yan Tai, Weichen Fan, Zhao Zhang, Feng Zhu, Rui Zhao, and Ziwei Liu. Link-context learning for multimodal llms. *Computer Vision and Pattern Recognition*, 2023.

[1013] Fahim Tajwar, Yiding Jiang, Abitha Thankaraj, Sumaita Sadia Rahman, J. Z. Kolter, Jeff Schneider, and Ruslan Salakhutdinov. Training a generally curious agent, arXiv preprint arXiv:2502.17543, 2025. URL https://arxiv.org/abs/2502.17543v3.

[1014] K. Tallam. From autonomous agents to integrated systems, a new paradigm: Orchestrated distributed intelligence, arXiv preprint arXiv:2503.13754, 2025. URL https://arxiv.org/abs/2503.13754v2.

[1015] A. Tan, Budhitama Subagdja, Di Wang, and Lei Meng. Self-organizing neural networks for universal learning and multimodal memory encoding. *Neural Networks*, 2019.

[1016] Chuanyuan Tan, Yuehe Chen, Wenbiao Shao, and Wenliang Chen. Make a choice! knowledge base question answering with in-context learning, arXiv preprint arXiv:2305.13972, 2023. URL https://arxiv.org/abs/2305.13972v1.

[1017] Sijun Tan, Xiuyu Li, Shishir G. Patil, Ziyang Wu, Tianjun Zhang, Kurt Keutzer, Joseph E. Gonzalez, and Raluca A. Popa. Lloco: Learning long contexts offline. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[1004] 孙立军, 杨一军, 段启奇, 石宇辉, 吕超, 张宇成, 林金腾, 和沈阳. 跨越多样化应用的多智能体协调: 一项调查, arXiv 预印本 arXiv:2502.14743, 2025. URLhttps://arxiv.org/abs/2502.14743v2.

[1005] 孙千如, 刘瑶瑶, Chua Tat-Seng, 和 Schiele B. 元迁移学习用于小样本学习.计算机视觉与模式识别, 2018.

[1006] 孙宇, 王舒欢, 李雨坤, 冯石昆, 田浩, 吴华, 和王海峰. Ernie 2.0: 一种用于语言理解的持续预训练框架. *AAAI* 人工智能会议, 2019.

[1007]Surapaneni Rao, Jha Miku, Vakoc Michael, 和 Segal Todd. 宣布 agent2agent 协议 (a2a). https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/, 2025年4月. [在线; 访问于 17-7-2025].

[1008] Stefan Szeider. Mcp-solver: 将语言模型与约束编程系统集成. arXiv 预印本, 2024.

[1009] Daniel Szelogowski. Engram 记忆编码与检索：一种神经计算视角, arXiv preprint arXiv:2506.01659, 2025. URLhttps://arxiv.org/abs/2506.01659v1.

[1010] N. Taatgen, David Huss, D. Dickison, 和 John R. Anderson. 鲁棒和灵活认知技能的习得. 实验心理学杂志. 一般, 2008.

[1011] Jihoon Tack, Jaehyung Kim, Eric Mitchell, Jinwoo Shin, Yee Whye Teh, 和 Jonathan Richard Schwarz. 具有摊销上下文记忆的语言模型的在线适应. 神经信息处理系统, 2024.

[1012] Yan Tai, Weichen Fan, Zhao Zhang, Feng Zhu, Rui Zhao, 和 Ziwei Liu. 多模态 llms 的链接上下文学习. 计算机视觉与模式识别, 2023.

[1013] Fahim Tajwar, Yiding Jiang, Abitha Thankaraj, Sumaita Sadia Rahman, J. Z. Kolter, Jeff Schneider, 和 Ruslan Salakhutdinov. 训练一个普遍好奇的智能体, arXiv 预印本 arXiv:2502.17543, 2025. URLhttps://arxiv.org/abs/2502.17543v3.

[1014] K. Tallam. 从自主代理到集成系统，一种新范式：编排式分布式智能，arXiv 预印本 arXiv:2503.13754，2025。URLhttps://arxiv.org/abs/2503.13754v2.

[1015] A. Tan, Budhitama Subagdja, Di Wang, 和 Lei Meng. 自组织神经网络用于通用学习和多模态记忆编码. *NeuralNetworks*, 2019.

[1016] 陈传元, 陈月和, 邵文标, 和 陈文亮. Make a choice! 基于情境学习的知识库问答, arXiv 预印本 arXiv:2305.13972, 2023. URLhttps://arxiv.org/abs/2305.13972v1.

[1017] 谭思军, 李秀宇, Patil Shishir G., 吴子阳, 张天军, Keutzer Kurt, Gonzalez Joseph E., 和 Popa Raluca A. Lloco: 离线学习长上下文. 自然语言处理经验方法会议, 2024.

[1018] Xiaoyu Tan, Haoyu Wang, Xihe Qiu, Yuan Cheng, Yinghui Xu, Wei Chu, and Yuan Qi. Struct-x: Enhancing large language models reasoning with structured data. arXiv preprint, 2024.

[1019] Zhaoxuan Tan and Meng Jiang. User modeling in the era of large language models: Current research and future directions. *IEEE Data Engineering Bulletin*, 2023.

[1020] Zhijie Tan, Xu Chu, Weiping Li, and Tong Mo. Order matters: Exploring order sensitivity in multimodal large language models, arXiv preprint arXiv:2410.16983v1, 2024. URL https://arxiv.org/abs/2410.16983v1.

[1021] Matthew Tancik, Pratul P. Srinivasan, B. Mildenhall, Sara Fridovich-Keil, N. Raghavan, Utkarsh Singhal, R. Ramamoorthi, J. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Neural Information Processing Systems*, 2020.

[1022] Fei Tang, Haolei Xu, Hang Zhang, Siqi Chen, Xingyu Wu, Yongliang Shen, Wenqi Zhang, Guiyang Hou, Zeqi Tan, Yuchen Yan, Kaitao Song, Jian Shao, Weiming Lu, Jun Xiao, and Yueting Zhuang. A survey on (m)llm-based gui agents, arXiv preprint arXiv:2504.13865, 2025. URL https://arxiv.org/abs/2504.13865v2.

[1023] Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

[1024] Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 491–500, 2024.

[1025] Jianheng Tang, Qifan Zhang, Yuhan Li, Nuo Chen, and Jia Li. Grapharena: Evaluating and exploring large language models on graph computation. *arXiv preprint arXiv:2407.00379*, 2024.

[1026] Xiangru Tang, Yiming Zong, Jason Phang, Yilun Zhao, Wangchunshu Zhou, Arman Cohan, and Mark Gerstein. Struc-bench: Are large language models good at generating complex structured tabular data? *North American Chapter of the Association for Computational Linguistics*, 2024.

[1027] Xiangru Tang, Tianyu Hu, Muyang Ye, Yanjun Shao, Xunjian Yin, Siru Ouyang, Wangchunshu Zhou, Pan Lu, Zhuosheng Zhang, Yilun Zhao, Arman Cohan, and Mark Gerstein. Chemagent: Self-updating library in large language models improves chemical reasoning, arXiv preprint arXiv:2501.06590, 2025. URL https://arxiv.org/abs/2501.06590v1.

[1028] Xuemei Tang, Jun Wang, and Q. Su. Chinese word segmentation with heterogeneous graph neural network, arXiv preprint arXiv:2201.08975, 2022. URL https://arxiv.org/abs/2201.08975v1.

[1029] Yiqing Tang, Xingyuan Dai, Chengchong Zhao, Qi Cheng, and Yisheng Lv. Large language model-driven urban traffic signal control. *Australian and New Zealand Control Conference*, 2024.

[1030] Yongjian Tang, Rakebul Hasan, and Thomas Runkler. Fsponer: Few-shot prompt optimization for named entity recognition in domain-specific scenarios. *European Conference on Artificial Intelligence*, 2024.

[1031] Yunlong Tang, Daiki Shimada, Jing Bi, Hang Hua, and Chenliang Xu. Empowering llms with pseudo-untrimmed videos for audio-visual temporal understanding. *AAAI Conference on Artificial Intelligence*, 2024.

[1018] Tan Xiaoyu, Wang Haoyu, Qiu Xihe, Cheng Yuan, Xu Yinghui, Chu Wei, and Qi Yuan. Struct-x: 使用结构化数据增强大型语言模型的推理能力. arXiv preprint, 2024.

[1019] Tan Zhaoxuan and Jiang Meng. 大型语言模型时代下的用户建模：当前研究及未来方向. *IEEE* 数据工程简报, 2023.

[1020] Tan Zhijie, Chu Xu, Li Weiping, and Mo Tong. 顺序很重要：探索多模态大型语言模型中的顺序敏感性, arXiv preprint arXiv:2410.16983v1, 2024. URL https://arxiv.org/abs/2410.16983v1.

[1021] Tancik Matthew, Srinivasan Pratul P., Mildenhall B., Fridovich-Keil Sara, Raghavan N., Singhal Utkarsh, Ramamoorthi R., Barron J., and Ng Ren. 傅里叶特征让网络在低维域学习高频函数. 神经信息处理系统, 2020.

[1022] 飞堂，浩雷徐，张航，陈思琪，吴兴宇，沈永亮，张文琪，侯贵阳，谭泽奇，闫宇辰，宋凯涛，邵健，陆为民，肖俊，庄宇婷. 基于大型语言模型的图形用户界面代理综述，arXiv预印本 arXiv:2504.13865，2025. URLhttps://arxiv.org/abs/2504.13865v2.

唐嘉斌, 杨宇豪, 魏伟, 石磊, 苏立新, 成苏琪, 尹大伟, 和黄超. GraphGPT: 针对大型语言模型的图指令微调. 国际信息检索研究与发展年度会议 *(ACM SIGIR)*, 2023.

唐嘉斌, 杨宇豪, 魏伟, 石磊, 苏立新, 成苏琪, 尹大伟, 和黄超. Graphgpt: 针对大型语言模型的图指令微调. 在《第 47 届国际 ACM SIGIR 信息检索研究与发展会议论文集》中 , 第 491-500 页 , 2024.

[1025] Jianheng Tang, Qifan Zhang, Yuhan Li, Nuo Chen, 和 Jia Li. Grapharena: 在图计算上评估和探索大型语言模型. *arXivpreprint arXiv:2407.00379*, 2024.

[1026] Xiangru Tang, Yiming Zong, Jason Phang, Yilun Zhao, Wangchunshu Zhou, Arman Cohan, 和 Mark Gerstein. Struc-bench: 大型语言模型在生成复杂结构化表格数据方面表现如何? *North American Chapter of theAssociationfor Computational Linguistics*, 2024.

[1027] Xiangru Tang, Tianyu Hu, Muyang Ye, Yanjun Shao, Xunjian Yin, Siru Ouyang, Wangchunshu Zhou, Pan Lu, Zhuosheng Zhang, Yilun Zhao, Arman Cohan, 和 MarkGerstein. Chemagent: 大型语言模型中的自更新库提高了化学推理能力, arXiv preprint arXiv:2501.06590, 2025. URLhttps://arxiv.org/abs/2501.06590v1.

[1028] 徐美唐，王军，和 Q. Su. 基于异构图神经网络的中文分词，arXiv 预印本 arXiv:2201.08975，2022。URL https://arxiv.org/abs/2201.08975v1.

[1029] 唐一青，戴兴元，赵成冲，程琪，和吕奕升。基于大型语言模型的交通信号控制。 澳大利亚和新西兰控制会议, 2024。

[1030] 唐永健，哈桑·拉克布尔，和托马斯·伦克勒。Fsponer：特定领域场景中命名实体识别的少样本提示优化。 欧洲人工智能会议, 2024。

[1031] 唐云龙，岛崎大记，毕静，华航，和徐晨亮。使用伪未修剪视频为大型语言模型赋能以实现视听时间理解。 *AAAI* 人工智能会议, 2024。

[1032] Yao Tao, Yehui Tang, Yun Wang, Mingjian Zhu, Hailin Hu, and Yunhe Wang. Saliency-driven dynamic token pruning for large language models, arXiv preprint arXiv:2504.04514, 2025. URL https://arxiv.org/abs/2504.04514v2.

[1033] Denis Tarasov and Kumar Shridhar. Distilling llms' decomposition abilities into compact language models, arXiv preprint arXiv:2402.01812, 2024. URL https://arxiv.org/abs/2402.01812v1.

[1034] Pittawat Taveekitworachai, Potsawee Manakul, Kasima Tharnpipitchai, and Kunat Pipatanakul. Typhoon t1: An open thai reasoning model, arXiv preprint arXiv:2502.09042, 2025. URL https://arxiv.org/abs/2502.09042v2.

[1035] Yi Tay, Anh Tuan Luu, Minh C. Phan, and S. Hui. Multi-task neural network for non-discrete attribute prediction in knowledge graphs. *International Conference on Information and Knowledge Management*, 2017.

[1036] 36Kr Editorial Team. The future of ai: From parameter scaling to context scaling. Online, 2025. URL https://36kr.com/p/3337269379328264. Chinese business and technology media publication discussing context scaling in large language models.

[1037] Junfeng Tian, Da Zheng, Yang Cheng, Rui Wang, Colin Zhang, and Debing Zhang. Untie the knots: An efficient data augmentation strategy for long-context pre-training in language models, arXiv preprint arXiv:2409.04774, 2024. URL https://arxiv.org/abs/2409.04774v1.

[1038] S Tian, R Wang, H Guo, P Wu, Y Dong, and X Wang.... Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. 2025. URL https://arxiv.org/abs/2506.13654.

[1039] Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. arXiv preprint, 2025.

[1040] Ramine Tinati, Xin Wang, Ian C. Brown, T. Tiropanis, and W. Hall. A streaming real-time web observatory architecture for monitoring the health of social machines. *The Web Conference*, 2015.

[1041] Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Neural Information Processing Systems*, 2022.

[1042] Despina Tomkou, George Fatouros, Andreas Andreou, Georgios Makridis, F. Liarokapis, Dimitrios Dardanis, Athanasios Kiourtis, John Soldatos, and D. Kyriazis. Bridging industrial expertise and xr with llm-powered conversational agents, arXiv preprint arXiv:2504.05527, 2025. URL https://arxiv.org/abs/2504.05527v1.

[1043] Sabrina Toro, A. V. Anagnostopoulos, Sue Bello, Kai Blumberg, Rhiannon Cameron, Leigh Carmody, A. Diehl, Damion M. Dooley, William Duncan, P. Fey, Pascale Gaudet, Nomi L. Harris, marcin p. joachimiak, Leila Kiani, Tiago Lubiana, M. Munoz-Torres, Shawn T. O'Neil, David Osumi-Sutherland, Aleix Puig, Justin Reese, L. Reiser, Sofia M C Robb, Troy Ruemping, James Seager, Eric Sid, Ray Stefancsik, Magalie Weber, Valerie Wood, M. Haendel, and Christopher J. Mungall. Dynamic retrieval augmented generation of ontologies using artificial intelligence (dragon-ai). *Journal of Biomedical Semantics*, 2023.

[1032] Yao Tao, Yehui Tang, Yun Wang, Mingjian Zhu, Hailin Hu, and Yunhe Wang. 基于显著性的动态标记剪枝用于大型语言模型，arXiv 预印本 arXiv:2504.04514，2025。URLhttps://arxiv.org/abs/2504.04514v2.

[1033] Denis Tarasov and Kumar Shridhar. 将大型语言模型的分解能力蒸馏到紧凑语言模型中，arXiv 预印本 arXiv:2402.01812，2024。URLhttps://arxiv.org/abs/2402.01812v1.

[1034] Pittawat Taveekitworachai, Potsawee Manakul, Kasima Tharnpipitchai, and Kunat Pipatanakul. 台风 t1：一个开放式的泰语推理模型，arXiv 预印本 arXiv:2502.09042，2025。URL https://arxiv.org/abs/2502.09042v2.

[1035] Yi Tay, Anh Tuan Luu, Minh C. Phan, 和 S. Hui. 用于知识图谱中非离散属性预测的多任务神经网络。 国际信息与知识管理会议, 2017。

[1036] 36Kr 编辑团队. 人工智能的未来：从参数扩展到上下文扩展。 在线, 2025. URL https://36kr.com/p/3337269379328264。 一家中文商业和技术媒体，讨论大型语言模型中的上下文扩展。

[1037] Tian Junfeng, Zheng Da, Cheng Yang, 王瑞, 张 Colin, 和 张德兵. 解开结：语言模型长上下文预训练的高效数据增强策略，arXiv 预印本 arXiv:2409.04774, 2024. URLhttps://arxiv.org/abs/2409.04774v1.

[1038] Tian S, Wang R, Guo H, Wu P, Dong Y, 和 王X. . . . Ego-r1：用于超长自我中心视频推理的工具链思维链。2025. URLhttps://arxiv.org/abs/2506.13654.

[1039] Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: 链式思维工具用于超长第一人称视频推理. arXiv preprint, 2025.

[1040] Ramine Tinati, Xin Wang, Ian C. Brown, T. Tiropanis, and W. Hall. 流式实时网络观测架构用于监控社交机器的健康状况. *The WebConference*, 2015.

[1041] Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 无过拟合的记忆：分析大型语言模型的训练动态. *Neural InformationProcessing Systems*, 2022.

[1042] Despina Tomkou, George Fatouros, Andreas Andreou, Georgios Makridis, F. Liarokapis, Dimitrios Dardanis, Athanasios Kiourtis, John Soldatos, and D. Kyriazis. 连接工业专长和xr与llm驱动的对话代理, arXiv preprint arXiv:2504.05527, 2025. URLhttps://arxiv.org/abs/2504.05527v1.

[1043] Sabrina Toro, A. V. Anagnostopoulos, Sue Bello, Kai Blumberg, Rhiannon Cameron, Leigh Carmody, A. Diehl, Damion M. Dooley, William Duncan, P. Fey, Pascale Gaudet, Nomi L. Harris, marcin p. joachimiak, Leila Kiani, Tiago Lubiana, M. Munoz-Torres, Shawn T. O' Neil, David Osumi-Sutherland, Aleix Puig, Justin Reese, L. Reiser, Sofia M C Robb, Troy Ruemping, James Seager, Eric Sid, Ray Stefancsik, Magalie Weber, Valerie Wood, M. Haendel, 和Christopher J. Mungall. 使用人工智能（dragon-ai）动态检索增强本体生成. *Journal of BiomedicalSemantics*, 2023.

[1044] Fernanda M De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. Llmr: Real-time prompting of interactive worlds using large language models. *International Conference on Human Factors in Computing Systems*, 2023.

[1045] Martina Toshevska and Sonja Gievska. Llm-based text style transfer: Have we taken a step forward? *IEEE Access*, 2025.

[1046] Fouad Trad and Ali Chehab. Evaluating the efficacy of prompt-engineered large multimodal models versus fine-tuned vision transformers in image-based security applications. *ACM Transactions on Intelligent Systems and Technology*, 2024.

[1047] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. Multi-agent collaboration mechanisms: A survey of llms, arXiv preprint arXiv:2501.06322, 2025. URL https://arxiv.org/abs/2501.06322v1.

[1048] Harold Triedman, Rishi Jha, and Vitaly Shmatikov. Multi-agent systems execute arbitrary malicious code, arXiv preprint arXiv:2503.12188, 2025. URL https://arxiv.org/abs/2503.12188v1.

[1049] H. Trivedi, Tushar Khot, Mareike Hartmann, R. Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[1050] Yun-Da Tsai, Ting-Yu Yen, Pei-Fu Guo, Zhe-Yan Li, and Shou-De Lin. Text-centric alignment for multi-modality learning, arXiv preprint arXiv:2402.08086v2, 2024. URL https://arxiv.org/abs/2402.08086v2.

[1051] Tao Tu, M. Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomašev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, A. Karthikesalingam, and Vivek Natarajan. Towards conversational diagnostic artificial intelligence. *Nature*, 2025.

[1052] Eduard Tulchinskii, Laida Kushnareva, Kristian Kuznetsov, Anastasia Voznyuk, Andrei Andriiainen, Irina Piontkovskaya, Evgeny Burnaev, and Serguei Barannikov. Listening to the wise few: Select-and-copy attention heads for multiple-choice qa, arXiv preprint arXiv:2410.02343, 2024. URL https://arxiv.org/abs/2410.02343v1.

[1053] Meet Udeshi, Minghao Shao, Haoran Xi, Nanda Rani, Kimberly Milner, Venkata Sai Charan Putrevu, Brendan Dolan-Gavitt, S. K. Shukla, P. Krishnamurthy, F. Khorrami, Ramesh Karri, and Muham- mad Shafique. D-cipher: Dynamic collaborative intelligent multi-agent system with planner and heterogeneous executors for offensive security. arXiv preprint, 2025.

[1054] M. Ursino, Nicole Cesaretti, and G. Pirazzini. A model of working memory for encoding multiple items and ordered sequences exploiting the theta-gamma code. *Cognitive Neurodynamics*, 2022.

[1055] D. H. V. Uytsel, Filip Van Aelten, and Dirk Van Compernolle. A structured language model based on context-sensitive probabilistic left-corner parsing. *North American Chapter of the Association for Computational Linguistics*, 2001.

[1044] Fernanda M De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. Llmr: 使用大型语言模型实时提示交互式世界。 国际人机交互会议, 2023。

[1045] Martina Toshevska and Sonja Gievska. 基于LLM的文本风格迁移：我们是否迈出了进步的一步？*IEEE Access*, 2025。

[1046] Fouad Trad and Ali Chehab. 评估提示工程的大型多模态模型与微调视觉变换器在基于图像的安全应用中的有效性。 *ACMTransactions onIntelligent Systemsand Technology*, 2024。

[1047] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O' Sullivan, and Hoang D. Nguyen. 多智能体协作机制：对LLM的调查，arXiv预印本 arXiv:2501.06322，2025。
URLhttps://arxiv.org/abs/2501.06322v1.

[1048] Harold Triedman, Rishi Jha, and Vitaly Shmatikov. Multi-agent systems execute arbitrary malicious code, arXiv preprint arXiv:2503.12188, 2025. URL https://arxiv.org/abs/2503.12188v1.

[1049] H. Trivedi, Tushar Khot, Mareike Hartmann, R. Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. Appworld: 一个用于测试交互式编码代理的应用和人可控世界. 计算语言学协会年度会议, 2024.

[1050] Yun-Da Tsai, Ting-Yu Yen, Pei-Fu Guo, Zhe-Yan Li, and Shou-De Lin. 面向多模态学习的文本中心对齐, arXiv 预印本 arXiv:2402.08086v2, 2024. URL https://arxiv.org/abs/2402.08086v2.

[1051] Tao Tu, M. Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomašev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, A. Karthikesalingam, and Vivek Natarajan. 趋向对话式诊断人工智能. 自然, 2025.

[1052] Eduard Tulchinskii, Laida Kushnareva, Kristian Kuznetsov, Anastasia Voznyuk, Andrei Andriiainen, Irina Piontkovskaya, Evgeny Burnaev, and Serguei Barannikov. 听从智者的声音：用于多项选择题问答的选中和复制注意力头，arXiv 预印本 arXiv:2410.02343, 2024. URLhttps://arxiv.org/abs/2410.02343v1.

[1053] Meet Udeshi, Minghao Shao, Haoran Xi, Nanda Rani, Kimberly Milner, Venkata Sai Charan Putrevu, Brendan Dolan-Gavitt, S. K. Shukla, P. Krishnamurthy, F. Khorrami, Ramesh Karri, and Muham- mad Shafique. D-cipher：具有规划和异构执行器的动态协作智能多智能体系统，用于进攻性安全。arXiv 预印本，2025。

[1054] M. Ursino, Nicole Cesaretti, and G. Pirazzini. 一种用于编码多个项目和有序序列的工作记忆模型，利用了theta-gamma代码。 <style id='1'>认知神经动力学</style>, 2022。

A Survey of Context Engineering for Large Language Models

[1055] D. H. V. Uytsel, Filip Van Aelten, and Dirk Van Compernolle. 一种基于上下文敏感概率左角解析的结构化语言模型。 美国计算语言学协会北美分会, 2001。

[1056] Saeid Ario Vaghefi, Aymane Hachcham, Veronica Grasso, Jiska Manicus, Nakiete Msemo, C. Senni, and Markus Leippold. Ai for climate finance: Agentic retrieval and multi-step reasoning for early warning system investments, arXiv preprint arXiv:2504.05104, 2025. URL https://arxiv.org/abs/2504.05104v2.

[1057] Priyan Vaithilingam, Tianyi Zhang, and Elena L. Glassman. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. *CHI Extended Abstracts*, 2022.

[1058] Phuc Phan Van, Dat Nguyen Minh, An Dinh Ngoc, and Huy-Phan Thanh. Rx strategist: Prescription verification using llm agents system, arXiv preprint arXiv:2409.03440, 2024. URL https://arxiv.org/abs/2409.03440v1.

[1059] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and I. Polosukhin. Attention is all you need. *Neural Information Processing Systems*, 2017.

[1060] J. D. Velásquez-Henao, Carlos Jaime Franco-Cardona, and Lorena Cadavid-Higuita. Prompt engineering: a methodology for optimizing interactions with ai-language models in the field of engineering. *DYNA*, 2023.

[1061] Gaurav Verma, Rachneet Kaur, Nishan Srishankar, Zhen Zeng, T. Balch, and Manuela Veloso. Adaptagent: Adapting multimodal web agents with few-shot learning from human demonstrations, arXiv preprint arXiv:2411.13451, 2024. URL https://arxiv.org/abs/2411.13451v1.

[1062] Aliaksei Vertsel and Mikhail Rumiantsau. Hybrid llm/rule-based approaches to business insights generation from structured data, arXiv preprint arXiv:2404.15604, 2024. URL https://arxiv.org/abs/2404.15604v1.

[1063] Aishwarya Vijayan. A prompt engineering approach for structured data extraction from unstructured text using conversational llms. *International Conference on Advances in Computing and Artificial Intelligence*, 2023.

[1064] Juraj Vladika, Alexander Fichtl, and Florian Matthes. Diversifying knowledge enhancement of biomedical language models using adapter modules and knowledge graphs. *International Conference on Agents and Artificial Intelligence*, 2023.

[1065] James Vo. Sparseaccelerate: Efficient long-context inference for mid-range gpus, arXiv preprint arXiv:2412.06198, 2024. URL https://arxiv.org/abs/2412.06198v1.

[1066] Blavz vSkrlj, Boshko Koloski, S. Pollak, and Nada Lavravc. From symbolic to neural and back: Exploring knowledge graph-large language model synergies. arXiv preprint, 2025.

[1067] Tom Völker, Jan Pfister, Tobias Koopmann, and Andreas Hotho. From chat to publication management: Organizing your related work using bibsonomy & llms. *Conference on Human Information Interaction and Retrieval*, 2024.

[1068] D. Walton. Using argumentation schemes to find motives and intentions of a rational agent. *Argument Comput.*, 2020.

[1069] Hanlong Wan, Jian Zhang, Yan Chen, Weili Xu, and Fan Feng. Generative ai application for building industry. *Building Simulation*, 2024.

[1056] Saeid Ario Vaghefi、Aymane Hachcham、Veronica Grasso、Jiska Manicus、Nakiete Msemo、C. Senni 和 Markus Leippold。人工智能在气候金融中的应用：用于早期预警系统投资的自主检索和多步推理，arXiv 预印本 arXiv:2504.05104，2025年。URL https://arxiv.org/abs/2504.05104v2。

[1057] Priyan Vaithilingam、Tianyi Zhang 和 Elena L. Glassman。期望与体验：评估由大型语言模型驱动的代码生成工具的可用性。*CHI Extended Abstracts*，2022。

[1058] Phuc Phan Van, Dat Nguyen Minh, An Dinh Ngoc, and Huy-Phan Thanh. Rx strategist: Prescription verification using llm agents system, arXiv preprint arXiv:2409.03440, 2024. URL https://arxiv.org/abs/2409.03440v1.

[1059] Ashish Vaswani、Noam M. Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser和I. Polosukhin。Attention is all you need。*Neural Information ProcessingSystems*，2017。

[1060] J. D. Velásquez-Henao、Carlos JaimeFranco-Cardona和Lorena Cadavid-Higuita。提示工程：一种优化工程领域中与人工智能语言模型交互的方法。*DYNA*，2023.

[1061] Gaurav Verma，Rachneet Kaur，Nishan Srishankar，Zhen Zeng，T. Balch 和 Manuela Veloso。Adaptagent：基于人类演示的少样本学习适应多模态网络代理，arXiv 预印本 arXiv:2411.13451，2024。URLhttps://arxiv.org/abs/2411.13451v1。

[1062] Aliaksei Vertsel and Mikhail Rumiantsau. 基于混合 llm/规则的方法从结构化数据生成商业洞察，arXiv 预印本 arXiv:2404.15604，2024。URLhttps://arxiv.org/abs/2404.15604v1.

[1063] Aishwarya Vijayan. 一种用于从非结构化文本中提取结构化数据的提示工程方法，使用对话式 llms。 国际会议先进计算与人工智能，2023。

[1064] Juraj Vladika, Alexander Fichtl, and Florian Matthes. 使用适配器模块和知识图谱多样化生物医学语言模型的知识增强。 国际会议智能体与人工智能，2023。

[1065] James Vo. Sparseaccelerate：为中端显卡的高效长上下文推理，arXiv 预印本 arXiv:2412.06198，2024。URLhttps://arxiv.org/abs/2412.06198v1.

[1066] Blavz vSkrlj, Boshko Koloski, S. Pollak, 和 Nada Lavravc. 从符号到神经再回到符号：探索知识图谱-大型语言模型的协同作用。arXiv 预印本，2025。

[1067] Tom Völker, JanPfister, Tobias Koopmann, 和 Andreas Hotho. 从聊天到论文管理：使用 bibsonomy & llms 组织你的相关工作。人类信息交互与检索会议，2024。

[1068] D. Walton. 使用论证模式来发现理性代理的动机和意图。论证计算，2020。

[1069] Hanlong Wan, Jian Zhang, Yan Chen, Weili Xu, 和 Fan Feng. 建筑行业的生成式人工智能应用。建筑模拟，2024。

[1070] Jun Wan and Lingrui Mei. Large language models as computable approximations to solomonoff induction, arXiv preprint arXiv:2505.15784, 2025. URL https://arxiv.org/abs/2505.15784.

[1071] Luanbo Wan and Weizhi Ma. Storybench: A dynamic benchmark for evaluating long-term memory with multi turns, arXiv preprint arXiv:2506.13356, 2025. URL https://arxiv.org/abs/2506.13356v1.

[1072] Bernie Wang, Si ting Xu, K. Keutzer, Yang Gao, and Bichen Wu. Improving context-based meta-reinforcement learning with self-supervised trajectory contrastive learning, arXiv preprint arXiv:2103.06386, 2021. URL https://arxiv.org/abs/2103.06386v1.

[1073] Bing Wang, Xinnian Liang, Jian Yang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. Scm: Enhancing large language model with self-controlled memory framework, arXiv preprint arXiv:2304.13343, 2025. URL https://arxiv.org/abs/2304.13343.

[1074] Cangqing Wang, Yutian Yang, Ruisi Li, Dan Sun, Ruicong Cai, Yuzhu Zhang, and Chengqian Fu. Adapting llms for efficient context processing through soft prompt compression. *Proceedings of the International Conference on Modeling, Natural Language Processing and Machine Learning*, 2024.

[1075] Chaozheng Wang, Yuanhang Yang, Cuiyun Gao, Yun Peng, Hongyu Zhang, and Michael R. Lyu. Prompt tuning in code intelligence: An experimental evaluation. *IEEE Transactions on Software Engineering*, 2023.

[1076] Dongsheng Wang, Zhiqiang Ma, Armineh Nourbakhsh, Kang Gu, and Sameena Shah. Docgraphlm: Documental graph language model for information extraction. *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

[1077] Fan Wang, Chuan Lin, Yang Cao, and Yu Kang. Benchmarking general purpose in-context learning, arXiv preprint arXiv:2405.17234, 2024. URL https://arxiv.org/abs/2405.17234v6.

[1078] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, arXiv preprint arXiv:2305.16291, 2023. URL https://arxiv.org/abs/2305.16291.

[1079] Guoqing Wang, Zeyu Sun, Zhihao Gong, Sixiang Ye, Yizhou Chen, Yifan Zhao, Qing-Lin Liang, and Dan Hao. Do advanced language models eliminate the need for prompt engineering in software engineering?, arXiv preprint arXiv:2411.02093, 2024. URL https://arxiv.org/abs/2411.02093v1.

[1080] Hanlin Wang, Zhan Tong, Kecheng Zheng, Yujun Shen, and Limin Wang. Contextual ad narration with interleaved multimodal sequence. *Computer Vision and Pattern Recognition*, 2024.

[1081] Hanrui Wang, Zhekai Zhang, and Song Han. Spatten: Efficient sparse attention architecture with cascade token and head pruning. *International Symposium on High-Performance Computer Architecture*, 2020.

[1082] Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. *arXiv preprint arXiv:2507.07999*, 2025.

[1070] 孙婉和梅凌瑞. 大型语言模型作为可计算的索洛蒙夫归纳的近似，arXiv预印本arXiv:2505.15784, 2025. URLhttps://arxiv.org/abs/2505.15784.

[1071] 万卯波和马伟智. Storybench: 一个用于评估多轮长期记忆的动态基准，arXiv预印本arXiv:2506.13356, 2025. URL https://arxiv.org/abs/2506.13356v1.

[1072] 王伯尼，徐思婷，K. Keutzer，高杨，和吴必辰. 基于上下文的元强化学习改进与自监督轨迹对比学习，arXiv预印本arXiv:2103.06386, 2021. URLhttps://arxiv.org/abs/2103.06386v1.

[1073] 王冰，梁欣年，杨健，黄辉，吴双知，吴培豪，卢路，马泽军，和李周军。Scm: 基于自控记忆框架增强大语言模型，arXiv预印本arXiv:2304.13343，2025年。URLhttps://arxiv.org/abs/2304.13343。

[1074] 王苍庆、杨宇天、李瑞思、孙丹、蔡瑞丛、张宇珠和付成倩。通过软提示压缩使LLM适应高效上下文处理。《国际建模、自然语言处理和机器学习会议论文集》，2024。

王超正, 杨元航, 高翠云, 彭云, 张红宇, 和 Michael R. Lyu. 代码智能中的提示微调：实验评估. *IEEE Transactions on SoftwareEngineering*, 2023.

[1076] 王冬生, 马志强, Nourbakhsh Armineh, 古康, 和 Shah Sameena. Docgraphlm: 用于信息抽取的文档图语言模型. 年度国际 *ACM SIGIR* 信息检索研究与发展会议, 2023.

[1077] 王帆, 林川, 曹杨, 和 康宇. 通用情境学习的基准测试, arXiv 预印本 arXiv:2405.17234, 2024. URLhttps://arxiv.org/abs/2405.17234v6.

[1078] 王观志, 谢宇琪, 姜云帆, Mandlekar Ajay, 小肖超伟, 朱玉克, 范林曦, 和 Anandkumar Anima. Voyager: 基于大型语言模型的开放式具身智能体, arXiv 预印本 arXiv:2305.16291, 2023. URLhttps://arxiv.org/abs/2305.16291.

[1079] 郭庆王, 孙泽宇, 龚志浩, 叶思祥, 陈一舟, 赵一帆, 梁清林, 和单浩. 先进语言模型是否消除了软件工程中提示工程的需求?, arXiv preprint arXiv:2411.02093, 2024. URL https://arxiv.org/abs/2411.02093v1.

[1080] 王汉林, 戚通, 郑克成, 沈宇君, 和王黎明. 上下文广告叙述与交错多模态序列. 计算机视觉与模式识别, 2024.

[1081] 王汉瑞, 张哲凯, 和韩松. Spatten: 带级联标记和头剪枝的高效稀疏注意力架构. 高性能计算机架构国际会议, 2020.

[1082] 王浩辰, 李向台, 黄子龙, 王安然, 王家聪, 张涛, 郑佳妮, 白苏蕾, 康子健, 冯嘉石, 等. 可追溯证据增强视觉推理: 评估与方法. *arXivpreprint arXiv:2507.07999*, 2025.

[1083] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning llama model with chinese medical knowledge, arXiv preprint arXiv:2304.06975, 2023. URL https://arxiv.org/abs/2304.06975.

[1084] Haoyu Wang, Tong Teng, Tianyu Guo, An Xiao, Duyu Tang, Hanting Chen, and Yunhe Wang. Unshackling context length: An efficient selective attention approach through query-key compression, arXiv preprint arXiv:2502.14477, 2025. URL https://arxiv.org/abs/2502.14477v1.

[1085] Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. Can language models solve graph problems in natural language? *Neural Information Processing Systems*, 2023.

[1086] Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, A. Nambi, T. Ganu, and Hao Wang. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. *North American Chapter of the Association for Computational Linguistics*, 2024.

[1087] Hongru Wang, Cheng Qian, Manling Li, Jiahao Qiu, Boyang Xue, Mengdi Wang, Heng Ji, and Kam-Fai Wong. Toward a theory of agents as tool-use decision-makers, arXiv preprint arXiv:2506.00886, 2025. URL https://arxiv.org/abs/2506.00886v1.

[1088] Jingjin Wang. Proprag: Guiding retrieval with beam search over proposition paths, arXiv preprint arXiv:2504.18070, 2025. URL https://arxiv.org/abs/2504.18070v1.

[1089] Jingyu Wang, Lu Zhang, Xueqing Li, Huazhong Yang, and Yongpan Liu. Ulseq-ta: Ultra-long sequence attention fusion transformer accelerator supporting grouped sparse softmax and dual-path sparse layernorm. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2024.

[1090] Jize Wang, Zerun Ma, Yining Li, Songyang Zhang, Cailian Chen, Kai Chen, and Xinyi Le. Gta: A benchmark for general tool agents. *Neural Information Processing Systems*, 2024.

[1091] Lei Wang, Chengbang Ma, Xueyang Feng, Zeyu Zhang, Hao ran Yang, Jingsen Zhang, Zhi-Yang Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji rong Wen. A survey on large language model based autonomous agents. *Frontiers Comput. Sci.*, 2023.

[1092] Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *Annual Meeting of the Association for Computational Linguistics*, 2023.

[1093] Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, et al. When large language model based agent meets user behavior analysis: A novel user simulation paradigm. 2023.

[1094] Libo Wang. Towards humanoid robot autonomy: A dynamic architecture integrating continuous thought machines (ctm) and model context protocol (mcp), arXiv preprint arXiv:2505.19339, 2025. URL https://arxiv.org/abs/2505.19339v1.

[1095] Liya Wang, Jason Chou, Xin Zhou, A. Tien, and Diane M. Baumgartner. Aviationgpt: A large language model for the aviation domain, arXiv preprint arXiv:2311.17686, 2023. URL https://arxiv.org/abs/2311.17686v1.

[1083] 王浩春, 刘驰, 西女, 强泽文, 赵 Sendong, 秦 Bing, 和 刘婷. 华佗: 使用中医知识调整 llama 模型, arXiv preprint arXiv:2304.06975, 2023. URLhttps://arxiv.org/abs/2304.06975.

[1084] 王浩宇, 奔腾, 国天宇, 肖安, 唐都宇, 陈汉庭, 和 王云和. 解放上下文长度: 通过查询键压缩实现高效的选择性注意力方法, arXiv preprint arXiv:2502.14477, 2025. URLhttps://arxiv.org/abs/2502.14477v1.

[1085] 王恒, 冯上斌, 何天行, 谭赵轩, 韩晓川, 和 切夫托娃 Yulia. 语言模型能否用自然语言解决图问题? 神经信息处理系统, 2023.

[1086] 王恒毅, 石海舟, 谭石伟, 秦伟毅, 王文远, 张屯宇, A. Nambi, T. Ganu, 和 王浩. 多模态大海捞针: 测试多模态大语言模型的长期上下文能力. 北美学术计算语言学协会, 2024.

[1087] 王红如, 钱成, 李曼玲, 邱嘉豪, 薛博阳, 王梦迪, 季恒, 和 邓肯飞. 朝向智能体作为工具使用决策者的理论, arXiv 预印本 arXiv:2506.00886, 2025. URLhttps://arxiv.org/abs/2506.00886v1.

[1088] 王景进. Proprag: 通过命题路径上的波束搜索指导检索, arXiv 预印本 arXiv:2504.18070, 2025. URLhttps://arxiv.org/abs/2504.18070v1.

王景宇, 张路, 李雪青, 杨华中, 刘永攀. Ulseq-ta: 支持分组稀疏softmax和双路径稀疏层归一化的超长序列注意力融合Transformer加速器. *IEEETransactionsonComputer-Aided Design of Integrated Circuits and Systems*, 2024.

[1090] 王继哲, 马泽润, 李一宁, 张宋阳, 陈才良, 陈凯, 和雷新怡. Gta: 通用工具代理的基准测试. 神经信息处理系统, 2024.

王蕾, 马成邦, 冯雪阳, 张泽宇, 杨浩然, 张景森, 陈志阳, 唐嘉凯, 陈旭, 林彦开, 赵文伟, 魏哲伟, 温继荣. 基于大型语言模型的自主代理综述. *FrontiersComput. Sci.*, 2023.

[1092] 王蕾, 徐婉宇, 兰亦怀, 胡志强, 兰云石, 李国威, 以及林伊平。计划和解决提示: 通过大型语言模型改进零样本思维链推理。计算语言学协会年会, 2023.

[1093] 雷王, 张景森, 杨浩, 陈志远, 唐嘉凯, 张泽宇, 陈旭, 林彦凯, 宋瑞华, 赵文新, 等. 当大型语言模型代理遇见用户行为分析: 一种新型用户模拟范式. 2023.

[1094] 王丽波. 迈向人形机器人自主性: 集成持续思维机器 (ctm) 和模型上下文协议 (mcp) 的动态架构, arXiv 预印本 arXiv:2505.19339, 2025. URL https://arxiv.org/abs/2505.19339v1.

[1095] 王丽亚, 陈杰森, 周欣, A. 廷, 和黛安·M. 鲍姆加特纳. Aviationgpt: 用于航空领域的一个人工智能语言模型, arXiv 预印本 arXiv:2311.17686, 2023. URL https://arxiv.org/abs/2311.17686v1.

[1096] Liyuan Wang, Bo Lei, Qian Li, Hang Su, Jun Zhu, and Yi Zhong. Triple-memory networks: A brain-inspired method for continual learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[1097] Lu Wang, Fangkai Yang, Chaoyun Zhang, Junting Lu, Jiaxu Qian, Shilin He, Pu Zhao, Bo Qiao, Ray Huang, Si Qin, Qisheng Su, Jiayi Ye, Yudi Zhang, Jian-Guang Lou, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. Large action models: From inception to implementation, arXiv preprint arXiv:2412.10047, 2025. URL https://arxiv.org/abs/2412.10047.

[1098] Peijie Wang, Zhong-Zhi Li, Fei Yin, Xin Yang, Dekang Ran, and Cheng-Lin Liu. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. 2025.

[1099] Qineng Wang, Zihao Wang, Ying Su, and Yangqiu Song. On the discussion of large language models: Symmetry of agents and interplay with prompts, arXiv preprint arXiv:2311.07076, 2023. URL https://arxiv.org/abs/2311.07076v1.

[1100] Rongzheng Wang, Shuang Liang, Qizhi Chen, Jiasheng Zhang, and Ke Qin. Graphtool-instruction: Revolutionizing graph reasoning in llms through decomposed subtask instruction. *Knowledge Discovery and Data Mining*, 2024.

[1101] Shengnan Wang, Youhui Bai, Lin Zhang, Pingyi Zhou, Shixiong Zhao, Gong Zhang, Sen Wang, Renhai Chen, Hua Xu, and Hongwei Sun. Xl3m: A training-free framework for llm length extension based on segment-wise inference, arXiv preprint arXiv:2405.17755, 2024. URL https://arxiv.org/abs/2405.17755v1.

[1102] Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 2023.

[1103] Song Wang, Junhong Lin, Xiaojie Guo, Julian Shun, Jundong Li, and Yada Zhu. Reasoning of large language models over knowledge graphs with super-relations. *International Conference on Learning Representations*, 2025.

[1104] Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang, Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao Xie, Chuou Xu, Jihong Dai, Yibin Liu, Jialong Wu, Shengwei Ding, Long Li, Zhiwei Huang, Xinle Deng, Teng Yu, Gangan Ma, Han Xiao, Z. Chen, Danjun Xiang, Yunxia Wang, Yuanyuan Zhu, Yichen Xiao, Jing Wang, Yiru Wang, Siran Ding, Jiayang Huang, Jiayi Xu, Yilihamujiang Tayier, Zhenyu Hu, Yuan Gao, Chengfeng Zheng, Yu-Jie Ye, Yihan Li, Lei Wan, Xinyue Jiang, Yujie Wang, Siyuan Cheng, Zhule Song, Xiangru Tang, Xiaohua Xu, Ningyu Zhang, Huajun Chen, Y. Jiang, and Wangchunshu Zhou. Weaver: Foundation models for creative writing, arXiv preprint arXiv:2401.17268, 2024. URL https://arxiv.org/abs/2401.17268v1.

[1105] Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Visually-augmented language modeling. *International Conference on Learning Representations*, 2022.

[1106] Xiang Wang, Tinglin Huang, Dingxian Wang, Yancheng Yuan, Zhenguang Liu, Xiangnan He, and Tat seng Chua. Learning intents behind interactions with knowledge graph for recommendation. *The Web Conference*, 2021.

[1107] Xiao Wang, Isaac Lyngaas, A. Tsaris, Peng Chen, Sajal Dash, Mayanka Chandra Shekar, Tao Luo, Hong-Jun Yoon, M. Wahib, and J. Gounley. Ultra-long sequence distributed transformer, arXiv preprint arXiv:2311.02382, 2023. URL https://arxiv.org/abs/2311.02382v2.

[1096] 王丽媛, 雷波, 李倩, 苏杭, 朱俊, 和钟毅. 三重记忆网络：一种受大脑启发的持续学习方法. *IEEE Transactionson NeuralNetworks and LearningSystems*, 2020.

[1097] 王鲁, 杨方凯, 张超云, 陆俊庭, 钱嘉旭, 何世林, 赵普, 邱波, 黄睿, 秦思, 苏启胜, 叶佳怡, 张宇迪, 刘建光, 林清伟, Rajmohan Saravan, 张冬梅, 和张琪. 大动作模型：从起源到实现, arXiv preprint arXiv:2412.10047, 2025. URLhttps://arxiv.org/abs/2412.10047.

[1098] 王培杰, 李中志, 尹飞, 杨欣, 韩德刚, 和刘成林. Mv-math: 在多视觉环境下评估多模态数学推理. 2025.

[1099] 王秦能, 王子豪, 苏颖, 和宋阳秋. 关于大语言模型的讨论：代理的对称性与提示的相互作用, arXiv 预印本 arXiv:2311.07076, 2023. URLhttps://arxiv.org/abs/2311.07076v1.

[1100] 王荣正, 梁双, 陈启之, 张嘉盛, 和秦科. Graphtool-instruction：通过分解子任务指令革新大语言模型中的图推理. 知识发现与数据挖掘, 2024.

[1101] 王胜男, 白友会, 张林, 周平怡, 赵世雄, 张公, 王森, 陈仁海, 许华, 和孙宏伟. Xl3m：基于分段推理的无训练大语言模型长度扩展框架, arXiv 预印本 arXiv:2405.17755, 2024. URL https://arxiv.org/abs/2405.17755v1.

[1102] 王松, 朱瑶辰, 刘浩辰, 郑再毅, 陈晨, 和李俊东. 大型语言模型的知识编辑：一项调查. *ACM ComputingSurveys*, 2023.

[1103] 王松, 林俊宏, 郭晓杰, Shun Julian, 李俊东, 和朱亚达. 基于超关系的知识图谱上大型语言模型的推理. *InternationalConference on LearningRepresentations*, 2025.

[1104] 王天南, 陈嘉明, 贾庆瑞, 王帅, 方若宇, 王慧琳, 高兆伟, 谢春昭, 许楚鸥, 戴继宏, 刘一宾, 吴嘉隆, 丁胜伟, 李龙, 黄志伟, 邓新乐, 余腾, 马千千, 肖汉, 陈Z., 向丹军, 王云霞, 朱媛媛, 肖一晨, 王英儒, 丁思然, 黄嘉阳, 许嘉怡, 泰亚尔·依力哈木江, 胡振宇, 高远, 郑成峰, 叶雨洁, 李一涵, 万雷, 蒋新悦, 王宇杰, 成媛程, 宋竹乐, 唐祥儒, 许晓华, 张宁宇, 陈华军, 蒋Y., 周王春舒. Weaver：创意写作的基础模型, arXiv预印本 arXiv:2401.17268, 2024. URLhttps://arxiv.org/abs/2401.17268v1.

[1105] Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 视觉增强语言建模. 国际学习表征会议, 2022.

[1106] Xiang Wang, Tinglin Huang, Dingxian Wang, Yancheng Yuan, Zhenguang Liu, Xiangnan He, and Tat seng Chua. 学习知识图谱交互背后的意图用于推荐. 网络会议, 2021.

[1107] Xiao Wang, Isaac Lyngaas, A. Tsaris, Peng Chen, Sajal Dash, Mayanka Chandra Shekar, Tao Luo, Hong-Jun Yoon, M. Wahib, and J. Gounley. 超长序列分布式Transformer, arXiv预印本 arXiv:2311.02382, 2023. URLhttps://arxiv.org/abs/2311.02382v2.

[1108] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and S. Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. *European Conference on Computer Vision*, 2024.

[1109] Xiaolong Wang, Zhaolu Kang, Wangyuxuan Zhai, Xinyue Lou, Yunghwei Lai, Ziyue Wang, Yawen Wang, Kaiyu Huang, Yile Wang, Peng Li, and Yang Liu. Mucar: Benchmarking multilingual cross-modal ambiguity resolution for multimodal large language models, arXiv preprint arXiv:2506.17046v1, 2025. URL https://arxiv.org/abs/2506.17046v1.

[1110] Xiaoqiang Wang, Suyuchen Wang, Yun Zhu, and Bang Liu. R3mem: Bridging memory retention and retrieval via reversible compression. arXiv preprint, 2025.

[1111] Xiaoyang Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, I. Abdelaziz, Maria Chang, Achille Fokoue, B. Makni, Nicholas Mattei, and M. Witbrock. Improving natural language inference using external knowledge in the science questions domain. *AAAI Conference on Artificial Intelligence*, 2018.

[1112] Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and A. Eshaghi. Beyond the limits: A survey of techniques to extend the context length in large language models. *International Joint Conference on Artificial Intelligence*, 2024.

[1113] Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better llm agents. *International Conference on Machine Learning*, 2024.

[1114] Xuezhi Wang, Jason Wei, D. Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations*, 2022.

[1115] Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. Recmind: Large language model powered agent for recommendation, arXiv preprint arXiv:2308.14296, 2024. URL https://arxiv.org/abs/2308.14296.

[1116] Yani Wang. Application of large language models based on knowledge graphs in question-answering systems: A review. *Applied and Computational Engineering*, 2024.

[1117] Yanlin Wang, Wanjun Zhong, Yanxian Huang, Ensheng Shi, Min Yang, Jiachi Chen, Hui Li, Yuchi Ma, Qianxiang Wang, and Zibin Zheng. Agents in software engineering: Survey, landscape, and vision, arXiv preprint arXiv:2409.09030, 2024. URL https://arxiv.org/abs/2409.09030v2.

[1118] Yaqi Wang and Haipei Xu. Srsa: A cost-efficient strategy-router search agent for real-world human-machine interactions. *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2024.

[1119] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyun Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kaiming Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling. arXiv preprint, 2025.

[1120] Yiming Wang, Zhuosheng Zhang, and Rui Wang. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. *Annual Meeting of the Association for Computational Linguistics*, 2023.

[1108] 王晓寒, 张宇辉, Orr Zohar, 和 S. Yeung-Levy. Videoagent: 以大型语言模型为代理的长视频理解. 欧洲计算机视觉会议, 2024.

[1109]王晓龙, 康兆录, 翟王宇轩, 刘新悦, 赖云辉, 王子越, 王亚文, 黄凯宇, 王逸, 李鹏, 和 刘杨. Mucar: 针对多模态大型语言模型的跨语言跨模态歧义解析基准测试, arXiv 预印本 arXiv:2506.17046v1, 2025. URLhttps://arxiv.org/abs/2506.17046v1.

[1110] 王晓强, 王素纯, 朱云, 和 刘邦. R3mem: 通过可逆压缩桥接记忆保持和检索. arXiv 预印本, 2025.

[1111] 王小阳, Pavan Kapanipathi, Ryan Musa, 余墨, Kartik Talamadupula, I. Abdelaziz, Maria Chang, Achille Fokoue, B. Makni, Nicholas Mattei, 和 M. Witbrock. 利用外部知识改进科学问题领域的自然语言推理. *AAAI* 人工智能会议, 2018.

[1112] 王馨迪, Mahsa Salmani, Parsa Omidi, 任祥宇, Mehdi Rezagholizadeh, 和 A. Eshaghi. 超越极限：一项关于扩展大型语言模型上下文长度的技术综述.国际人工智能联合会议,2024.

[1113] 王行瑶, 陈阳怡, 袁立帆, 张一哲, 李云珠, 彭浩, 和 Ji Heng. 可执行代码动作激发更好的 llm 代理. 国际机器学习会议, 2024.

[1114] Xuezhi Wang, Jason Wei, D. Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. 自洽性提升语言模型中的思维链推理. 国际学习表征会议, 2022.

[1115] Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. Recmind: 基于大型语言模型的推荐代理, arXiv 预印本 arXiv:2308.14296, 2024. URL https://arxiv.org/abs/2308.14296.

[1116] Yani Wang. 基于知识图谱的大型语言模型在问答系统中的应用：综述. 应用与计算工程, 2024.

[1117] Yanlin Wang, Wanjun Zhong, Yanxian Huang, Ensheng Shi, Min Yang, Jiachi Chen, Hui Li, Yuchi Ma, Qianxiang Wang, and Zibin Zheng. 软件工程中的代理：调查、格局与愿景, arXiv 预印本 arXiv:2409.09030, 2024. URLhttps://arxiv.org/abs/2409.09030v2.

[1118] 王亚琪和徐海培. Srsa: 一种面向真实人机交互的高效策略路由搜索代理. *2024 IEEE International Conference onData Mining Workshops (ICDMW)*, 2024.

[1119] 王毅, 李新浩, 严壮, 何一南, 余嘉树, 曾祥云, 王澄澄, 马长莲, 黄海安, 高建飞, 陶敏, 陈凯明, 王文海, 乔宇, 王亚丽, 和王黎明. Internvideo2.5: 赋能视频大语言模型的长且丰富的上下文建模. arXiv preprint, 2025.

[1120] 王一鸣, 张卓升, 和王瑞. 基于大语言模型的元素感知摘要: 专家对齐评估和思维链方法. 计算语言学协会年度会议, 2023.

[1121] Yingming Wang and Pepa Atanasova. Self-critique and refinement for faithful natural language explanations, arXiv preprint arXiv:2505.22823, 2025. URL https://arxiv.org/abs/2505.22823v1.

[1122] Yiwei Wang, Wei Wang, Yuxuan Liang, Yujun Cai, and Bryan Hooi. Mixup for node and graph classification. In *Proceedings of the Web Conference 2021*, pages 3663–3674, 2021.

[1123] Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, Jingbo Shang, and Julian McAuley. Memoryllm: Towards self-updatable large language models, arXiv preprint arXiv:2402.04624, 2024. URL https://arxiv.org/abs/2402.04624.

[1124] Yu Wang, Dmitry Krotov, Yuanzhe Hu, Yifan Gao, Wangchunshu Zhou, Julian McAuley, Dan Gutfreund, Rogério Feris, and Zexue He. M+: Extending memoryllm with scalable long-term memory. arXiv preprint, 2025.

[1125] Yubin Wang, Xinyang Jiang, De Cheng, Wenli Sun, Dongsheng Li, and Cairong Zhao. Hpt++: Hierarchically prompting vision-language models with multi-granularity knowledge generation and improved structure modeling. arXiv preprint, 2024.

[1126] Yujie Wang, Shiju Wang, Shenhan Zhu, Fangcheng Fu, Xinyi Liu, Xuefeng Xiao, Huixia Li, Jiashi Li, Faming Wu, and Bin Cui. Flexsp: Accelerating large language model training via flexible sequence parallelism. *International Conference on Architectural Support for Programming Languages and Operating Systems*, 2024.

[1127] Yuntao Wang, Yanghe Pan, Zhou Su, Yi Deng, Quan Zhao, L. Du, Tom H. Luan, Jiawen Kang, and D. Niyato. Large model based agents: State-of-the-art, cooperation paradigms, security and privacy, and future trends. *IEEE Communications Surveys & Tutorials*, 2024.

[1128] Yuntao Wang, Shaolong Guo, Yanghe Pan, Zhou Su, Fahao Chen, Tom H. Luan, Peng Li, Jiawen Kang, and Dusit Niyato. Internet of agents: Fundamentals, applications, and challenges, arXiv preprint arXiv:2505.07176, 2025. URL https://arxiv.org/abs/2505.07176v1.

[1129] Yuxiang Wang, Xinnan Dai, Wenqi Fan, and Yao Ma. Exploring graph tasks with pure llms: A comprehensive benchmark and investigation, arXiv preprint arXiv:2502.18771v1, 2025. URL https://arxiv.org/abs/2502.18771v1.

[1130] Z. Wang, King Zhu, Chunpu Xu, Wangchunshu Zhou, Jiaheng Liu, Yibo Zhang, Jiashuo Wang, Ning Shi, Siyu Li, Yizhi Li, Haoran Que, Zhaoxiang Zhang, Yuanxing Zhang, Ge Zhang, Ke Xu, Jie Fu, and Wenhao Huang. Mio: A foundation model on multimodal tokens, arXiv preprint arXiv:2409.17692v3, 2024. URL https://arxiv.org/abs/2409.17692v3.

[1131] Zheng Wang, Shu Xian Teo, Jieer Ouyang, Yongjun Xu, and Wei Shi. M-rag: Reinforcing large language model performance through retrieval-augmented generation with multiple partitions. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[1132] Zhiruo Wang, Zhoujun Cheng, Hao Zhu, Daniel Fried, and Graham Neubig. What are tools anyway? a survey from the language model perspective, arXiv preprint arXiv:2403.15452, 2024. URL https://arxiv.org/abs/2403.15452v1.

[1121] Yingming Wang 和 Pepa Atanasova. 自我批评与改进以实现忠实自然语言解释，arXiv 预印本 arXiv:2505.22823，2025。URLhttps://arxiv.org/abs/2505.22823v1。

王怡伟, 王伟, 梁宇轩, 蔡宇君, 和胡斌. Mixup用于节点和图分类. 在《2021年网络会议论文集》中, 第3663-3674页, 2021.

[1123] 王宇，高一帆，陈秀思，蒋浩明，李时阳，杨景峰，尹清宇，李政，李娴，尹冰，商景波，以及朱利安·麦克劳利。Memoryllm：面向可自我更新的大型语言模型，arXiv预印本arXiv:2402.04624，2024年。URLhttps://arxiv.org/abs/2402.04624。

[1124] 王宇, Dmitry Krotov, 胡元哲, 高一帆, 周王春舒, Julian McAuley, Dan Gut- freund, Rogério Feris, 和 何泽学. M+: 扩展 memoryllm 的可扩展长期记忆. arXiv 预印本, 2025.

[1125] 王宇斌, 姜新阳, 成德, 孙文莉, 李东升, 和 赵才荣. Hpt++: 基于多粒度知识生成和改进结构建模的分层提示视觉语言模型. arXiv 预印本, 2024.

[1126] 王宇洁, 王世举, 朱神翰, 傅方成, 刘新怡, 肖雪峰, 李会霞, 李嘉石, 吴发明, 和 崔斌. Flexsp: 通过灵活序列并行性加速大型语言模型训练. 国际会议：编程语言与操作系统架构支持, 2024.

[1127] 王云涛, 潘阳和, 苏周, 邓毅, 赵全, L. 杜, 陆汤姆·H., 康嘉文, 和 Niyato·D. 基于大型模型的智能体：最新进展、合作范式、安全和隐私以及未来趋势。*IEEE* 通信调查与教程, 2024。

[1128] 王云涛, 郭少龙, 潘阳和, 苏周, 陈发浩, 陆汤姆·H., 李鹏, 康嘉文, 和 Niyato·杜西特. 智能体互联网：基础、应用和挑战, arXiv 预印本 arXiv:2505.07176, 2025. URLhttps://arxiv.org/abs/2505.07176v1.

[1129] 王宇翔, 戴欣南, 范文奇, 和 马瑶. 使用纯 llms 探索图任务：综合基准和调查, arXiv 预印本 arXiv:2502.18771v1, 2025. URLhttps://arxiv.org/abs/2502.18771v1.

[1130] 王志，朱 king，徐春pu，周王chunshu，刘 jiaheng，张 yibo，王 jiashuo，石 ning，李 siyu，李 yizhi，钱 haoran，张 zhaoxiang，张 yuanxing，张 ge，徐 ke，付 jie，黄 wenhao。Mio: 一个基于多模态 token 的基础模型，arXiv 预印本 arXiv:2409.17692v3，2024。URLhttps://arxiv.org/abs/2409.17692v3。

[1131] 王正，谢水 shu，欧 jieer，徐 yongjun，和石 wei。M-rag: 通过多分区检索增强生成来强化大型语言模型性能。计算语言学协会年度会议，2024。

[1132] 王志ruo，程 Zhoujun，朱 hao，弗里德 Daniel，和纽 big Graham。工具到底是什么？从语言模型的角度来看的调查，arXiv 预印本 arXiv:2403.15452，2024。URLhttps://arxiv.org/abs/2403.15452v1。

[1133] Ziyang Wang, Jianzhou You, Haining Wang, Tianwei Yuan, Shichao Lv, Yang Wang, and Limin Sun. Honeygpt: Breaking the trilemma in terminal honeypots with large language model, arXiv preprint arXiv:2406.01882, 2024. URL https://arxiv.org/abs/2406.01882v2.

[1134] Ziyue Wang, Chi Chen, Yiqi Zhu, Fuwen Luo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Maosong Sun, and Yang Liu. Browse and concentrate: Comprehending multimodal content via prior-llm context fusion. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[1135] Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory, arXiv preprint arXiv:2409.07429, 2024. URL https://arxiv.org/abs/2409.07429.

[1136] Irene Weber. Large language models are pattern matchers: Editing semi-structured and structured documents with chatgpt. *AKWI Jahrestagung*, 2024.

[1137] Hui Wei, Chenyue Feng, and Jianning Zhang. Modeling of memory mechanisms in cerebral cortex and simulation of storage performance, arXiv preprint arXiv:2401.00381, 2023. URL https://arxiv.org/abs/2401.00381v2.

[1138] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *Neural Information Processing Systems*, 2022.

[1139] Jerry W. Wei, Le Hou, Andrew Kyle Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc V. Le. Symbol tuning improves in-context learning in language models. *Conference on Empirical Methods in Natural Language Processing*, 2023.

[1140] Shaopeng Wei, Yu Zhao, Xingyan Chen, Qing Li, Fuzhen Zhuang, Ji Liu, and Gang Kou. Graph learning and its advancements on large language models: A holistic survey, arXiv preprint arXiv:2212.08966, 2022. URL https://arxiv.org/abs/2212.08966v5.

[1141] Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, Hyokun Yun, and Lihong Li. Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning. arXiv preprint, 2025.

[1142] Zhiyuan Wei, Jing Sun, Zijian Zhang, and Xianhao Zhang. Llm-smartaudit: Advanced smart contract vulnerability detection. arXiv preprint, 2024.

[1143] Rebecca Westhäußer, Frederik Berenz, Wolfgang Minker, and Sebastian Zepf. Caim: Development and evaluation of a cognitive ai memory framework for long-term interaction with intelligent agents. arXiv preprint, 2025.

[1144] Danny Weyns and F. Oquendo. An architectural style for self-adaptive multi-agent systems, arXiv preprint arXiv:1909.03475, 2019. URL https://arxiv.org/abs/1909.03475v1.

[1145] Erik Wijmans, Brody Huval, Alexander Hertzberg, V. Koltun, and Philipp Krähenbühl. Cut your losses in large-vocabulary language models. *International Conference on Learning Representations*, 2024.

[1146] Wikipedia contributors. Agent communications language — Wikipedia, the free encyclopedia, 2025. URL https://en.wikipedia.org/wiki/Agent_Communications_Language. [Online; accessed 17-July-2025].

[1133] 张一阳, 游建舟, 王海宁, 袁天伟, 吕时超, 王阳, 和孙立明. Honeygpt: 使用大型语言模型打破终端蜜罐的三难困境, arXiv 预印本 arXiv:2406.01882, 2024. URLhttps://arxiv.org/abs/2406.01882v2.

[1134] 王子越, 陈驰, 朱一奇, 罗福文, 李鹏, 阎明, 张继, 黄飞, 孙茂松, 和刘杨. 浏览和聚焦：通过先验-llm 上下文融合理解多模态内容. 计算语言学协会年度会议, 2024.

[1135] 王卓如, 毛嘉源, 弗里德丹尼尔, 和纽比格格雷厄姆. 代理工作流内存, arXiv 预印本 arXiv:2409.07429, 2024. URLhttps://arxiv.org/abs/2409.07429.

[1136] Irene Weber. 大型语言模型是模式匹配器：使用ChatGPT编辑半结构化和结构化文档。 *AKWI Jahrestagung*, 2024。

[1137] Hui Wei, Chenyue Feng, and Jianning Zhang. 大脑皮层记忆机制的建模和存储性能的模拟, arXiv预印本 arXiv:2401.00381，2023。URL https://arxiv.org/abs/2401.00381v2.

[1138] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 思维链提示在大语言模型中引发推理。 *Neural InformationProcessing Systems*, 2022。

[1139] Jerry W. Wei, Le Hou, Andrew Kyle Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc V. Le. 符号微调提高了语言模型中的情境学习。 自然语言处理经验方法会议, 2023。

[1140] 魏少鹏, 赵宇, 陈行言, 李清, 庄福珍, 刘继, 和寇刚. 图学习及其在大语言模型上的进展：一项整体性调查, arXiv 预印本 arXiv:2212.08966, 2022. URLhttps://arxiv.org/abs/2212.08966v5.

[1141] 魏哲培, 姚文林, 刘瑶, 张伟志, 陆秦, 邱良, 余长龙, 许普阳, 张超, 尹冰, 尹赫坤, 和李立宏. Webagent-r1：通过端到端多轮强化学习训练网络代理. arXiv 预印本, 2025.

[1142] 魏志远, 孙静, 张子健, 和张先浩. Llm-smartaudit：高级智能合约漏洞检测. arXiv 预印本, 2024.

[1143] Rebecca Westhäußer, Frederik Berenz, Wolfgang Minker, 和 Sebastian Zepf. Caim: 一种用于与智能代理进行长期交互的认知式AI记忆框架的开发与评估。arXiv 预印本，2025年。

[1144] Danny Weyns 和 F. Oquendo. 一种自适应多智能体系统的架构风格，arXiv 预印本 arXiv:1909.03475，2019年。URLhttps://arxiv.org/abs/1909.03475v1。

[1145] Erik Wijmans, Brody Huval, AlexanderHertzberg, V.Koltun, 和 PhilippKrähenbühl. 大词汇量语言模型中的Cutyourlosses。国际学习表示会议, 2024年。

[1146] Wikipedia 贡献者。智能体通信语言 — 维基百科，自由百科全书，2025年。URLhttps://en.wikipedia.org/wiki/Agent_Communications_Language。 [在线；访问于17-July-2025]。

[1147] Genta Indra Winata, Lingjue Xie, Karthik Radhakrishnan, Shijie Wu, Xisen Jin, Pengxiang Cheng, Mayank Kulkarni, and Daniel Preotiuc-Pietro. Overcoming catastrophic forgetting in massively multilingual continual learning, arXiv preprint arXiv:2305.16252, 2023. URL https://arxiv.org/abs/2305.16252.

[1148] Beong woo Kwak, Minju Kim, Dongha Lim, Hyungjoo Chae, Dongjin Kang, Sunghwan Kim, Dongil Yang, and Jinyoung Yeo. Toolhaystack: Stress-testing tool-augmented language models in realistic long-term interactions, arXiv preprint arXiv:2505.23662, 2025. URL https://arxiv.org/abs/2505.23662v1.

[1149] Biao Wu, Yanda Li, Meng Fang, Zirui Song, Zhiwei Zhang, Yunchao Wei, and Ling Chen. Foundations and recent trends in multimodal mobile agents: A survey, arXiv preprint arXiv:2411.02006, 2024. URL https://arxiv.org/abs/2411.02006v2.

[1150] Cheng-Kuang Wu, Zhi Rui Tam, Chieh-Yen Lin, Yun-Nung Chen, and Hung yi Lee. Streambench: Towards benchmarking continuous improvement of language agents. *Neural Information Processing Systems*, 2024.

[1151] Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. Webdancer: Towards autonomous information seeking agency, arXiv preprint arXiv:2505.22648, 2025. URL https://arxiv.org/abs/2505.22648v2.

[1152] Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Deyu Zhou, Pengjun Xie, and Fei Huang. Webwalker: Benchmarking llms in web traversal, arXiv preprint arXiv:2501.07572, 2025. URL https://arxiv.org/abs/2501.07572v2.

[1153] Junde Wu, Jiayuan Zhu, and Yuyuan Liu. Agentic reasoning: Reasoning llms with tools for the deep research, arXiv preprint arXiv:2502.04644, 2025. URL https://arxiv.org/abs/2502.04644v1.

[1154] Likang Wu, Zhilan Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. A survey on large language models for recommendation. *World wide web (Bussum)*, 2023.

[1155] M Wu, J Yang, J Jiang, M Li, K Yan, and H Yu.... Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. 2025. URL https://arxiv.org/abs/2505.19255.

[1156] Mengsong Wu, Tong Zhu, Han Han, Chuanyuan Tan, Xiang Zhang, and Wenliang Chen. Seal-tools: Self-instruct tool learning dataset for agent tuning and detailed benchmark. *Natural Language Processing and Chinese Computing*, 2024.

[1157] Panlong Wu, Ting Wang, Yifei Zhong, Haoqi Zhang, Zitong Wang, and Fangxin Wang. Deep-form: Reasoning large language model for communication system formulation, arXiv preprint arXiv:2506.08551, 2025. URL https://arxiv.org/abs/2506.08551v2.

[1158] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, A. Awadallah, Ryen W. White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, arXiv preprint arXiv:2308.08155, 2023. URL https://arxiv.org/abs/2308.08155v2.

[1147] Genta Indra Winata, Lingjue Xie, Karthik Radhakrishnan, Shijie Wu, Xisen Jin, Pengxiang Cheng, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 克服大规模多语言持续学习中的灾难性遗忘, arXiv preprint arXiv:2305.16252，2023。URLhttps://arxiv.org/abs/2305.16252.

[1148] Beong woo Kwak, Minju Kim, Dongha Lim, Hyungjoo Chae, Dongjin Kang, Sunghwan Kim, Dongil Yang, and Jinyoung Yeo. Toolhaystack：在真实长期交互中对工具增强语言模型进行压力测试, arXiv preprint arXiv:2505.23662，2025。URLhttps://arxiv.org/abs/2505.23662v1.

[1149] Biao Wu, Yanda Li, Meng Fang, ZiruiSong, ZhiweiZhang, Yunchao Wei, and LingChen. 多模态移动代理的基础和最新趋势：一项调查，arXiv preprint arXiv:2411.02006，2024。URL https://arxiv.org/abs/2411.02006v2.

[1150] 吴成光，谭志瑞，林杰延，陈云农，以及李洪毅。Streambench：迈向语言代理的持续改进基准测试。神经信息处理系统, 2024.

[1151] 吴家龙, 李百选, 方润南, 尹文标, 张立文, 陶正伟, 张定初, 石泽坤, 姜勇, 谢鹏军, 黄飞, 周景仁. Webdancer:迈向自主信息获取代理, arXiv预印本arXiv:2505.22648, 2025. URLhttps://arxiv.org/abs/2505.22648v2.

[1152] 吴家龙，尹文标，蒋勇，王正林，席泽坤，方润南，周德宇，谢鹏军，和黄飞. Webwalker: 在网络遍历中基准测试大型语言模型，arXiv预印本arXiv:2501.07572, 2025. URLhttps://arxiv.org/abs/2501.07572v2.

[1153] 吴俊德, 朱嘉源, 和刘宇源. 自主推理：使用工具进行深度研究的推理大语言模型, arXiv 预印本 arXiv:2502.04644, 2025. URLhttps://arxiv.org/abs/2502.04644v1.

[1154] 吴立康, 郑志兰, 邱兆鹏, 王浩, 顾洪超, 沈庭嘉, 秦川, 朱晨, 朱恒舒, 刘琪, 熊辉, 和陈恩宏. 推荐用大语言模型的综述. *万维网 (Bussum)*, 2023.

[1155] M Wu, J Yang, J Jiang, M Li, K Yan, 和 H Yu... . Vtool-r1: Vlms 通过多模态工具使用的强化学习学习用图像思考. 2025. URL https://arxiv.org/abs/2505.19255.

[1156] 吴梦松, 朱通, 韩汉, 谭传元, 张翔, 和陈文亮. Seal-tools: 用于代理调优和详细基准测试的自指令工具学习数据集. 自然语言处理与中文计算, 2024.

[1157] 吴盘龙, 王挺, 中怡飞, 张浩奇, 王子通, 和王方欣. Deep- form: 用于通信系统公式的推理大语言模型, arXiv 预印本 arXiv:2506.08551, 2025. URLhttps://arxiv.org/abs/2506.08551v2.

[1158] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, A. Awadallah, Ryen W. White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, arXiv preprint arXiv:2308.08155, 2023. URL https://arxiv.org/abs/2308.08155v2.

[1159] Ruofan Wu, Youngwon Lee, Fan Shu, Danmei Xu, Seung won Hwang, Zhewei Yao, Yuxiong He, and Feng Yan. Composerag: A modular and composable rag for corpus-grounded multi-hop question answering, arXiv preprint arXiv:2506.00232, 2025. URL https://arxiv.org/abs/2506.00232v1.

[1160] Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, and C. Xue. Retrieval-augmented generation for natural language processing: A survey, arXiv preprint arXiv:2407.13193, 2024. URL https://arxiv.org/abs/2407.13193v3.

[1161] Shirley Wu, Shiyu Zhao, Qian Huang, Kexin Huang, Michihiro Yasunaga, V. Ioannidis, Karthik Subbian, J. Leskovec, and James Zou. Avatar: Optimizing llm agents for tool usage via contrastive reasoning. *Neural Information Processing Systems*, 2024.

[1162] Suhang Wu, Minlong Peng, Yue Chen, Jinsong Su, and Mingming Sun. Eva-kellm: A new benchmark for evaluating knowledge editing of llms, arXiv preprint arXiv:2308.09954, 2023. URL https://arxiv.org/abs/2308.09954.

[1163] Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason E Weston, and Sainbayar Sukhbaatar. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. arXiv preprint, 2024.

[1164] Tong Wu, Chong Xiang, Jiachen T. Wang, and Prateek Mittal. Effectively controlling reasoning models through thinking intervention, arXiv preprint arXiv:2503.24370, 2025. URL https://arxiv.org/abs/2503.24370v3.

[1165] Xinbo Wu and L. Varshney. A meta-learning perspective on transformers for causal language modeling. *Annual Meeting of the Association for Computational Linguistics*, 2023.

[1166] Xue Wu and Kostas Tsioutsiouliklis. Thinking with knowledge graphs: Enhancing llm reasoning through structured data, arXiv preprint arXiv:2412.10654, 2024. URL https://arxiv.org/abs/2412.10654v1.

[1167] Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang, Huifeng Guo, Ruiming Tang, and Yong Liu. From human memory to ai memory: A survey on memory mechanisms in the era of llms, arXiv preprint arXiv:2504.15965, 2025. URL https://arxiv.org/abs/2504.15965v2.

[1168] Zengqing Wu and Takayuki Ito. The hidden strength of disagreement: Unraveling the consensus-diversity tradeoff in adaptive multi-agent systems, arXiv preprint arXiv:2502.16565, 2025. URL https://arxiv.org/abs/2502.16565v2.

[1169] Zihao Wu, Lu Zhang, Chao-Yang Cao, Xiao-Xing Yu, Haixing Dai, Chong-Yi Ma, Zheng Liu, Lin Zhao, Gang Li, Wei Liu, Quanzheng Li, Dinggang Shen, Xiang Li, Dajiang Zhu, and Tianming Liu. Exploring the trade-offs: Unified large language models vs local fine-tuned models for highly-specific radiology nli task. *IEEE Transactions on Big Data*, 2023.

[1170] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Qin Liu, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huan, and Tao Gui. The rise and potential of large language model based agents: A survey, arXiv preprint arXiv:2309.07864, 2023. URL https://arxiv.org/abs/2309.07864v3.

[1159] 吴若凡, 李永元, 舒帆, 许丹梅, 黄胜温, 姚哲伟, 何宇雄, 和冯岩. Composerag: 一种面向语料库驱动的多跳问答的模块化和可组合式检索增强生成, arXiv preprint arXiv:2506.00232, 2025. URL https://arxiv.org/abs/2506.00232v1.

[1160] 吴尚宇, 邢颖, 崔宇飞, 吴浩伦, 陈灿, 袁叶, 黄连明, 刘雪, 郭铁伟, 郭楠, 和薛雪. 自然语言处理的检索增强生成: 一篇综述, arXiv preprint arXiv:2407.13193, 2024. URL https://arxiv.org/abs/2407.13193v3.

[1161] Shirley Wu, Shiyu Zhao, Qian Huang, Kexin Huang, Michihiro Yasunaga, V. Ioannidis, Karthik Subbian, J. Leskovec, and James Zou. Avatar: 通过对比推理优化 llm 代理的工具使用. *Neural Information ProcessingSystems*, 2024.

[1162] Suhang Wu, Minlong Peng, Yue Chen, Jinsong Su, and Mingming Sun. Eva-kellm: 一个评估 llm 知识编辑的新基准, arXiv preprint arXiv:2308.09954, 2023. URL https://arxiv.org/abs/2308.09954.

[1163] Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason E Weston, and Sainbayar Sukhbaatar. Meta-rewarding language models: 通过 llm 作为元裁判实现自我改进的 alignment. arXiv preprint, 2024.

[1164] 童武, 程翔, 王嘉晨, 和 Prateek Mittal. 通过思维干预有效控制推理模型, arXiv preprint arXiv:2503.24370, 2025. URL https://arxiv.org/abs/2503.24370v3.

[1165] 吴新波和L. Varshney. 基于元学习的因果语言模型Transformer视角.计算语言学协会年会, 2023.

[1166] 吴雪和Kostas Tsioutsiouliklis. 基于知识图谱的思维：通过结构化数据增强llm推理, arXiv preprint arXiv:2412.10654, 2024. URLhttps://arxiv.org/abs/2412.10654v1.

[1167] 吴亚雄, 梁胜, 张晨, 王一超, 张永越, 郭会峰, 唐瑞明, 和 刘勇. 从人类记忆到AI记忆：大语言模型时代记忆机制的综述, arXiv preprint arXiv:2504.15965, 2025. URL https://arxiv.org/abs/2504.15965v2.

[1168] 曾庆清和伊藤隆之. 不同意的隐藏力量：解开自适应多智能体系统中的共识-多样性权衡，arXiv预印本 arXiv:2502.16565，2025年。URLhttps://arxiv.org/abs/2502.16565v2.

[1169] 朱兆华, 张鲁, 曹超阳, 余晓星, 戴海星, 马崇毅, 刘政, 赵琳, 李刚, 刘伟, 李全正, 沈定刚, 肖立, 朱大江, 刘天明. 探索权衡：统一大语言模型与本地微调模型在高度特定放射学NLI任务中的应用. *IEEE Transactions on Big Data*, 2023.

[1170] 张志恒，陈文祥，郭欣，何伟，丁依文，Hong Boyang，张明，王俊哲，金森杰，周恩宇，郑瑞，范晓然，王晓，熊利毛，刘秦，周宇浩，王伟然，蒋长浩，邹一成，刘祥阳，尹张跃，窦石涵，翁荣祥，程文森，张琪，秦文娟，郑永岩，邱希鹏，宦宣景，桂涛。基于大型语言模型的智能体崛起与潜力：一项调查，arXiv预印本arXiv:2309.07864，2023。URLhttps://arxiv.org/abs/2309.07864v3.

[1171] Menglin Xia, Victor Ruehle, Saravan Rajmohan, and Reza Shokri. Minerva: A programmable memory test benchmark for language models, arXiv preprint arXiv:2502.03358, 2025. URL https://arxiv.org/abs/2502.03358v2.

[1172] Yuchen Xia, Manthan Shenoy, N. Jazdi, and M. Weyrich. Towards autonomous system: flexible modular production system enhanced with large language model agents. *IEEE International Conference on Emerging Technologies and Factory Automation*, 2023.

[1173] Yutong Xia, Ao Qu, Yunhan Zheng, Yihong Tang, Dingyi Zhuang, Yuxuan Liang, Cathy Wu, Roger Zimmermann, and Jinhua Zhao. Reimagining urban science: Scaling causal inference with large language models, arXiv preprint arXiv:2504.12345v3, 2025. URL https://arxiv.org/abs/2504.12345v3.

[1174] Zhishang Xiang, Chuanjie Wu, Qinggang Zhang, Shengyuan Chen, Zijin Hong, Xiao Huang, and Jinsong Su. When to use graphs in rag: A comprehensive analysis for graph retrieval-augmented generation, arXiv preprint arXiv:2506.05690, 2025. URL https://arxiv.org/abs/2506.05690v1.

[1175] Chaojun Xiao, Pengle Zhang, Xu Han, Guangxuan Xiao, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, Song Han, and Maosong Sun. Infllm: Training-free long-context extrapolation for llms with an efficient context memory. *Neural Information Processing Systems*, 2024.

[1176] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *International Conference on Learning Representations*, 2023.

[1177] MiniCPM Team Chaojun Xiao, Yuxuan Li, Xu Han, Yuzhuo Bai, Jie Cai, Haotian Chen, Wentong Chen, Xin Cong, Ganqu Cui, Ning Ding, Shengda Fan, Yewei Fang, Zixuan Fu, Wenyu Guan, Yitong Guan, Junshao Guo, Yu-Xuan Han, Bingxiang He, Yuxian Huang, Cunliang Kong, Qiu-Tong Li, Siyuan Li, Wenhao Li, Yanghao Li, Yishan Li, Zhen Li, Dan Liu, Biyuan Lin, Yankai Lin, Xiang Long, Quanyu Lu, Ya-Ting Lu, Pei Luo, Hongya Lyu, Litu Ou, Yinxu Pan, Zekai Qu, Qundong Shi, Zijun Song, Jiayu Su, Zhou Su, Ao Sun, Xiang ping Sun, Peijun Tang, Fang-Ming Wang, Feng Wang, Shuo Wang, Yudong Wang, Yesai Wu, Zhenyu Xiao, Jie Xie, Zi-Kang Xie, Yukun Yan, Jia-Li Yuan, Kai Zhang, Lei Zhang, Linyu Zhang, Xueren Zhang, Yudi Zhang, Hengyu Zhao, Weilin Zhao, Weilun Zhao, Yuanqian Zhao, Zhijun Zheng, Ge Zhou, Jie Zhou, Wei Zhou, Zihan Zhou, Zi-An Zhou, Zhiyuan Liu, Guoyang Zeng, Chaochao Jia, Dahai Li, and Maosong Sun. Minicpm4: Ultra-efficient llms on end devices, arXiv preprint arXiv:2506.07900, 2025. URL https://arxiv.org/abs/2506.07900v1.

[1178] Yang Xiao, Jiashuo Wang, Ruifeng Yuan, Chunpu Xu, Kaishuai Xu, Wenjie Li, and Pengfei Liu. Limopro: Reasoning refinement for efficient and effective test-time scaling, arXiv preprint arXiv:2505.19187, 2025. URL https://arxiv.org/abs/2505.19187v1.

[1179] Yilin Xiao, Chuang Zhou, Qinggang Zhang, Bo Li, Qing Li, and Xiao Huang. Reliable reasoning path: Distilling effective guidance for llm reasoning with knowledge graphs, arXiv preprint arXiv:2506.10508, 2025. URL https://arxiv.org/abs/2506.10508v1.

[1180] Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. Travelplanner: A benchmark for real-world planning with language agents. *International Conference on Machine Learning*, 2024.

[1181] Yuxi Xie, Anirudh Goyal, Xiaobao Wu, Xunjian Yin, Xiao Xu, Min-Yen Kan, Liangming Pan, and William Yang Wang. Coral: Order-agnostic language modeling for efficient iterative refinement, arXiv preprint arXiv:2410.09675, 2024. URL https://arxiv.org/abs/2410.09675v1.

[1171] 夏梦琳, 维克多·鲁埃尔, 萨拉万·拉贾莫汉和 丽萨·肖克里. Minerva: 一种用于语言模型的可编程内存测试基准, arXiv 预印本 arXiv:2502.03358, 2025. URLhttps://arxiv.org/abs/2502.03358v2.

[1172] 夏宇辰, 沈曼森, N.贾兹迪和M.魏里希.迈向自主系统：增强有大型语言模型代理的柔性模块化生产系统.*IEEE*国际会议新兴技术与工厂自动化, 2023.

[1173] 余通霞, 阿曲, 郑云涵, 唐一红, 庄定一, 梁宇轩, 吴凯蒂, 罗杰·齐默曼, 和赵金花. 重塑城市科学：利用大型语言模型扩展因果推理, arXiv 预印本 arXiv:2504.12345v3, 2025. URL https://arxiv.org/abs/2504.12345v3.

[1174] 张志上, 吴传杰, 张庆刚, 陈胜元, 洪子金, 黄晓, 和苏晋松. 何时使用图在rag: 图检索增强生成的一个全面分析, arXiv预印本 arXiv:2506.05690, 2025. URL https://arxiv.org/abs/2506.05690v1.

[1175] 肖超军, 张鹏磊, 韩旭, 肖广宣, 林彦开, 张铮岩, 刘志远, 韩松, 和孙茂松. Infllm: 为llms提供高效上下文记忆的无训练长上下文外推. 神经信息处理系统, 2024.

[1176] 肖广宣, 田元东, 陈北地, 韩松, 和利克·刘易斯. 带有注意力陷阱的高效流式语言模型. 学习表示国际会议, 2023.

[1177] MiniCPM 团队 肖超军, 李宇轩, 韩旭, 白宇卓, 蔡杰, 陈浩天, 陈文通, 从丛, 崔甘泉, 丁宁, 樊胜达, 方晔伟, 傅子轩, 关文宇, 关奕通, 郭俊超, 韩宇轩, 何炳祥, 黄宇娴, 孔存亮, 李求通, 李思源, 李文昊, 李阳浩, 李奕山 , 李振 , 刘丹 , 林比远 , 林彦凯 , 龙翔 , 陆泉宇 , 陆亚婷 , 罗培 , 吕红亚 , 欧立图 , 潘银旭 , 钱凯 , 石昆东, 宋子军, 苏嘉宇, 苏舟, 孙澳, 孙翔平, 唐培军, 王方明, 王峰, 王硕, 王宇东, 吴叶赛, 肖振宇, 谢杰, 谢子康, 闫雨坤, 袁家丽, 张凯, 张雷, 张琳宇, 张学仁, 张雨迪, 赵恒宇, 赵伟林, 赵伟伦, 赵元倩, 郑志军, 周格, 周杰, 周伟, 周子涵, 周子安, 刘志远, 曾国阳, 贾超超, 李大海, 孙茂松. Minicpm4: 终端设备上的超高效llms, arXiv preprint arXiv:2506.07900, 2025. URLhttps://arxiv.org/abs/2506.07900v1.

[1178] 杨晓，嘉树王，袁瑞峰，徐春普，徐凯帅，李文杰，和刘鹏飞. Limopro: 用于高效和有效测试时扩展的推理细化，arXiv 预印本 arXiv:2505.19187, 2025. URLhttps://arxiv.org/abs/2505.19187v1.

[1179] 谢一林，周创，张庆刚，李波，李清，和黄晓. 可靠推理路径：利用知识图谱为 LLM 推理提取有效指导，arXiv 预印本 arXiv:2506.10508, 2025. URLhttps://arxiv.org/abs/2506.10508v1.

[1180] 谢建，张凯，陈江杰，朱停辉，刘仁泽，田元东，谢阳华，和苏宇. Travelplanner: 语言代理现实世界规划基准. 机器学习国际会议, 2024.

[1181] 谢宇曦, Anirudh Goyal, 吴晓宝, 尹训健, 许晓, Kan Min-Yen, 潘良明, 和 王威廉杨. Coral: 非顺序语言建模用于高效的迭代细化, arXiv 预印本 arXiv:2410.09675, 2024. URLhttps://arxiv.org/abs/2410.09675v1.

[1182] Yue Xing, Tao Yang, Yijiashun Qi, Minggu Wei, Yu Cheng, and Honghui Xin. Structured memory mechanisms for stable context representation in large language models, arXiv preprint arXiv:2505.22921, 2025. URL https://arxiv.org/abs/2505.22921v1.

[1183] Guangzhi Xiong, Qiao Jin, Xiao Wang, Yin Fang, Haolin Liu, Yifan Yang, Fangyuan Chen, Zhixing Song, Dengyu Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. Rag-gym: Systematic optimization of language agents for retrieval-augmented generation. arXiv preprint, 2025.

[1184] Haoyi Xiong, Zhiyuan Wang, Xuhong Li, Jiang Bian, Zeke Xie, Shahid Mumtaz, and Laura E. Barnes. Converging paradigms: The synergy of symbolic and connectionist ai in llm-empowered autonomous agents, arXiv preprint arXiv:2407.08516, 2024. URL https://arxiv.org/abs/2407.08516v5.

[1185] Junjie Xiong, Changjia Zhu, Shuhang Lin, Chong Zhang, Yongfeng Zhang, Yao Liu, and Lingyao Li. Invisible prompts, visible threats: Malicious font injection in external resources for large language models, arXiv preprint arXiv:2505.16957, 2025. URL https://arxiv.org/abs/2505.16957v1.

[1186] Zhen Xiong, Yujun Cai, Bryan Hooi, Nanyun Peng, Zhecheng Li, and Yiwei Wang. Enhancing llm character-level manipulation via divide and conquer. 2025.

[1187] Zhen Xiong, Yujun Cai, Zhecheng Li, and Yiwei Wang. Mapping the minds of llms: A graph-based analysis of reasoning llm. 2025.

[1188] Zhen Xiong, Yujun Cai, Zhecheng Li, and Yiwei Wang. Unveiling the potential of diffusion large language model in controllable generation. 2025.

[1189] Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts llm agents: An empirical study of experience-following behavior, arXiv preprint arXiv:2505.16067, 2025. URL https://arxiv.org/abs/2505.16067v1.

[1190] Chunmei Xu, Shengheng Liu, Cheng Zhang, Yongming Huang, Zhaohua Lu, and Luxi Yang. Multi-agent reinforcement learning based distributed transmission in collaborative cloud-edge systems. *IEEE Transactions on Vehicular Technology*, 2021.

[1191] Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, et al. Redstar: Does scaling long-cot data unlock better slow-reasoning systems? 2025.

[1192] Hongshen Xu, Su Zhu, Zihan Wang, Hang Zheng, Da Ma, Ruisheng Cao, Shuai Fan, Lu Chen, and Kai Yu. Reducing tool hallucination via reliability alignment, arXiv preprint arXiv:2412.04141, 2024. URL https://arxiv.org/abs/2412.04141v3.

[1193] Hu Xu, Gargi Ghosh, Po-Yao (Bernie) Huang, Dmytro Okhonko, Armen Aghajanyan, and Florian Metze Luke Zettlemoyer Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *Conference on Empirical Methods in Natural Language Processing*, 2021.

[1194] Mengjia Xu. Understanding graph embedding methods and their applications. *SIAM Review*, 2020.

[1195] Minrui Xu, Hongyang Du, Dusist Niyato, Jiawen Kang, Zehui Xiong, Shiwen Mao, Zhu Han, A. Jamalipour, Dong In Kim, X. Shen, Victor C. M. Leung, and H. Poor. Unleashing the power of edge-cloud generative ai in mobile networks: A survey of aigc services. *IEEE Communications Surveys and Tutorials*, 2023.

[1182] 余兴, 杨涛, 齐一霁, 魏明谷, 余成, 和辛洪辉. 大型语言模型中稳定的上下文表示的 结构化内存机制, arXiv 预印本 arXiv:2505.22921, 2025. URLhttps://arxiv.org/abs/2505.22921v1.

[1183] 熊广智, 金乔, 王晓, 方寅, 刘浩林, 杨奕帆, 陈方圆, 宋志兴, 王登宇, 张敏嘉, 陆志勇, 和张爱东. Rag-gym: 针对检索增强生成的语言代理的系统优化. arXiv 预印本, 2025.

[1184] 熊浩毅, 王志远, 李旭红, 潘江, 谢泽凯, Mumtaz Shahid, 和 Laura E. Barnes. 趋同范式：符号 AI 与连接主义 AI 在 LLM 赋能的自主代理中的协同作用, arXiv 预印本 arXiv:2407.08516, 2024. URLhttps://arxiv.org/abs/2407.08516v5.

[1185] 熊俊杰, 朱长嘉, 林树航, 张崇, 张永峰, 刘瑶, 和 李凌瑶. 不可见的提示，可见的威胁：大型语言模型外部资源中的恶意字体注入, arXiv 预印本 arXiv:2505.16957, 2025. URLhttps://arxiv.org/abs/2505.16957v1.

[1186] 熊俊杰, 蔡宇君, 胡锦阳, 彭南云, 李哲成, 和 王怡伟. 通过分而治之增强 llm 字符级操控. 2025.

[1187] 熊俊杰, 蔡宇君, 李哲成, 和 王怡伟. 揭示 llm 的思维：基于图的推理 llm 分析. 2025.

[1188] 熊俊杰, 蔡宇君, 李哲成, 和 王怡伟. 揭示扩散大型语言模型在可控生成中的潜力. 2025.

[1189] Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. 如何内存管理影响 LLM 代理：对经验跟随行为的实证研究，arXiv preprint arXiv:2505.16067，2025。URLhttps://arxiv.org/abs/2505.16067v1.

[1190] Chunmei Xu, Shengheng Liu, Cheng Zhang, Yongming Huang, Zhaohua Lu, 和 Luxi Yang. 基于多智能体强化学习的协同云边系统中的分布式传输。*IEEE Transactions onVehicularTechnology*, 2021。

[1191] Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, 等。Redstar：扩展长文本数据是否解锁更好的慢推理系统？2025。

[1192] 徐红森, 朱苏, 王志涵, 郑航, 马达, 曹瑞生, 范帅, 陈路, 和余凯. 通过可靠性对齐减少工具幻觉, arXiv预印本 arXiv:2412.04141, 2024. URL https://arxiv.org/abs/2412.04141v3.

[1193] 胡旭，Gargi Ghosh，Po-Yao（Bernie）黄，Dmytro Okhonko，Armen Aghajanyan，以及Florian Metze Luke Zettlemoyer Christoph Feichtenhofer。Videoclip：用于零样本视频-文本理解的对比预训练。自然语言处理经验方法会议，2021。

徐梦嘉. 理解图嵌入方法及其应用. *SIAM Review*, 2020.

[1195] Minrui Xu, Hongyang Du, Dusist Niyato, Jiawen Kang, Zehui Xiong, Shiwen Mao, Zhu Han, A. Jamalipour, Dong In Kim, X. Shen, Victor C. M. Leung, and H. Poor. Unleashing the power of edge-cloud generative ai in mobile networks: A survey of aigc services. *IEEE Communications Surveys and Tutorials*, 2023.

[1196] Minrui Xu, D. Niyato, and Christopher G. Brinton. Serving long-context llms at the mobile edge: Test-time reinforcement learning-based model caching and inference offloading, arXiv preprint arXiv:2501.14205, 2025. URL https://arxiv.org/abs/2501.14205v1.

[1197] Nan Xu, Fei Wang, Sheng Zhang, Hoifung Poon, and Muhao Chen. From introspection to best practices: Principled analysis of demonstrations in multimodal in-context learning. *North American Chapter of the Association for Computational Linguistics*, 2024.

[1198] Shuhang Xu and Fangwei Zhong. Comet: Metaphor-driven covert communication for multi-agent language games, arXiv preprint arXiv:2505.18218, 2025. URL https://arxiv.org/abs/2505.18218v1.

[1199] Tianyang Xu, Haojie Zheng, Chengze Li, Haoxiang Chen, Yixin Liu, Ruoxi Chen, and Lichao Sun. Noderag: Structuring graph-based rag with heterogeneous nodes, arXiv preprint arXiv:2504.11544, 2025. URL https://arxiv.org/abs/2504.11544v1.

[1200] Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and W. Wang. Pride and prejudice: Llm amplifies self-bias in self-refinement. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[1201] Wenrui Xu and Keshab K. Parhi. A survey of attacks on large language models, arXiv preprint arXiv:2505.12567, 2025. URL https://arxiv.org/abs/2505.12567v1.

[1202] Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. arXiv preprint, 2025.

[1203] Wujiang Xu, Kai Mei, Hang Gao, Juntao Tan, Zujie Liang, and Yongfeng Zhang. A-mem: Agentic memory for llm agents, arXiv preprint arXiv:2502.12110, 2025. URL https://arxiv.org/abs/2502.12110.

[1204] Yifei Xu, Jingqiao Zhang, Ru He, Liangzhu Ge, Chao Yang, Cheng Yang, and Ying Wu. Sas: Self-augmentation strategy for language model pre-training. *AAAI Conference on Artificial Intelligence*, 2021.

[1205] Zhe Xu, Daoyuan Chen, Zhenqing Ling, Yaliang Li, and Ying Shen. Mindgym: What matters in question synthesis for thinking-centric fine-tuning?, arXiv preprint arXiv:2503.09499, 2025. URL https://arxiv.org/abs/2503.09499v2.

[1206] Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. Retrieval-augmented generation with knowledge graphs for customer service question answering. *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

[1207] Eric Xue, Ke Chen, Zeyi Huang, Yuyang Ji, Yong Jae Lee, and Haohan Wang. Improve: Iterative model pipeline refinement and optimization leveraging llm experts, arXiv preprint arXiv:2502.18530, 2025. URL https://arxiv.org/abs/2502.18530v2.

[1208] Huiyin Xue and Nikolaos Aletras. Pit one against many: Leveraging attention-head embeddings for parameter-efficient multi-head attention. *Conference on Empirical Methods in Natural Language Processing*, 2023.

[1196] 徐敏睿, Niyato D., 和 Christopher G. Brinton. 在移动边缘服务长上下文 llms: 基于测试时强化学习的模型缓存和推理卸载, arXiv 预印本 arXiv:2501.14205, 2025. URLhttps://arxiv.org/abs/2501.14205v1.

[1197] 徐楠, 王飞, 张胜, Poon Hoifung, 和 陈穆浩. 从内省到最佳实践: 多模态情境学习中的演示原则性分析. 美国计算语言学协会北美分会,2024.

[1198] 徐树航 和 钟方伟. Comet: 驱动隐喻的隐蔽通信用于多智能体语言游戏, arXiv 预印本 arXiv:2505.18218, 2025. URLhttps://arxiv.org/abs/2505.18218v1.

[1199] 徐天阳, 郑浩杰, 李成哲, 陈浩翔, 刘依欣, 陈若曦, 和孙立超. Noderag: 基于异构节点的图结构rag构建, arXiv预印本arXiv:2504.11544, 2025. URLhttps://arxiv.org/abs/2504.11544v1.

[1200] 徐文达, 朱国磊, 赵宣东, 潘良明, 李雷, 和W. 王 Pride and prejudice: llm在自我完善中放大自我偏见. 计算语言学协会年会, 2024.

[1201] Wenrui Xu and Keshab K. Parhi. A survey of attacks on large language models, arXiv preprint arXiv:2505.12567, 2025. URL https://arxiv.org/abs/2505.12567v1.

[1202] 徐武江, 梁祖杰, 梅凯, 高航, 谭俊涛, 和张永峰. A-mem: 为llm代理的智能记忆. arXiv预印本, 2025.

[1203] 徐武江, 梅凯, 高航, 谭俊涛, 梁祖杰, 和张永峰. A-mem: 为llm代理的智能记忆, arXiv预印本 arXiv:2502.12110, 2025. URL https://arxiv.org/abs/2502.12110.

[1204] 徐一飞, 张静桥, 何如, 葛良珠, 杨超, 杨成, 和吴颖. Sas: 语言模型预训练的自增强策略. *AAAI*人工智能会议, 2021.

[1205] 徐哲, 陈道远, 凌振清, 李亚亮, 和沈颖. Mindgym: 思维中心微调中问题合成的重要性?, arXiv预印本 arXiv:2503.09499, 2025. URLhttps://arxiv.org/abs/2503.09499v2.

[1206] 徐振涛, Mark Jerome Cruz, Matthew Guevara, 王铁, Manasi Deshpande, 王晓峰, 和李铮. 基于知识图谱的检索增强生成用于客户服务问题回答. 年度国际*ACM SIGIR*信息检索研究与发展会议, 2024.

[1207] 徐Eric, 陈Ke, 黄Zeyi, 季Yuyang, 李Yong Jae, 和王Haohan. 提高：利用llm专家进行迭代模型管道细化和优化, arXiv预印本arXiv:2502.18530, 2025. URLhttps://arxiv.org/abs/2502.18530v2.

[1208] Huiyin Xue 和 Nikolaos Aletras. 以一敌众：利用注意力头嵌入实现参数高效的多头注意力机制。自然语言处理经验方法会议, 2023.

[1209] Xiangyuan Xue, Zeyu Lu, Di Huang, Zidong Wang, Wanli Ouyang, and Lei Bai. Comfybench: Benchmarking llm-based agents in comfyui for autonomously designing collaborative ai systems, arXiv preprint arXiv:2409.01392, 2024. URL https://arxiv.org/abs/2409.01392v2.

[1210] Bingyu Yan, Xiaoming Zhang, Litian Zhang, Lian Zhang, Ziyi Zhou, Dezhuang Miao, and Chaozhuo Li. Beyond self-talk: A communication-centric survey of llm-based multi-agent systems. arXiv preprint, 2025.

[1211] Tianqiang Yan and Tiansheng Xu. Refining the responses of llms by themselves, arXiv preprint arXiv:2305.04039, 2023. URL https://arxiv.org/abs/2305.04039v1.

[1212] Xu Yan, Junliang Du, Lun Wang, Yingbin Liang, Jiacheng Hu, and Bingxing Wang. The synergistic role of deep learning and neural architecture search in advancing artificial intelligence. *2024 International Conference on Electronics and Devices, Computational Science (ICEDCS)*, 2024.

[1213] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. *The Web Conference*, 2023.

[1214] Yuchen Yan, Yongliang Shen, Yang Liu, Jin Jiang, Mengdi Zhang, Jian Shao, and Yueting Zhuang. Inftythink: Breaking the length limits of long-context reasoning in large language models, arXiv preprint arXiv:2503.06692, 2025. URL https://arxiv.org/abs/2503.06692v3.

[1215] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. *International Conference on Learning Representations*, 2023.

[1216] Hongkang Yang, Zehao Lin, Wenjin Wang, Hao Wu, Zhiyu Li, Bo Tang, Wenqiang Wei, Jinbo Wang, Zeyun Tang, Shichao Song, Chenyang Xi, Yu Yu, Kai Chen, Feiyu Xiong, Linpeng Tang, and E. Weinan. Memory3: Language modeling with explicit memory. *Journal of Machine Learning*, 2024.

[1217] Jianxin Yang. Longqlora: Efficient and effective method to extend context length of large language models, arXiv preprint arXiv:2311.04879, 2023. URL https://arxiv.org/abs/2311.04879v2.

[1218] Jinghan Yang, Shuming Ma, and Furu Wei. Auto-icl: In-context learning without human supervision. arXiv preprint, 2023.

[1219] Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Octopus: Embodied vision-language programmer from environmental feedback. *European Conference on Computer Vision*, 2023.

[1220] John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Adriano Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. Swe-agent: Agent-computer interfaces enable automated software engineering. *Neural Information Processing Systems*, 2024.

[1221] Junhan Yang, Zheng Liu, Shitao Xiao, Chaozhuo Li, Defu Lian, Sanjay Agrawal, Amit Singh, Guangzhong Sun, and Xing Xie. Graphformers: Gnn-nested transformers for representation learning on textual graph. *Neural Information Processing Systems*, 2021.

[1222] Ke Yang, Yao Liu, Sapana Chaudhary, Rasool Fakoor, Pratik Chaudhari, George Karypis, and Huzefa Rangwala. Agentoccam: A simple yet strong baseline for llm-based web agents. 2024.

[1209] 向远学，陆泽宇，黄迪，王梓东，欧阳万利，白雷. Comfybench: 在Comfyui中测试基于llm的代理，用于自主设计协作式人工智能系统，arXiv预印本arXiv:2409.01392，2024年. URLhttps://arxiv.org/abs/2409.01392v2.

[1210] Bingyu Yan, Xiaoming Zhang, Litian Zhang, Lian Zhang, Ziyi Zhou, Dezhuang Miao, and Chaozhuo Li. 超越自我对话：以通信为中心的基于LLM的多智能体系统综述。arXiv预印本，2025.

[1211] 严天强和徐天盛. llms自我优化回答, arXiv预印本arXiv:2305.04039, 2023. URLhttps://arxiv.org/abs/2305.04039v1.

[1212] 徐岩, 杜俊良, 王伦, 梁英斌, 胡家成, 和王兵兴. 深度学习与神经架构搜索在推进人工智能中的协同作用. *2024* 国际电子与器件会议、计算科学 *(ICEDCS)*, 2024.

[1213] 闫奕博, 温豪明, 钟思儒, 陈伟, 陈浩东, 温清松, Roger Zimmermann, 和梁宇轩. Urbanclip: 从网络中通过对比语言-图像预训练学习文本增强的城市区域表征. *TheWebConference*, 2023.

[1214] 闫宇辰, 沈永亮, 刘杨, 蒋金, 张梦迪, 邵健, 和庄宇庭. Inftythink: 打破大语言模型长上下文推理的长度限制, arXiv 预印本 arXiv:2503.06692, 2025. URLhttps://arxiv.org/abs/2503.06692v3.

[1215] 杨成润, 王学志, 陆一峰, 刘汉晓, Le Quoc V., 周登, 陈新云. 大型语言模型作为优化器. 国际学习表征会议, 2023.

[1216] 杨宏康, 林泽浩, 王文进, 吴浩, 李志宇, 唐波, 魏文强, 王金波, 唐泽云, 宋时超, 西陈阳, 余宇, 陈凯, 邢飞宇, 唐林鹏, 和E. Weinan. Memory3: 带显式记忆的语言建模. 机器学习杂志, 2024.

[1217] 杨建新. Longqlora: 扩展大型语言模型上下文长度的有效方法, arXiv预印本 arXiv:2311.04879, 2023. URL https://arxiv.org/abs/2311.04879v2.

杨景涵、马树铭和魏福儒. Auto-icl：无需人工监督的情境学习. arXiv预印本, 2023.

[1219] 杨景康，董宇浩，刘帅，李波，王梓越，蒋晨程，谭浩然，康嘉木，张元汉，周凯阳，刘子伟。Octopus：基于环境反馈的具身视觉语言程序员。欧洲计算机视觉会议，2023。

杨强，卡洛斯·E·希门内斯，亚历山大·韦蒂格，基利安·阿德里亚诺·利雷特，姚顺宇，卡尔蒂克·纳拉辛汉，和奥菲尔·普雷斯。Swe-agent：代理-计算机界面实现自动化软件工程。《神经信息处理系统》，2024。

杨俊涵，刘正，肖时涛，李超卓，连德福，Sanjay Agrawal，Amit Singh，孙光中，谢行. Graphformers：用于文本图表示学习的GNN嵌套Transformer. 神经信息处理系统, 2021.

[1222] 柯杨, 刘瑶, Sapana Chaudhary, Rasool Fakoor, Pratik Chaudhari, George Karypis, 和 Huzefa Rangwala. Agentoccam: 一个简单但强大的基于 llm 的网络代理基线. 2024.

[1223] Lin F. Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[1224] Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models, arXiv preprint arXiv:2505.15809v1, 2025. URL https://arxiv.org/abs/2505.15809v1.

[1225] R Yang, L Song, Y Li, S Zhao, and Y Ge.... Gpt4tools: Teaching large language model to use tools via self-instruction. 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e393677793767624f2821cec8bdd02f1-Abstract-Conference.html?utm_campaign=Artificial%2BIntelligence%2BWeekly&utm_medium=email&utm_source=Artificial_Intelligence_Weekly_411.

[1226] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. Gpt4tools: Teaching large language model to use tools via self-instruction. *Neural Information Processing Systems*, 2023.

[1227] Shang Yang, Junxian Guo, Haotian Tang, Qinghao Hu, Guangxuan Xiao, Jiaming Tang, Yujun Lin, Zhijian Liu, Yao Lu, and Song Han. Lserve: Efficient long-sequence llm serving with unified sparse attention, arXiv preprint arXiv:2502.14866, 2025. URL https://arxiv.org/abs/2502.14866v2.

[1228] Shanglong Yang, Zhipeng Yuan, Shunbao Li, Ruoling Peng, Kang Liu, and Po Yang. Gpt-4 as evaluator: Evaluating large language models on pest management in agriculture. arXiv preprint, 2024.

[1229] Wang Yang, Zirui Liu, Hongye Jin, Qingyu Yin, Vipin Chaudhary, and Xiaotian Han. Longer context, deeper thinking: Uncovering the role of long-context ability in reasoning, arXiv preprint arXiv:2505.17315, 2025. URL https://arxiv.org/abs/2505.17315v1.

[1230] Wen Yang, Kai Fan, and Minpeng Liao. Markov chain of thought for efficient mathematical reasoning. *North American Chapter of the Association for Computational Linguistics*, 2024.

[1231] Yaodong Yang, Chengdong Ma, Zihan Ding, S. McAleer, Chi Jin, and Jun Wang. Game-theoretic multiagent reinforcement learning, arXiv preprint arXiv:2011.00583, 2020. URL https://arxiv.org/abs/2011.00583v4.

[1232] Yazheng Yang, Yuqi Wang, Sankalok Sen, Lei Li, and Qi Liu. Unleashing the potential of large language models for predictive tabular tasks in data science, arXiv preprint arXiv:2403.20208, 2024. URL https://arxiv.org/abs/2403.20208v7.

[1233] Yi Yang, Yixuan Tang, and Kar Yan Tam. Investlm: A large language model for investment using financial domain instruction tuning, arXiv preprint arXiv:2309.13064, 2023. URL https://arxiv.org/abs/2309.13064.

[1234] Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. G-daug: Generative data augmentation for commonsense reasoning. *Findings*, 2020.

[1235] Yingxuan Yang, Huacan Chai, Yuanyi Song, Siyuan Qi, Muning Wen, Ning Li, Junwei Liao, Haoyi Hu, Jianghao Lin, Gaowei Chang, Weiwen Liu, Ying Wen, Yong Yu, and Weinan Zhang. A survey of ai agent protocols, arXiv preprint arXiv:2504.16736, 2025. URL https://arxiv.org/abs/2504.16736v3.

[1223] 林馥娜, 陈红阳, 李兆, 丁晓, 和 吴欣东. 给我们事实：使用知识图谱增强大型语言模型以实现事实感知语言建模. *IEEE Transactionson Knowledge and DataEngineering*, 2023.

[1224] 杨玲, 田叶, 李 Bowen, 张欣辰, 沈可, 邓云海, 和 王梦迪. Mmada: 多模态大型扩散语言模型, arXiv 预印本 arXiv:2505.15809v1, 2025. URLhttps://arxiv.org/abs/2505.15809v1.

[1225] 杨瑞, 宋丽, 李岩, 赵思, 和 葛岩. ... Gpt4tools: 通过自我指导教大型语言模型使用工具. 2023. URLhttps://proceedings.neurips.cc/paper_files/paper/2023/hash/e393677793767624f2821cec8bdd02f1-Abstract-Conference.html?utm_campaign=Artificial%2BIntelligence%2BWeekly&utm_medium=email&utm_source=Artificial_Intelligence_Weekly_411.

[1226] 阮杨, 宋琳, 李延伟, 赵思捷, 葛亦潇, 李秀, 和山英. Gpt4tools: 通过自我指导让大型语言模型使用工具. 神经信息处理系统, 2023.

[1227] 商阳, 郭俊贤, 唐浩天, 胡庆浩, 肖广轩, 唐嘉明, 林宇君, 刘志坚, 陆瑶, 韩松. Lserve: 统一稀疏注意力的高效长序列llm服务, arXiv预印本 arXiv:2502.14866, 2025. URL https://arxiv.org/abs/2502.14866v2.

[1228] 杨尚龙, 袁志鹏, 李顺宝, 彭若玲, 刘康, 和杨波. Gpt-4作为评估器: 在农业病虫害管理上评估大型语言模型. arXiv预印本, 2024.

[1229] 王洋, 刘子睿, 金红叶, 尹清宇, Vipin Chaudhary, 和 韩晓天. 更长的上下文，更深的思考：揭示长上下文能力在推理中的作用, arXiv 预印本 arXiv:2505.17315, 2025. URLhttps://arxiv.org/abs/2505.17315v1.

[1230] 杨文, KaiFan, 和 廖明鹏. 马尔可夫思维链用于高效的数学推理.美国计算语言学协会北美分会, 2024.

[1231] 杨亚东, 马成东, 丁志涵, S. McAleer, 金池, 和 王军. 基于博弈论的多智能体强化学习, arXiv 预印本 arXiv:2011.00583, 2020. URL https://arxiv.org/abs/2011.00583v4.

[1232] 杨亚争，王宇琪，森桑卡拉，李雷，刘奇。释放大型语言模型在数据科学中的预测表格任务潜力，arXiv 预印本 arXiv:2403.20208，2024年。URL https://arxiv.org/abs/2403.20208v7。

[1233] 易阳，唐亦璇，和谭家岩。Investlm：一个使用金融领域指令微调的投资大语言模型，arXiv预印本arXiv:2309.13064，2023年。URLhttps://arxiv.org/abs/2309.13064。

杨一本，查蒂亚·马拉维亚，贾雷德·费尔南德斯，斯瓦哈·斯瓦米迪普塔，罗南·勒·布拉，王继平，钱德拉·巴加瓦图拉，赵叶金，以及道格·唐尼。G-daug：常识推理的生成数据增强。发现，2020。

[1235] Yingxuan Yang, Huacan Chai, Yuanyi Song, Siyuan Qi, Muning Wen, Ning Li, Junwei Liao, Haoyi Hu, Jianghao Lin, Gaowei Chang, Weiwen Liu, Ying Wen, Yong Yu, and Weinan Zhang. A survey of ai agent protocols, arXiv preprint arXiv:2504.16736, 2025. URL https://arxiv.org/abs/2504.16736v3.

[1236] Yuan Yang, Siheng Xiong, Ehsan Shareghi, and F. Fekri. The compressor-retriever architecture for language model os, arXiv preprint arXiv:2409.01495, 2024. URL https://arxiv.org/abs/2409.01495v1.

[1237] Yuxin Yang, Haoyang Wu, Tao Wang, Jia Yang, Hao Ma, and Guojie Luo. Pseudo-knowledge graph: Meta-path guided retrieval and in-graph text for rag-equipped llm, arXiv preprint arXiv:2503.00309, 2025. URL https://arxiv.org/abs/2503.00309v1.

[1238] Zhen Yang, Fang Liu, Zhongxing Yu, J. Keung, Jia Li, Shuo Liu, Yifan Hong, Xiaoxue Ma, Zhi Jin, and Ge Li. Exploring and unleashing the power of large language models in automated code translation. *Proc. ACM Softw. Eng.*, 2024.

[1239] Chengyuan Yao and Satoshi Fujita. Adaptive control of retrieval-augmented generation for large language models through reflective tags. *Electronics*, 2024.

[1240] Huaiyuan Yao, Longchao Da, Vishnu Nandam, J. Turnau, Zhiwei Liu, Linsey Pang, and Hua Wei. Comal: Collaborative multi-agent large language models for mixed-autonomy traffic, arXiv preprint arXiv:2410.14368, 2024. URL https://arxiv.org/abs/2410.14368v2.

[1241] Jiayu Yao, Shenghua Liu, Yiwei Wang, Lingrui Mei, Baolong Bi, Yuyao Ge, Zhecheng Li, and Xueqi Cheng. Who is in the spotlight: The hidden bias undermining multimodal retrieval-augmented generation. 2025.

[1242] Jinghan Yao, Sam Ade Jacobs, Masahiro Tanaka, Olatunji Ruwase, A. Shafi, H. Subramoni, and Dhabaleswar K. Panda. Training ultra long context language model with fully pipelined distributed transformer, arXiv preprint arXiv:2408.16978, 2024. URL https://arxiv.org/abs/2408.16978v2.

[1243] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. *AAAI Conference on Artificial Intelligence*, 2018.

[1244] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Neural Information Processing Systems*, 2022.

[1245] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *International Conference on Learning Representations*, 2022.

[1246] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, T. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Neural Information Processing Systems*, 2023.

[1247] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, arXiv preprint arXiv:2210.03629, 2023. URL https://arxiv.org/abs/2210.03629.

[1248] Shunyu Yao, Noah Shinn, P. Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. arXiv preprint, 2024.

[1236] 袁洋, 熊时恒, Shareghi Ehsan, 和 F. Fekri. 语言模型的压缩器-检索器架构 os, arXiv 预印本 arXiv:2409.01495, 2024. URLhttps://arxiv.org/abs/2409.01495v1.

[1237] 杨宇欣, 吴浩阳, 王涛, 杨嘉, 马浩, 和 罗国杰. 伪知识图谱：元路径引导检索和图内文本为配备 RAG 的大语言模型, arXiv 预印本 arXiv:2503.00309, 2025. URLhttps://arxiv.org/abs/2503.00309v1.

[1238] 杨振, 刘方, 余中行, Keung J., 李佳, 刘硕, 韩一帆, 马晓雪, 金智, 和 李格. 在自动代码翻译中探索和释放大语言模型的力量.*ACM* 软件工程会议录, 2024.

[1239] 程远姚和藤田智。通过反射标签对大型语言模型的检索增强生成进行自适应控制。电子学, 2024。

[1240] 姚怀远, 大龙超, Vishnu Nandam, J. Turnau, 刘志伟, Pang Linsey, 和 魏华. Comal: 用于混合自主交通的协作多智能体大型语言模型, arXiv 预印本 arXiv:2410.14368, 2024. URLhttps://arxiv.org/abs/2410.14368v2.

[1241] 姚嘉宇, 刘胜华, 王怡伟, 梅凌瑞, 毕宝龙, 郭雨瑶, 李哲成, 和 程雪琪。谁在聚光灯下：损害多模态检索增强生成的隐藏偏见。2025。

[1242] 姚景涵, Sam Ade Jacobs, 藤田正弘, Ruwase Olatunji, A. Shafi, H. Subramoni, 和 Dhabaleswar K. Panda. 使用完全流水线分布式 Transformer 训练超长上下文语言模型, arXiv 预印本 arXiv:2408.16978, 2024. URLhttps://arxiv.org/abs/2408.16978v2.

[1243] 梁瑶, 唐盛, 和 罗元. 用于文本分类的图卷积网络。*AAAI*人工智能会议, 2018。

[1244] 姚顺宇, 陈浩, 杨约翰, 和 Narasimhan Karthik. Webshop: 基于语言代理的可扩展现实世界网络交互。 神经信息处理系统会议, 2022。

[1245] 姚顺宇, 赵建宇, 余狄, 杜南, Shafran Izhak, Narasimhan Karthik, 和 曹元. React: 在语言模型中协同推理和行动。 国际学习表示会议, 2022。

[1246] 姚顺宇, 余狄, 赵建宇, Shafran Izhak, Griffiths T., 曹元, 和 Narasimhan Karthik. 思维树: 使用大型语言模型的深思熟虑问题解决。 神经信息处理系统会议, 2023。

[1247] 孙宇，赵建平，于迪安，杜南，Izhak Shafran，Karthik Narasimhan，和曹元。React：在语言模型中协同推理和行动，arXiv预印本arXiv:2210.03629，2023年。URL https://arxiv.org/abs/2210.03629。

[1248] 孙宇，Noah Shinn，P. Razavi，和Karthik Narasimhan。 τ-bench：现实世界领域中工具-代理-用户交互的基准。arXiv预印本，2024年。

[1249] Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Murthy, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, and Silvio Savarese. Retroformer: Retrospective large language agents with policy gradient optimization, arXiv preprint arXiv:2308.02151, 2024. URL https://arxiv.org/abs/2308.02151.

[1250] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and J. Leskovec. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. *North American Chapter of the Association for Computational Linguistics*, 2021.

[1251] Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and J. Leskovec. Deep bidirectional language-knowledge graph pretraining. *Neural Information Processing Systems*, 2022.

[1252] Fahd Yazin, Moumita Das, A. Banerjee, and Dipanjan Roy. Contextual prediction errors reorganize naturalistic episodic memories in time. *Scientific Reports*, 2021.

[1253] J Ye, G Li, S Gao, C Huang, Y Wu, S Li, and X Fan.... Tooleyes: Fine-grained evaluation for tool learning capabilities of large language models in real-world scenarios. 2024. URL https://arxiv.org/abs/2401.00741.

[1254] J Ye, S Li, G Li, C Huang, S Gao, and Y Wu.... Toolsword: Unveiling safety issues of large language models in tool learning across three stages. 2024. URL https://arxiv.org/abs/2402.10753.

[1255] Junjie Ye, Zhengyin Du, Xuesong Yao, Weijian Lin, Yufei Xu, Zehui Chen, Zaiyuan Wang, Sining Zhu, Zhiheng Xi, Siyu Yuan, Tao Gui, Qi Zhang, Xuanjing Huang, and Jiechao Chen. Toolhop: A query-driven benchmark for evaluating large language models in multi-hop tool use, arXiv preprint arXiv:2501.02506, 2025. URL https://arxiv.org/abs/2501.02506v4.

[1256] Qinyuan Ye, Maxamed Axmed, Reid Pryzant, and Fereshte Khani. Prompt engineering a prompt engineer. *Annual Meeting of the Association for Computational Linguistics*, 2023.

[1257] Sixiang Ye, Zeyu Sun, Guoqing Wang, Liwei Guo, Qing-Lin Liang, Zheng Li, and Yong Liu. Prompt alchemy: Automatic prompt refinement for enhancing code generation, arXiv preprint arXiv:2503.11085, 2025. URL https://arxiv.org/abs/2503.11085v1.

[1258] Zhifan Ye, Kejing Xia, Yonggan Fu, Xin Dong, Jihoon Hong, Xiangchi Yuan, Shizhe Diao, Jan Kautz, Pavlo Molchanov, and Y. Lin. Longmamba: Enhancing mamba's long-context capabilities via training-free receptive field enlargement. *International Conference on Learning Representations*, 2025.

[1259] Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. Survey on evaluation of llm-based agents, arXiv preprint arXiv:2503.16416, 2025. URL https://arxiv.org/abs/2503.16416v1.

[1260] Peiling Yi and Yuhan Xia. Irony detection, reasoning and understanding in zero-shot learning. *IEEE Transactions on Artificial Intelligence*, 2025.

[1261] Da Yin, Faeze Brahman, Abhilasha Ravichander, Khyathi Raghavi Chandu, Kai-Wei Chang, Yejin Choi, and Bill Yuchen Lin. Agent lumos: Unified and modular training for open-source language agents. *Annual Meeting of the Association for Computational Linguistics*, 2023.

[1249] 姚巍然，何谢莉，胡安·卡洛斯·涅布勒斯，刘志伟，冯一浩，薛乐，里特什·穆尔蒂，陈泽远，张建国，阿维什·阿普里特，徐然，梅菲尔，王欢，熊才明，以及西尔维奥·萨瓦雷斯。Retroformer：具有策略梯度优化的回顾式大型语言智能体，arXiv 预印本 arXiv:2308.02151，2024年。URLhttps://arxiv.org/abs/2308.02151。

[1250] Yasunaga Michihiro, Ren Hongyu, Bosselut Antoine, Liang Percy, 和 J. Leskovec. Qa-gnn: 使用语言模型和知识图谱进行问答推理。美国计算语言学协会北美分会, 2021.

[1251] Yasunaga Michihiro, Bosselut Antoine, Ren Hongyu, Zhang Xikun, Christopher D. Manning, Liang Percy, and J.Leskovec. 深度双向语言知识图谱预训练. 神经信息处理系统, 2022.

[1252] Fahd Yazin, Moumita Das, A. Banerjee, and Dipanjan Roy. 上下文预测误差重组时间内的自然主义情景记忆。*ScientificReports*, 2021。

[1253] J Ye, G Li, S Gao, C Huang, Y Wu, S Li, and X Fan. .. . Tooleyes: 评估大型语言模型在现实场景中工具学习能力的细粒度方法。 2024. URL https://arxiv.org/abs/2401.00741.

[1254] J Ye, S Li, G Li, C Huang, S Gao, and Y Wu.... Toolsword: Unveiling safety issues of large language models in tool learning across three stages. 2024. URL https://arxiv.org/abs/2402.10753.

[1255] Junjie Ye, Zhengyin Du, Xuesong Yao, Weijian Lin, Yufei Xu, Zehui Chen, Zaiyuan Wang, Sining Zhu, Zhiheng Xi, Siyu Yuan, Tao Gui, Qi Zhang, Xuanjing Huang, and Jiechao Chen. Toolhop: 一个基于查询的多跳工具使用评估大型语言模型的基准，arXiv 预印本 arXiv:2501.02506，2025。URLhttps://arxiv.org/abs/2501.02506v4.

[1256] 叶勤元, 马克萨迈德·阿克迈德, 里德·普兰赞特, 和 费尔希特·哈尼. 提示工程：提示工程师的提示工程. 计算语言学协会年会, 2023.

[1257] 叶思想, 孙泽宇, 王国清, 郭立伟, 梁清林, 李铮, 和 刘勇. 提示炼金术：用于增强代码生成的自动提示优化, arXiv 预印本 arXiv:2503.11085, 2025. URLhttps://arxiv.org/abs/2503.11085v1.

[1258] 叶志凡, 夏克敬, 傅永甘, 董欣, 韩基雄, 袁祥池, 肖世哲, 卡茨·扬, 摩尔恰诺夫·帕夫洛, 和 林怡. 长毛蚱：通过无训练感受野扩大增强毛蚱的长上下文能力. 学习表示国际会议, 2025.

[1259] Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. 大型语言模型代理的评估调查，arXiv 预印本 arXiv:2503.16416，2025。URL https://arxiv.org/abs/2503.16416v1.

[1260] Peiling Yi and Yuhan Xia. 零样本学习中的讽刺检测、推理和理解。 *IEEETransactions on Artificial Intelligence*, 2025。

[1261] Da Yin, Faeze Brahman, Abhilasha Ravichander, Khyathi Raghavi Chandu, Kai-Wei Chang, Yejin Choi, and Bill Yuchen Lin。代理 lumos：开源语言代理的统一和模块化训练。*Association for Computational Linguistics* 年会, 2023。

[1262] Fan Yin, Zifeng Wang, I-Hung Hsu, Jun Yan, Ke Jiang, Yanfei Chen, Jindong Gu, Long T. Le, Kai-Wei Chang, Chen-Yu Lee, Hamid Palangi, and Tomas Pfister. Magnet: Multi-turn tool-use data synthesis and distillation via graph translation, arXiv preprint arXiv:2503.07826, 2025. URL https://arxiv.org/abs/2503.07826v1.

[1263] Guoli Yin, Haoping Bai, Shuang Ma, Feng Nan, Yanchao Sun, Zhaoyang Xu, Shen Ma, Jiarui Lu, Xiang Kong, Aonan Zhang, Dian Ang Yap, Yizhe Zhang, K. Ahnert, Vik Kamath, Mathias Berglund, Dominic Walsh, Tobias Gindele, Juergen Wiest, Zhengfeng Lai, Xiaoming Wang, Jiulong Shan, Meng Cao, Ruoming Pang, and Zirui Wang. Mmau: A holistic benchmark of agent capabilities across diverse domains. *North American Chapter of the Association for Computational Linguistics*, 2024.

[1264] Pengcheng Yin, Graham Neubig, Wen tau Yih, and Sebastian Riedel. Tabert: Pretraining for joint understanding of textual and tabular data. *Annual Meeting of the Association for Computational Linguistics*, 2020.

[1265] S Yin, W You, Z Ji, G Zhong, and J Bai. Mumath-code: Combining tool-use large language models with multi-perspective data augmentation for mathematical reasoning. 2024. URL https://arxiv.org/abs/2405.07551.

[1266] Gunwoo Yong, Kahyun Jeon, Daeyoung Gil, and Ghang Lee. Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. *Comput. Aided Civ. Infrastructure Eng.*, 2022.

[1267] Aspen H. Yoo and A. Collins. How working memory and reinforcement learning are intertwined: A cognitive, neural, and computational perspective. *Journal of Cognitive Neuroscience*, 2021.

[1268] Chanwoong Yoon, Taewhoo Lee, Hyeon Hwang, Minbyul Jeong, and Jaewoo Kang. Compact: Compressing retrieved documents actively for question answering. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[1269] Jiaxuan You, Mingjie Liu, Shrimai Prabhumoye, M. Patwary, M. Shoeybi, and Bryan Catanzaro. Llm-evolve: Evaluation for llm's evolving capability on benchmarks. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[1270] Yuxin You, Zhen Liu, Xiangchao Wen, Yongtao Zhang, and Wei Ai. Large language models meet graph neural networks: A perspective of graph mining. *Mathematics*, 2024.

[1271] Dian Yu, Yuheng Zhang, Jiahao Xu, Tian Liang, Linfeng Song, Zhaopeng Tu, Haitao Mi, and Dong Yu. Teaching llms to refine with tools, arXiv preprint arXiv:2412.16871, 2024. URL https://arxiv.org/abs/2412.16871v1.

[1272] Miao Yu, Fanci Meng, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pang, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, Bo An, and Qingsong Wen. A survey on trustworthy llm agents: Threats and countermeasures, arXiv preprint arXiv:2503.09648, 2025. URL https://arxiv.org/abs/2503.09648v1.

[1273] Ye Yu, Yaoning Yu, and Haohan Wang. Premise: Scalable and strategic prompt optimization for efficient mathematical reasoning in large models, arXiv preprint arXiv:2506.10716, 2025. URL https://arxiv.org/abs/2506.10716v1.

[1262] 范寅, 王紫峰, 许一鸿, 阎俊, 姜科, 陈岩飞, 顾晋东, 李隆泰, 张凯伟, 李陈宇, 哈米德·帕兰吉, 和托马斯·普菲斯特. Magnet: 基于图翻译的多轮工具使用数据合成与蒸馏, arXiv 预印本 arXiv:2503.07826, 2025. URLhttps://arxiv.org/abs/2503.07826v1.

[1263] 尹国利, 白海平, 马双, 难峰, 孙言超, 许昭阳, 马神, 陆嘉瑞, 孔翔, 张澳南, 叶典安·亚柏, 张易哲, K. Ahnert, 维克·卡玛斯, 马蒂亚斯·贝格伦德, 多米尼克·沃尔什, 托马斯·金德尔, 贾根·魏斯特, 赖正峰, 王小明, 山九隆, 曹梦, 庞若鸣, 王子睿. Mmau: 跨越多样化领域的智能体能力综合基准. 美国计算语言学协会北美分会, 2024.

[1264] 尹鹏程, Graham Neubig, 梁文韬, 和 Sebastian Riedel. Tabert: 预训练用于联合理解文本和表格数据. 计算语言学协会年会, 2020.

[1265] S Yin, W You, Z Ji, G Zhong, 和 J Bai. Mumath-code: 结合工具使用大型语言模型与多视角数据增强进行数学推理. 2024. URLhttps://arxiv.org/abs/2405.07551.

[1266] Yong Gunwoo, Jeon Kahyun, Gil Daeyoung, 和 Lee Ghang. 使用视觉语言预训练模型进行零样本和少样本缺陷检测与分类的提示工程. 计算辅助民用基础设施工程, 2022.

[1267] Yoo Aspen H. 和 A. Collins. 工作记忆和强化学习是如何交织的：认知、神经和计算视角. 认知神经科学杂志, 2021.

[1268] Chanwoong Yoon, Taewhoo Lee, Hyeon Hwang, Minbyul Jeong, and Jaewoo Kang. Compact: 主动压缩检索到的文档用于问答. 自然语言处理经验方法会议, 2024.

[1269] Jiaxuan You, Mingjie Liu, Shrimai Prabhumoye, M. Patwary, M. Shoeybi, and Bryan Catanzaro. Llm-evolve: 评估 LLM 在基准测试上的进化能力. 自然语言处理经验方法会议, 2024.

[1270] Yuxin You, Zhen Liu, Xiangchao Wen, Yongtao Zhang, and Wei Ai. 大型语言模型与图神经网络：图挖掘的视角. 数学, 2024.

[1271] Dian Yu, Yuheng Zhang, Jiahao Xu, Tian Liang, Linfeng Song, Zhaopeng Tu, Haitao Mi, and Dong Yu. 教会 LLM 使用工具, arXiv 预印本 arXiv:2412.16871, 2024. URL https://arxiv.org/abs/2412.16871v1.

[1272] 苗宇, 方茜, 周新云, 王世龙, 毛军元, 庞琳西, 陈天隆, 王坤, 李新峰, 张永峰, 安波, 文清松. 可信大语言模型代理的综述：威胁与对策, arXiv 预印本 arXiv:2503.09648, 2025. URLhttps://arxiv.org/abs/2503.09648v1.

[1273] 叶宇, 余亚宁, 王浩瀚. 前提：大规模和策略性提示优化，用于大模型中的高效数学推理, arXiv 预印本 arXiv:2506.10716, 2025. URLhttps://arxiv.org/abs/2506.10716v1.

[1274] Zeping Yu and Sophia Ananiadou. Understanding multimodal llms: the mechanistic interpretability of llava in visual question answering, arXiv preprint arXiv:2411.10950v2, 2024. URL https://arxiv.org/abs/2411.10950v2.

[1275] Zishun Yu, Tengyu Xu, Di Jin, Karthik Abinav Sankararaman, Yun He, Wenxuan Zhou, Zhouhao Zeng, Eryk Helenowski, Chen Zhu, Si-Yuan Wang, Hao Ma, and Han Fang. Think smarter not harder: Adaptive reasoning with inference aware optimization, arXiv preprint arXiv:2501.17974, 2025. URL https://arxiv.org/abs/2501.17974v2.

[1276] Zhao yu Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, and Yu Cheng. Openthinkimg: Learning to think with images via visual tool reinforcement learning, arXiv preprint arXiv:2505.08617, 2025. URL https://arxiv.org/abs/2505.08617v1.

[1277] Siyu Yuan, Zehui Chen, Zhiheng Xi, Junjie Ye, Zhengyin Du, and Jiecao Chen. Agent-r: Training language model agents to reflect via iterative self-training. arXiv preprint, 2025.

[1278] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. *International Conference on Machine Learning*, 2024.

[1279] Xiaowei Yuan, Zhao Yang, Ziyang Huang, Yequan Wang, Siqi Fan, Yiming Ju, Jun Zhao, and Kang Liu. Exploiting contextual knowledge in llms through v-usable information based layer enhancement. arXiv preprint, 2025.

[1280] Xinbin Yuan, Jian Zhang, Kaixin Li, Zhuoxuan Cai, Lujian Yao, Jie Chen, Enguang Wang, Qibin Hou, Jinwei Chen, Peng-Tao Jiang, and Bo Li. Enhancing visual grounding for gui agents via self-evolutionary reinforcement learning, arXiv preprint arXiv:2505.12370, 2025. URL https://arxiv.org/abs/2505.12370v2.

[1281] Murong Yue. A survey of large language model agents for question answering, arXiv preprint arXiv:2503.19213, 2025. URL https://arxiv.org/abs/2503.19213v1.

[1282] Xihang Yue, Linchao Zhu, and Yi Yang. Fragrel: Exploiting fragment-level relations in the external memory of large language models. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[1283] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J. Kim. Graph transformer networks. *Neural Information Processing Systems*, 2019.

[1284] Ge Yuyao, Cheng Yiting, Wang Jia, Zhou Hanlin, and Chen Lizhe. Vision transformer based on knowledge distillation in tcm image classification. In *2022 IEEE 5th International Conference on Computer and Communication Engineering Technology (CCET)*, pages 120–125. IEEE, 2022.

[1285] M. Zaheer, Guru Guruganesh, Kumar Avinava Dubey, J. Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. *Neural Information Processing Systems*, 2020.

[1286] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *International Journal of Computer Vision*, 2023.

[1287] E. Zelikman, Yuhuai Wu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning, arXiv preprint arXiv:2203.14465, 2022. URL https://arxiv.org/abs/2203.14465v2.

[1274] Zeping Yu and Sophia Ananiadou. 理解多模态大语言模型：视觉问答中 Llava 的机制可解释性，arXiv 预印本 arXiv:2411.10950v2，2024。URL https://arxiv.org/abs/2411.10950v2.

[1275] Zishun Yu, Tengyu Xu, Di Jin, Karthik Abinav Sankararaman, Yun He, Wenxuan Zhou, Zhouhao Zeng, Eryk Helenowski, Chen Zhu, Si-Yuan Wang, Hao Ma, and Han Fang. Think smarter not harder: Adaptive reasoning with inference aware optimization, arXiv preprint arXiv:2501.17974, 2025. URL https://arxiv.org/abs/2501.17974v2.

[1276] 赵宇苏，李林杰，宋明阳，郝云卓，杨正源，张俊，陈冠杰，顾佳伟，李俊涛，曲晓叶，程宇. Openthinkimg: 通过视觉工具强化学习学习用图像思考，arXiv预印本arXiv:2505.08617，2025. URLhttps://arxiv.org/abs/2505.08617v1.

[1277] 袁思宇，陈哲辉，席志恒，叶俊杰，杜正寅，陈杰超. Agent-r: 通过迭代自训练训练语言模型代理以进行反思. arXiv 预印本, 2025.

[1278] 魏哲，袁元哲，Cho Kyunghyun，Sainbayar Sukhbaatar，徐静，和 Jason E Weston。Self-rewarding language models。国际机器学习会议，2024。

[1279] 袁晓伟，杨兆，黄子阳，王逸群，范思琪，朱毅明，赵军，和 刘康。通过基于v-可用信息层的增强利用llms中的上下文知识。arXiv preprint，2025。

[1280] 袁新斌，张健，李凯欣，蔡卓璇，姚路建，陈杰，王恩广，侯启斌，陈金伟，蒋鹏涛，和 李波。通过自进化强化学习增强gui代理的视觉定位，arXiv preprint arXiv:2505.12370，2025。URLhttps://arxiv.org/abs/2505.12370v2。

[1281] Murong Yue. 大型语言模型问答代理的调查，arXiv 预印本 arXiv:2503.19213，2025。URLhttps://arxiv.org/abs/2503.19213v1。

[1282] Xihang Yue, Linchao Zhu, 和 Yi Yang. Fragrel: 利用大型语言模型外部内存中的片段级关系。计算语言学协会年度会议, 2024。

[1283] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, 和 Hyunwoo J. Kim. 图形转换网络。 神经信息处理系统, 2019。

[1284] Ge Yuyao, Cheng Yiting, Wang Jia, Zhou Hanlin, 和 Chen Lizhe. 基于知识蒸馏的 tcm 图像分类视觉转换器。在 *2022 IEEE* 第 *5* 届计算机与通信工程技术会议 *(CCET)*，第 120–125 页。IEEE, 2022。

[1285] M. Zaheer, Guru Guruganesh, Kumar Avinava Dubey, J. Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, 和 Amr Ahmed. Big Bird: 用于更长序列的 Transformer. 神经信息处理系统, 2020.

[1286] 张宇航、李伟、韩俊、周凯阳和罗晨昌。基于多模态大语言模型的情境目标检测。《国际计算机视觉杂志》，2023年。

[1287] E. Zelikman, Yuhuai Wu, and Noah D. Goodman. Star: 用推理进行推理引导，arXiv 预印本 arXiv:2203.14465，2022。URLhttps://arxiv.org/abs/2203.14465v2.

[1288] E. Zelikman, Eliana Lorch, Lester Mackey, and A. Kalai. Self-taught optimizer (stop): Recursively self-improving code generation, arXiv preprint arXiv:2310.02304, 2023. URL https://arxiv.org/abs/2310.02304v3.

[1289] Pai Zeng, Zhenyu Ning, Jieru Zhao, Weihao Cui, Mengwei Xu, Liwei Guo, XuSheng Chen, and Yizhou Shan. The cap principle for llm serving: A survey of long-context large language model serving, arXiv preprint arXiv:2405.11299, 2024. URL https://arxiv.org/abs/2405.11299v2.

[1290] Ruihong Zeng, Jinyuan Fang, Siwei Liu, and Zaiqiao Meng. On the structural memory of llm agents, arXiv preprint arXiv:2412.15266, 2024. URL https://arxiv.org/abs/2412.15266v1.

[1291] Yirong Zeng, Xiao Ding, Yuxian Wang, Weiwen Liu, Wu Ning, Yutai Hou, Xu Huang, Bing Qin, and Ting Liu. itool: Reinforced fine-tuning with dynamic deficiency calibration for advanced tool use, arXiv preprint arXiv:2501.09766, 2025. URL https://arxiv.org/abs/2501.09766v4.

[1292] Yongcheng Zeng, Xinyu Cui, Xuanfa Jin, Guoqing Liu, Zexu Sun, Dong Li, Ning Yang, Jianye Hao, Haifeng Zhang, and Jun Wang. Evolving llms' self-refinement capability via iterative preference optimization, arXiv preprint arXiv:2502.05605, 2025. URL https://arxiv.org/abs/2502.05605v3.

[1293] An Zhang, Leheng Sheng, Yuxin Chen, Hao Li, Yang Deng, Xiang Wang, and Tat-Seng Chua. On generative agents in recommendation. *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

[1294] B Zhang, K Zhou, X Wei, and X Zhao. . . . Evaluating and improving tool-augmented computation-intensive math reasoning. 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4a47dd69242d5af908cdd5d51c971cbf-Abstract-Datasets_and_Benchmarks.html.

[1295] Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Xiangyu Zhao, Yan Gao, Yao Hu, and Enhong Chen. Notellm-2: Multimodal large representation models for recommendation. *Knowledge Discovery and Data Mining*, 2024.

[1296] Chaoyun Zhang, He Huang, Chiming Ni, Jian Mu, Si Qin, Shilin He, Lu Wang, Fangkai Yang, Pu Zhao, Chao Du, et al. Ufo2: The desktop agentos. *arXiv preprint arXiv:2504.14603*, 2025.

[1297] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12203–12213, 2020.

[1298] Dan Zhang, G. Feng, Yang Shi, and D. Srinivasan. Physical safety and cyber security analysis of multi-agent systems: A survey of recent advances. *IEEE/CAA Journal of Automatica Sinica*, 2021.

[1299] Danyang Zhang, Lu Chen, Situo Zhang, Hongshen Xu, Zihan Zhao, and Kai Yu. Large language models are semi-parametric reinforcement learning agents. *Neural Information Processing Systems*, 2023.

[1300] Daoan Zhang, Weitong Zhang, Bing He, Jiang Zhang, Chenchen Qin, and Jianhua Yao. Dnagpt: A generalized pre-trained tool for multiple dna sequence analysis tasks. *bioRxiv*, 2024.

[1288] E. Zelikman, Eliana Lorch, Lester Mackey, and A. Kalai. Self-taught optimizer (stop): Recursively self-improving code generation, arXiv preprint arXiv:2310.02304, 2023. URL https://arxiv.org/abs/2310.02304v3.

[1289] 裴增，宁振宇，赵继如，崔伟豪，徐梦伟，郭立伟，陈旭生，和山一舟。面向LLM服务的覆盖原则：长上下文大语言模型服务综述，arXiv预印本arXiv:2405.11299，2024。URLhttps://arxiv.org/abs/2405.11299v2。

[1290] 曾瑞红，方晋元，刘思伟，孟再桥. 关于LLM代理的结构记忆，arXiv预印本arXiv:2412.15266，2024年。URLhttps://arxiv.org/abs/2412.15266v1.

[1291] Yirong Zeng, Xiao Ding, Yuxian Wang, Weiwen Liu, Wu Ning, Yutai Hou, Xu Huang, Bing Qin, and Ting Liu. itool: 强化微调与动态缺陷校准的高级工具使用, arXiv preprint arXiv:2501.09766, 2025. URLhttps://arxiv.org/abs/2501.09766v4.

[1292] Yongcheng Zeng, Xinyu Cui, Xuanfa Jin, Guoqing Liu, Zexu Sun, Dong Li, Ning Yang, Jianye Hao, Haifeng Zhang, and Jun Wang. 通过迭代偏好优化进化 llms 的自我完善能力, arXiv preprint arXiv:2502.05605, 2025. URLhttps://arxiv.org/abs/2502.05605v3.

[1293] An Zhang, Leheng Sheng, Yuxin Chen, Hao Li, Yang Deng, Xiang Wang, and Tat-Seng Chua. 关于推荐中的生成式代理. 年度国际 *ACM SIGIR* 信息检索研究与发展会议, 2023.

[1294] B 张, K 周X 魏X 赵. . . . 评估和改进工具增强的计算密集型数学推理. 2023. URL
h
t
tps://proceedings.neurips.cc/paper_files/paper/2023/hash/4a47dd69242d5af908cdd5d51c971cbf-Abstract-Datasets_and_Benchmarks.html.

[1295] 张超, 张浩鑫, 吴世伟, 吴迪, 许通, 赵祥宇, 高岩, 胡瑶, 和陈恩宏. Notellm-2: 用于推荐的 multimodal 大型表示模型. 知识发现与数据挖掘, 2024.

[1296] 张超云, 黄鹤, 倪晨明, 姜健, 秦思, 何世林, 王陆, 杨方凯, 赵普, 杜超, 等. Ufo2: 桌面代理os. *arXivpreprint arXiv:2504.14603*, 2025.

[1297] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: 少样本图像分类与可微地球移动距离和结构化分类器. 在 *IEEE/CVF*计算机视觉与模式识别会议论文集，第12203–12213页，2020年。

[1298] Dan Zhang, G. Feng, Yang Shi, and D. Srinivasan. 多智能体系统的物理安全与网络安全分析：近期进展综述. *IEEE/CAA*自动化学会期刊, 2021年。

[1299] Danyang Zhang, Lu Chen, Situo Zhang, Hongshen Xu, Zihan Zhao, and Kai Yu. 大型语言模型是半参数强化学习智能体. 神经信息处理系统, 2023年。

[1300] Daoan Zhang, Weitong Zhang, Bing He, Jiang Zhang, Chenchen Qin, and Jianhua Yao. Dnagpt: 一种用于多种DNA序列分析任务的通用预训练工具. *bioRxiv*, 2024年。

[1301] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12401–12430, 2024.

[1302] Duzhen Zhang, Yong Ren, Zhong-Zhi Li, Yahan Yu, Jiahua Dong, Chenxing Li, Zhilong Ji, and Jinfeng Bai. Enhancing multimodal continual instruction tuning with branchlora. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.

[1303] Han Zhang, Langshi Zhou, and Hanfang Yang. Learning to retrieve and reason on knowledge graph through active self-reflection, arXiv preprint arXiv:2502.14932, 2025. URL https://arxiv.org/abs/2502.14932v1.

[1304] Hengyu Zhang. Sinklora: Enhanced efficiency and chat capabilities for long-context large language models, arXiv preprint arXiv:2406.05678, 2024. URL https://arxiv.org/abs/2406.05678v1.

[1305] Jiaxin Zhang, Zhongzhi Li, Mingliang Zhang, Fei Yin, Chenglin Liu, and Yashar Moshfeghi. Geoeval: benchmark for evaluating llms and multi-modal models on geometry problem-solving. 2024.

[1306] Jing Zhang, Xiaokang Zhang, Jifan Yu, Jian Tang, Jie Tang, Cuiping Li, and Hong Chen. Subgraph retrieval enhanced model for multi-hop knowledge base question answering. *Annual Meeting of the Association for Computational Linguistics*, 2022.

[1307] Kai Zhang, Fubang Zhao, Yangyang Kang, and Xiaozhong Liu. Llm-based medical assistant personalization with short- and long-term memory coordination. *North American Chapter of the Association for Computational Linguistics*, 2023.

[1308] Kai Zhang, Yejin Kim, and Xiaozhong Liu. Personalized llm response generation with parameterized memory injection, arXiv preprint arXiv:2404.03565, 2025. URL https://arxiv.org/abs/2404.03565.

[1309] Kechi Zhang, Zhuo Li, Jia Li, Ge Li, and Zhi Jin. Self-edit: Fault-aware code editor for code generation. *Annual Meeting of the Association for Computational Linguistics*, 2023.

[1310] Kechi Zhang, Jia Li, Ge Li, Xianjie Shi, and Zhi Jin. Codeagent: Enhancing code generation with tool-integrated agent systems for real-world repo-level coding challenges. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[1311] Kechi Zhang, Ge Li, Jia Li, Huangzhao Zhang, Jingjing Xu, Hao Zhu, Lecheng Wang, Yihong Dong, Jing Mai, Bin Gu, and Zhi Jin. Computational thinking reasoning in large language models, arXiv preprint arXiv:2506.02658, 2025. URL https://arxiv.org/abs/2506.02658v2.

[1312] Ming-Liang Zhang, Zhong-Zhi Li, Fei Yin, Liang Lin, and Cheng-Lin Liu. Fuse, reason and verify: Geometry problem solving with parsed clauses from diagram. 2024.

[1313] Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Yufei Wang, Irwin King, Xue Liu, and Chen Ma. A survey on test-time scaling in large language models: What, how, where, and how well?, arXiv preprint arXiv:2503.24235, 2025. URL https://arxiv.org/abs/2503.24235v3.

[1301] 张杜珍, 余亚涵, 董嘉华, 李晨星, 苏丹, 崔晨辉, 和于东. Mm- llms: 多模态大语言模型的最新进展. 载于 计算语言学协会 *(ACL) 2024* 年会论文集, 第 12401–12430 页, 2024.

[1302] 张杜珍, 任勇, 李中志, 余亚涵, 董嘉华, 李晨星, 季志龙, 和白金峰. 使用 branchlora 增强多模态持续指令微调. 载于 计算语言学协会第 *63* 届年会论文集（第一卷：长论文）, 2025.

[1303] 张汉, 周朗石, 和杨汉方. 通过主动自我反思学习检索和推理知识图谱, arXiv 预印本 arXiv:2502.14932, 2025. URL https://arxiv.org/abs/2502.14932v1.

[1304] 张恒宇. Sinklora: 增强长上下文大语言模型的效率和聊天能力，arXiv 预印本 arXiv:2406.05678, 2024. URL https://arxiv.org/abs/2406.05678v1.

[1305] 张嘉欣, 李钟智, 张明良, 尹飞, 刘成林, 和 Yashar Moshfeghi. Geoeval: 用于评估大语言模型和多模态模型在几何问题解决上的基准. 2024.

[1306] 张静, 张晓康, 余继帆, 唐健, 唐杰, 李翠萍, 和 陈红. 用于多跳知识库问答的子图检索增强模型. 计算语言学协会年会, 2022.

[1307] 张凯, 赵福邦, 康洋洋, 和 刘晓中. 基于大语言模型的医疗助手个性化，结合短期和长期记忆协调. 计算语言学协会北美分会, 2023.

[1308] 张凯，金冶珍，刘晓中. 基于参数化记忆注入的个性化LLM响应生成，arXiv预印本 arXiv:2404.03565，2025年。URLhttps://arxiv.org/abs/2404.03565。

[1309] 张克奇, 李卓, 李佳, 李格, 和ZhiJin. 自我编辑：用于代码生成的故障感知代码编辑器。计算语言学协会年会, 2023.

[1310] 张克奇, 李佳, 李格, 石先杰, 金智. Codeagent: 基于工具集成代理系统增强代码生成，应对现实世界仓库级别的编码挑战. 计算语言学协会年会, 2024.

[1311] 张克奇, 李格, 李佳, 张黄超, 许晶晶, 朱浩, 王乐成, 董一红, 麦静, 古斌, 金智. 大型语言模型中的计算思维推理, arXiv 预印本 arXiv:2506.02658, 2025. URLhttps://arxiv.org/abs/2506.02658v2.

[1312] 张明良, 李中志, 尹飞, 林亮, 和 刘成林. 融合、推理和验证：基于图解析子句的几何问题求解. 2024.

[1313] 张启元, 吕福元, 孙泽旭, 王雷, 张伟旭, 郭志涵, 王宇飞, King Irwin, 刘雪, 和 马晨. 大型语言模型中的测试时扩展：是什么、如何、在哪里以及效果如何？arXiv 预印本 arXiv:2503.24235, 2025. URL https://arxiv.org/abs/2503.24235v3.

[1314] Ruichen Zhang, Mufan Qiu, Zhen Tan, Mohan Zhang, Vincent Lu, Jie Peng, Kaidi Xu, Leandro Z. Agudelo, Peter Qian, and Tianlong Chen. Symbiotic cooperation for web agents: Harnessing complementary strengths of large and small llms, arXiv preprint arXiv:2502.07942, 2025. URL https://arxiv.org/abs/2502.07942v2.

[1315] Tengchao Zhang, Yonglin Tian, Fei Lin, Jun Huang, Patrik P. Süli, Rui Qin, and Fei-Yue Wang. Coordfield: Coordination field for agentic uav task allocation in low-altitude urban scenarios, arXiv preprint arXiv:2505.00091, 2025. URL https://arxiv.org/abs/2505.00091v3.

[1316] Weizhi Zhang, Xinyang Zhang, Chenwei Zhang, Liangwei Yang, Jingbo Shang, Zhepei Wei, Henry Peng Zou, Zijie Huang, Zhengyang Wang, Yifan Gao, Xiaoman Pan, Lian Xiong, Jingguo Liu, Philip S. Yu, and Xian Li. Personaagent: When large language model agents meet personalization at test time, arXiv preprint arXiv:2506.06254, 2025. URL https://arxiv.org/abs/2506.06254v1.

[1317] Wen Zhang, Long Jin, Yushan Zhu, Jiaoyan Chen, Zhiwei Huang, Junjie Wang, Yin Hua, Lei Liang, and Hua zeng Chen. Trustuqa: A trustful framework for unified structured data question answering. *AAAI Conference on Artificial Intelligence*, 2024.

[1318] Wenlin Zhang, Xiangyang Li, Kuicai Dong, Yichao Wang, Pengyue Jia, Xiaopeng Li, Yingyi Zhang, Derong Xu, Zhaochen Du, Huifeng Guo, Ruiming Tang, and Xiangyu Zhao. Process vs. outcome reward: Which is better for agentic rag reinforcement learning, arXiv preprint arXiv:2505.14069, 2025. URL https://arxiv.org/abs/2505.14069v2.

[1319] Wentao Zhang, Ce Cui, Yilei Zhao, Rui Hu, Yang Liu, Yahui Zhou, and Bo An. Agentorchestra: A hierarchical multi-agent framework for general-purpose task solving, arXiv preprint arXiv:2506.12508, 2025. URL https://arxiv.org/abs/2506.12508v2.

[1320] Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding, arXiv preprint arXiv:2505.18079, 2025. URL https://arxiv.org/abs/2505.18079v2.

[1321] Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D. Manning, and J. Leskovec. Greaselm: Graph reasoning enhanced language models for question answering. *International Conference on Learning Representations*, 2022.

[1322] Yao Zhang, Zijian Ma, Yunpu Ma, Zhen Han, Yu Wu, and Volker Tresp. Webpilot: A versatile and autonomous multi-agent system for web task execution with strategic exploration, arXiv preprint arXiv:2408.15978, 2024. URL https://arxiv.org/abs/2408.15978.

[1323] Yinger Zhang, Hui Cai, Yicheng Chen, Rui Sun, and Jing Zheng. Reverse chain: A generic-rule for llms to master multi-api planning, arXiv preprint arXiv:2310.04474, 2023. URL https://arxiv.org/abs/2310.04474v3.

[1324] Youjia Zhang, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. Ma-bert: Learning representation by incorporating multi-attribute knowledge in transformers. *Findings*, 2021.

[1325] Yu Zhang, Jinlong Ma, Yongshuai Hou, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. Evaluating and steering modality preferences in multimodal large language model, arXiv preprint arXiv:2505.20977v1, 2025. URL https://arxiv.org/abs/2505.20977v1.

[1314] 张瑞辰, 邱沐凡, 谭振, 张漠, 陆文森, 彭杰, 许凯迪, Leandro Z. Agudelo, 钱佩, 陈天隆. 网络代理的共生合作：利用大小大语言模型的互补优势, arXiv 预印本 arXiv:2502.07942, 2025. URLhttps://arxiv.org/abs/2502.07942v2.

[1315] 张腾超, 田永林, 林飞, 黄俊, Patrik P. Süli, 秦瑞, 王飞越. Coordfield: 低空城市场景中自主无人机任务分配的协调场, arXiv 预印本 arXiv:2505.00091, 2025. URLhttps://arxiv.org/abs/2505.00091v3.

[1316] 张伟志, 张新阳, 张晨伟, 杨亮伟, 尚景波, 魏哲培, 张亨利, 黄子杰, 王正阳, 高一帆, 潘晓曼, 熊连, 刘景国, Philip S. Yu, 李先. Personaagent: 当大语言模型代理在测试时遇到个性化, arXiv 预印本 arXiv:2506.06254, 2025. URL https://arxiv.org/abs/2506.06254v1.

[1317] 张文，金龙，朱玉山，陈娇艳，黄志伟，王俊杰，华花，梁雷，陈华增. Trustuqa: 一个用于统一结构化数据问答的可信框架.*AAAI*人工智能会议, 2024.

[1318] 张文林, 李向阳, 董奎才, 王一超, 贾鹏越, 李晓鹏, 张英毅, 许德荣, 杜赵晨, 郭惠峰, 唐瑞明, 和赵祥宇. 过程与结果奖励：哪种更适合代理式RAG强化学习, arXiv预印本arXiv:2505.14069, 2025. URLhttps://arxiv.org/abs/2505.14069v2.

[1319] 张文涛，崔策，赵一蕾，胡瑞，刘杨，周亚辉，和安波。Agentorchestra：一个用于通用任务解决的层次式多智能体框架，arXiv预印本 arXiv:2506.12508，2025。URLhttps://arxiv.org/abs/2506.12508v2.

[1320] 张晓怡, 贾朝阳, 郭宗宇, 李嘉豪, 李斌, 李厚强, 和 陆岩. 深度视频发现：具有工具使用的代理搜索用于长视频理解, arXiv 预印本 arXiv:2505.18079, 2025. URLhttps://arxiv.org/abs/2505.18079v2.

[1321] 张锡坤, Antoine Bosselut, Yasunaga Michihiro, 任洪宇, Liang Percy, Christopher D. Manning, 和 J. Leskovec. Greaselm: 用于问答的图推理增强语言模型. 学习表示国际会议, 2022.

[1322] 张瑶, 马子健, 马云普, 韩振, 吴宇, 和 Volker Tresp. Webpilot: 用于网页任务执行的通用和自主多智能体系统，具有战略探索, arXiv 预印本 arXiv:2408.15978, 2024. URLhttps://arxiv.org/abs/2408.15978.

[1323] 张莺，蔡辉，陈一成，孙睿，郑静. 反向链：一种通用的规则，使大型语言模型掌握多API规划, arXiv预印本arXiv:2310.04474，2023. URLhttps://arxiv.org/abs/2310.04474v3.

张友嘉、王金、余良之、张雪捷。Ma-bert：通过在 Transformer 中整合多属性知识来学习表示。发现，2021。

[1325] 张宇，马金隆，侯永帅，白雪峰，陈可海，向阳，余俊，张敏. 评估和引导多模态大语言模型中的模态偏好，arXiv 预印本 arXiv:2505.20977v1, 2025. URLhttps://arxiv.org/abs/2505.20977v1.

[1326] Yunyi Zhang, Ming Zhong, Siru Ouyang, Yizhu Jiao, Sizhe Zhou, Linyi Ding, and Jiawei Han. Automated mining of structured knowledge from text in the era of large language models. *Knowledge Discovery and Data Mining*, 2024.

[1327] Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö. Arik. Chain of agents: Large language models collaborating on long-context tasks. *Neural Information Processing Systems*, 2024.

[1328] Yuxiang Zhang, Yuqi Yang, Jiangming Shu, Xinyan Wen, and Jitao Sang. Agent models: Internalizing chain-of-action generation into reasoning models, arXiv preprint arXiv:2503.06580, 2025. URL https://arxiv.org/abs/2503.06580v1.

[1329] Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents, arXiv preprint arXiv:2404.13501, 2024. URL https://arxiv.org/abs/2404.13501v1.

[1330] Zeyu Zhang, Quanyu Dai, Luyu Chen, Zeren Jiang, Rui Li, Jieming Zhu, Xu Chen, Yi Xie, Zhenhua Dong, and Ji-Rong Wen. Memsim: A bayesian simulator for evaluating memory of llm-based personal assistants, arXiv preprint arXiv:2409.20163, 2024. URL https://arxiv.org/abs/2409.20163v1.

[1331] Zeyu Zhang, Quanyu Dai, Xu Chen, Rui Li, Zhongyang Li, and Zhenhua Dong. Memengine: A unified and modular library for developing advanced memory of llm-based agents. *The Web Conference*, 2025.

[1332] Zheng Zhang, Liang Ding, Dazhao Cheng, Xuebo Liu, Min Zhang, and Dacheng Tao. Bliss: Robust sequence-to-sequence learning via self-supervised input representation, arXiv preprint arXiv:2204.07837, 2022. URL https://arxiv.org/abs/2204.07837v2.

[1333] Zhenyu (Allen) Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark W. Barrett, Zhangyang Wang, and Beidi Chen. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Neural Information Processing Systems*, 2023.

[1334] Zhihan Zhang, Zhenwen Liang, Wenhao Yu, Dian Yu, Mengzhao Jia, Dong Yu, and Meng Jiang. Learn beyond the answer: Training language models with reflection for mathematical reasoning. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[1335] Zhuosheng Zhang and Aston Zhang. You only look at screens: Multimodal chain-of-action agents. *Annual Meeting of the Association for Computational Linguistics*, 2023.

[1336] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Y. Liu, and Gao Huang. Expel: Llm agents are experiential learners. *AAAI Conference on Artificial Intelligence*, 2023.

[1337] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners, arXiv preprint arXiv:2308.10144, 2024. URL https://arxiv.org/abs/2308.10144.

[1338] Jianan Zhao, Meng Qu, Chaozhuo Li, Hao Yan, Qian Liu, Rui Li, Xing Xie, and Jian Tang. Learning on large-scale text-attributed graphs via variational inference. *International Conference on Learning Representations*, 2022.

[1326] 张云毅, 中明, 欧阳思儒, 焦一珠, 周思哲, 丁林怡, 和韩家伟. 大语言模型时代从文本中自动挖掘结构化知识. 知识发现与数据挖掘, 2024.

[1327] 张宇森, 孙若曦, 陈岩飞, Tomas Pfister, 张瑞, 和 Ö. Arik Sercan. 智能体链: 大语言模型在长上下文任务中的协作. 神经信息处理系统, 2024.

[1328] 张宇翔, 杨宇奇, 舒江明, 温新言, 和桑继涛. 智能体模型: 将行动链生成内化到推理模型中, arXiv 预印本 arXiv:2503.06580, 2025. URLhttps://arxiv.org/abs/2503.06580v1.

[1329] 张泽宇, 博晓和, 马晨, 李瑞, 陈旭, 戴全宇, 朱继明, 董振华, 和文继荣. 基于大语言模型智能体的记忆机制调查, arXiv 预印本 arXiv:2404.13501, 2024. URLhttps://arxiv.org/abs/2404.13501v1.

[1330] 张泽宇, 戴全宇, 陈路宇, 姜泽仁, 李瑞, 朱继明, 陈旭, 谢奕, 董振华, 文继荣. Memsim: 一个用于评估基于LLM的个人助理记忆的贝叶斯模拟器, arXiv预印本 arXiv:2409.20163, 2024. URLhttps://arxiv.org/abs/2409.20163v1.

[1331] 张泽宇, 戴全宇, 陈旭, 李瑞, 李中阳, 和董振华. Memengine: 一个用于开发基于LLM的智能体高级记忆的统一和模块化库. *The Web Conference*, 2025.

[1332] 张铮, 丁亮, 成大兆, 刘雪波, 张敏, 和陶大程. Bliss: 通过自监督输入表示实现鲁棒的序列到序列学习, arXiv预印本 arXiv:2204.07837, 2022. URLhttps://arxiv.org/abs/2204.07837v2.

[1333] 张振宇（张 Allen），盛英，周天一，陈天隆，郑连明，蔡瑞思，宋赵，田元东，Christopher Ré，Clark W. Barrett，王张阳，陈北地. H2o：用于高效生成式推理的大型语言模型的重量级预言机。神经信息处理系统, 2023.

[1334] 张志汉，梁振文，余文豪，余狄，贾梦昭，余东，蒋梦。超越答案学习：使用反思训练用于数学推理的语言模型。自然语言处理经验方法会议,2024。

[1335] 张卓升和张 Aston。你只看屏幕：多模态行动链智能体。计算语言学协会年度会议, 2023。

[1336] 赵安迪，黄丹尼尔，徐文沁，林马修，刘一，和高黄。Expel：Llm智能体是体验式学习者。人工智能协会会议, 2023。

[1337] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners, arXiv preprint arXiv:2308.10144, 2024. URL https://arxiv.org/abs/2308.10144.

[1338] 赵建安，曲萌，李超卓，闫浩，刘倩，李瑞，谢星，唐健。基于变分推理的大规模文本属性图学习。国际学习表征会议, 2022.

[1339] Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey, arXiv preprint arXiv:2402.19473, 2024. URL https://arxiv.org/abs/2402.19473v6.

[1340] Pengyu Zhao, Zijian Jin, and Ning Cheng. An in-depth survey of large language model-based artificial intelligence agents, arXiv preprint arXiv:2309.14365, 2023. URL https://arxiv.org/abs/2309.14365v1.

[1341] Pu Zhao, Xuan Shen, Zhenglun Kong, Yixin Shen, Sung-En Chang, Timothy Rupprecht, Lei Lu, Enfu Nan, Changdi Yang, Yumei He, Xingchen Xu, Yu Huang, Wei Wang, Yue Chen, Yongchun He, and Yanzhi Wang. 7b fully open source moxin-llm/vlm – from pretraining to grpo-based reinforcement learning enhancement. arXiv preprint, 2024.

[1342] Qi Zhao, Hongyu Yang, Qi Song, Xinwei Yao, and Xiangyang Li. Knowpath: Knowledge-enhanced reasoning via llm-generated inference paths over knowledge graphs, arXiv preprint arXiv:2502.12029, 2025. URL https://arxiv.org/abs/2502.12029v3.

[1343] Qifang Zhao, Weidong Ren, Tianyu Li, Xiaoxiao Xu, and Hong Liu. Graphgpt: Generative pre-trained graph eulerian transformer, arXiv preprint arXiv:2401.00529v3, 2023. URL https://arxiv.org/abs/2401.00529v3.

[1344] Ruilin Zhao, Feng Zhao, Long Wang, Xianzhi Wang, and Guandong Xu. Kg-cot: Chain-of-thought prompting of large language models over knowledge graphs for knowledge-aware question answering. *International Joint Conference on Artificial Intelligence*, 2024.

[1345] Shangziqi Zhao, Jiahao Yuan, Guisong Yang, and Usman Naseem. Can pruning improve reasoning? revisiting long-cot compression with capability in mind for better reasoning, arXiv preprint arXiv:2505.14582, 2025. URL https://arxiv.org/abs/2505.14582v1.

[1346] Shitian Zhao, Zhuowan Li, Yadong Lu, Alan L. Yuille, and Yan Wang. Causal-cog: A causal-effect look at context generation for boosting multi-modal language models. *Computer Vision and Pattern Recognition*, 2023.

[1347] Tony Zhao, Eric Wallace, Shi Feng, D. Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. *International Conference on Machine Learning*, 2021.

[1348] Weixiang Zhao, Xingyu Sui, Jiahe Guo, Yulin Hu, Yang Deng, Yanyan Zhao, Bing Qin, Wanxiang Che, Tat-Seng Chua, and Ting Liu. Trade-offs in large reasoning models: An empirical analysis of deliberative and adaptive reasoning over foundational capabilities, arXiv preprint arXiv:2503.17979, 2025. URL https://arxiv.org/abs/2503.17979v1.

[1349] Yibo Zhao, Jiapeng Zhu, Ye Guo, Kangkang He, and Xiang Li. E^{2}graphrag: Streamlining graph-based rag for high efficiency and effectiveness. arXiv preprint, 2025.

[1350] Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v(ision) is a generalist web agent, if grounded. *International Conference on Machine Learning*, 2024.

[1351] Changmeng Zheng, Dayong Liang, Wengyu Zhang, Xiao Wei, Tat seng Chua, and Qing Li. A picture is worth a graph: A blueprint debate paradigm for multimodal reasoning. *ACM Multimedia*, 2024.

[1339] 赵鹏浩, 张海林, 余庆汉, 王正仁, 耿云腾, 傅方成, 杨凌, 张文涛, 和崔斌. 基于检索增强的生成: 人工智能生成内容的调查, arXiv 预印本 arXiv:2402.19473, 2024. URLhttps://arxiv.org/abs/2402.19473v6.

[1340] 赵鹏宇, 金子健, 和程宁. 基于大型语言模型的人工智能智能体的深入调查, arXiv 预印本 arXiv:2309.14365, 2023. URLhttps://arxiv.org/abs/2309.14365v1.

[1341] 赵普, 沈璇, 孔正伦, 沈奕欣, 张崇恩, 伦纳德·R·鲁普雷希特, 陆雷, 难恩福, 杨长地, 何雨梅, 许行尘, 黄宇, 王伟, 陈岳, 何永春, 和王彦之. 7b 完全开源的摩辛-llm/vlm——从预训练到基于 grpo 的强化学习增强. arXiv 预印本, 2024.

[1342] 赵琦, 杨红宇, 宋奇, 姚新伟, 和 李向阳. Knowpath: 基于知识图谱的LLM生成推理路径的知识增强推理, arXiv预印本arXiv:2502.12029, 2025. URLhttps://arxiv.org/abs/2502.12029v3.

[1343] 赵启方, 任伟东, 李天宇, 许晓晓, 刘红. GraphGPT: 生成式预训练图欧拉变换器, arXiv 预印本 arXiv:2401.00529v3, 2023. URL https://arxiv.org/abs/2401.00529v3.

赵瑞林, 赵峰, 王龙, 王先志, 许冠东. Kg-cot: 基于知识图谱的大语言模型的思维链提示用于知识感知问答. 国际人工智能联合会议, 2024.

[1345] 赵上奇, 袁嘉浩, 杨贵松, 和 Usman Naseem. 能否通过剪枝来提升推理能力？重新审视长程压缩， 考虑能力以实现更好的推理, arXiv 预印本 arXiv:2505.14582, 2025. URLhttps://arxiv.org/abs/2505.14582v1.

[1346] 赵天奇, 李 Zhuowan, 陆 Yadong, Alan L. Yuille, 和 王岩. Causal-cog: 从因果关系视角审视上下文生成以提升多模态语言模型. 计算机视觉与模式识别, 2023.

[1347] 赵 Tony, Eric Wallace, 石 丰, D. Klein, 和 Sameer Singh. 使用前校准：提升语言模型的少样本性能. 机器学习国际会议, 2021.

[1348] 赵伟祥, 隋兴宇, 郭佳禾, 胡雨林, 邓阳, 赵艳艳, 秦冰, 车万祥, Chua Tat-Seng, 和 刘婷. 大型推理模型的权衡：对基础能力上审议和自适应推理的经验分析, arXiv 预印本 arXiv:2503.17979, 2025. URLhttps://arxiv.org/abs/2503.17979v1.

[1349] 赵一博, 朱嘉鹏, 郭叶, 何康康, 和李翔. E2graphrag: 高效且有效的基于图的rag流式处理. arXiv预印本, 2025.

[1350] 郑博文, 郭伯宇, 金基雄, 孙欢, 和苏宇. Gpt-4v(ision)是一个通才式网络代理, 如果有基础. 机器学习国际会议, 2024.

[1351] 郑长萌, 梁大勇, 张文宇, 魏晓, 蔡天锡, 和李清. 一图胜千言: 多模态推理的蓝图辩论范式. *ACM Multimedia*, 2024.

[1352] Chuanyang Zheng, Yihang Gao, Han Shi, Minbin Huang, Jingyao Li, Jing Xiong, Xiaozhe Ren, Michael Ng, Xin Jiang, Zhenguo Li, and Yu Li. Dape: Data-adaptive positional encoding for length extrapolation. *Neural Information Processing Systems*, 2024.

[1353] Chunmo Zheng, Saika Wong, Xing Su, Yinqiu Tang, Ahsan Nawaz, and Mohamad Kassem. Automating construction contract review using knowledge graph-enhanced large language models. *Automation in Construction*, 2023.

[1354] Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. Towards lifelong learning of large language models: A survey. *ACM Computing Surveys*, 2024.

[1355] Junhao Zheng, Xidi Cai, Qiuke Li, Duzhen Zhang, Zhongzhi Li, Yingying Zhang, Le Song, and Qianli Ma. Lifelongagentbench: Evaluating llm agents as lifelong learners, arXiv preprint arXiv:2505.11942, 2025. URL https://arxiv.org/abs/2505.11942v3.

[1356] Longtao Zheng, R. Wang, and Bo An. Synapse: Trajectory-as-exemplar prompting with memory for computer control. *International Conference on Learning Representations*, 2023.

[1357] Longtao Zheng, Rundong Wang, Xinrun Wang, and Bo An. Synapse: Trajectory-as-exemplar prompting with memory for computer control, arXiv preprint arXiv:2306.07863, 2024. URL https://arxiv.org/abs/2306.07863.

[1358] Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutao Jiang, Bin Ren, Jialei Chen, Jiawen Wang, Chengxin Li, Linfeng Zhang, D. Paudel, Xuanjing Huang, Yu-Gang Jiang, N. Sebe, Dacheng Tao, L. V. Gool, and Xuming Hu. Mllms are deeply affected by modality bias, arXiv preprint arXiv:2505.18657v1, 2025. URL https://arxiv.org/abs/2505.18657v1.

[1359] Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments, arXiv preprint arXiv:2504.03160, 2025. URL https://arxiv.org/abs/2504.03160v4.

[1360] Li Zhong, Zilong Wang, and Jingbo Shang. Debug like a human: A large language model debugger via verifying runtime execution step by step. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[1361] Rui Zhong, Yang Cao, Jun Yu, and M. Munetomo. Large language model assisted adversarial robustness neural architecture search. *2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, 2024.

[1362] Wanjun Zhong, Lianghong Guo, Qi-Fei Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. *AAAI Conference on Artificial Intelligence*, 2023.

[1363] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory, arXiv preprint arXiv:2305.10250, 2023. URL https://arxiv.org/abs/2305.10250.

[1364] Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models. *International Conference on Machine Learning*, 2023.

[1352] Chuanyang Zheng, Yihang Gao, Han Shi, Minbin Huang, Jingyao Li, Jing Xiong, Xiaozhe Ren, Michael Ng, Xin Jiang, Zhenguo Li, and Yu Li. Dape: 数据自适应位置编码用于长度外推. *Neural InformationProcessingSystems*, 2024.

[1353] Chunmo Zheng, Saika Wong, Xing Su, Yinqiu Tang, Ahsan Nawaz, andMohamadKassem. 使用知识图谱增强的大型语言模型自动化构建合同审查. *Automationin Construction*, 2023.

[1354] Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. 大型语言模型终身学习：一项调查. *ACM ComputingSurveys*, 2024.

[1355] Junhao Zheng, Xidi Cai, Qiuke Li, Duzhen Zhang, Zhongzhi Li, Yingying Zhang, Le Song, and Qianli Ma. Lifelongagentbench: 评估 llmagents 作为终身学习者, arXiv 预印本 arXiv:2505.11942, 2025. URLhttps://arxiv.org/abs/2505.11942v3.

[1356] Longtao Zheng, R. Wang, and Bo An. Synapse: Trajectory-as-exemplar prompting with memory for computer control. *International Conference on Learning Representations*, 2023.

[1357] 郑龙涛, 王润东, 王新润, 和安波. Synapse: 基于轨迹作为范例的提示与记忆的计算机控制, arXiv 预印本 arXiv:2306.07863, 2024. URLhttps://arxiv.org/abs/2306.07863.

[1358] 郑旭, 廖晨飞, 傅宇倩, 雷凯宇, 吕元慧, 江路涛, 任斌, 陈佳蕾, 王嘉雯, 李成新, 张林峰, Paudel D., 黄宣敬, 蒋宇光, Sebe N., 陶大程, Gool L. V., 和胡旭明. Mllms 深受模态偏差影响, arXiv 预印本 arXiv:2505.18657v1, 2025. URLhttps://arxiv.org/abs/2505.18657v1.

[1359] 朱宇翔, 傅大元, 胡祥坤, 蔡晓杰, 叶雨珊, 陆鹏瑞, 和 刘鹏飞. Deepresearcher: 通过强化学习在真实环境中扩展深度研究, arXiv 预印本 arXiv:2504.03160, 2025. URLhttps://arxiv.org/abs/2504.03160v4.

[1360] 李中, 王子龙, 和 尚景波. 像人类一样调试: 通过逐步验证运行时执行的大型语言模型调试器. 计算语言学协会年度会议, 2024.

[1361] 钟睿, 曹阳, 余俊, 和 M. Munetomo. 大型语言模型辅助对抗鲁棒性神经架构搜索. *2024* 年第 *6* 届复杂数据驱动优化系统国际会议 *(DOCS)*, 2024.

[1362] 钟万军, 郭良红, 高启飞, 叶鹤, 和 王艳林. Memorybank: 通过长期记忆增强大型语言模型. 人工智能协会会议, 2023.

[1363] 钟万军, 郭良红, 高琪琪, 叶鹤, 王艳林. Memorybank: 为大型语言模型增强长期记忆, arXiv 预印本 arXiv:2305.10250, 2023. URLhttps://arxiv.org/abs/2305.10250.

[1364] 周安, 严凯, Michal Shlapentokh-Rothman, 王浩瀚, 和 王宇雄. 语言代理树搜索统一了语言模型中的推理、行动和规划. 机器学习国际会议, 2023.

[1365] Bin Zhou, Xingwang Shen, Yuqian Lu, Xinyu Li, B. Hua, Tianyuan Liu, and Jinsong Bao. Semantic-aware event link reasoning over industrial knowledge graph embedding time series data. *International Journal of Production Research*, 2022.

[1366] Denny Zhou, Nathanael Scharli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, D. Schuurmans, O. Bousquet, Quoc Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. *International Conference on Learning Representations*, 2022.

[1367] Huichi Zhou, Kin-Hei Lee, Zhonghao Zhan, Yue Chen, Zhenhao Li, Zhaoyang Wang, Hamed Haddadi, and Emine Yilmaz. Trustrag: Enhancing robustness and trustworthiness in rag, arXiv preprint arXiv:2501.00879, 2025. URL https://arxiv.org/abs/2501.00879.

[1368] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. *International Conference on Learning Representations*, 2023.

[1369] Wangchunshu Zhou, Y. Jiang, Long Li, Jialong Wu, Tiannan Wang, Shi Qiu, Jintian Zhang, Jing Chen, Ruipu Wu, Shuai Wang, Shiding Zhu, Jiyu Chen, Wentao Zhang, Xiangru Tang, Ningyu Zhang, Huajun Chen, Peng Cui, and Mrinmaya Sachan. Agents: An open-source framework for autonomous language agents, arXiv preprint arXiv:2309.07870, 2023. URL https://arxiv.org/abs/2309.07870v3.

[1370] Yingli Zhou, Yaodong Su, Youran Sun, Shu Wang, Taotao Wang, Runyuan He, Yongwei Zhang, Sicong Liang, Xilin Liu, Yuchi Ma, and Yixiang Fang. In-depth analysis of graph-based rag in a unified framework, arXiv preprint arXiv:2503.04338, 2025. URL https://arxiv.org/abs/2503.04338v1.

[1371] Yuhang Zhou and Wei Ai. Teaching-assistant-in-the-loop: Improving knowledge distillation from imperfect teacher models in low-budget scenarios. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[1372] Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S. Yu. Trustworthiness in retrieval-augmented generation systems: A survey, arXiv preprint arXiv:2409.10102, 2024. URL https://arxiv.org/abs/2409.10102v1.

[1373] Zhehua Zhou, Jiayang Song, Kunpeng Yao, Zhan Shu, and Lei Ma. Isr-llm: Iterative self-refined large language model for long-horizon sequential task planning. *IEEE International Conference on Robotics and Automation*, 2023.

[1374] Zihan Zhou, Chong Li, Xinyi Chen, Shuo Wang, Yu Chao, Zhili Li, Haoyu Wang, Rongqiao An, Qi Shi, Zhixing Tan, Xu Han, Xiaodong Shi, Zhiyuan Liu, and Maosong Sun. Llm×mapreduce: Simplified long-sequence processing using large language models, arXiv preprint arXiv:2410.09342, 2024. URL https://arxiv.org/abs/2410.09342v1.

[1375] Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, B. Low, and P. Liang. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents, arXiv preprint arXiv:2506.15841, 2025. URL https://arxiv.org/abs/2506.15841v1.

[1376] Andrew Zhu, Liam Dugan, and Christopher Callison-Burch. Redel: A toolkit for llm-powered recursive multi-agent systems. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[1365] 周斌, 沈兴旺, 陆宇倩, 李新宇, Hua B., 刘天元, 和包金松. 基于工业知识图谱嵌入时间序列数据的语义感知事件链接推理. *InternationalJournal ofProductionResearch*, 2022.

[1366] 周丹尼, Scharli Nathanael, 侯乐, 魏建森, Scales Nathan, 王学志, Schuurmans D., Bousquet O., Le Quoc, 和 Chi Ed H. 从少到多的提示使大型语言模型能够进行复杂推理. *International ConferenceonLearningRepresentations*, 2022.

[1367] 周慧, Lee Kin-Hei, 翟中浩, 陈越, 李振豪, 王朝阳, Haddadi Hamed, 和 Yilmaz Emine. Trustrag: 增强rag的鲁棒性和可信度, arXiv预印本 arXiv:2501.00879, 2025. URLhttps://arxiv.org/abs/2501.00879.

[1368] 周舒岩, 徐弗兰克, 朱浩, 周旭辉, 罗罗伯特, 斯里德哈克阿比什克, 成先毅, 比斯卡约纳坦, 弗里德丹尼尔, 阿隆Uri, 和纽比格格雷厄姆. Webarena: 一个用于构建自主代理的现实网络环境. 国际学习表征会议, 2023.

[1369] 周王春树, 蒋颖, 李龙, 吴嘉龙, 王天南, 邱石, 张金田, 陈静, 吴瑞普, 王帅, 朱世定, 陈继宇, 张文涛, 唐祥如, 张宁宇, 陈华军, 崔鹏, 和萨奇南米林玛亚. Agents: 一个用于自主语言代理的开源框架, arXiv预印本arXiv:2309.07870, 2023. URLhttps://arxiv.org/abs/2309.07870v3.

[1370] 周英丽, 苏瑶东, 孙宇然, 王舒, 王涛涛, 何润元, 张永伟, 梁思聪, 刘希林, 马雨琪, 方奕翔. 基于统一框架的图结构rag深入分析, arXiv预印本 arXiv:2503.04338, 2025. URLhttps://arxiv.org/abs/2503.04338v1.

[1371] 周宇航和阿伟. 环教学助理: 在预算有限场景下从不完美教师模型中改进知识蒸馏. 计算语言学协会年会, 2024.

[1372] 周宇嘉, 刘岩, 李晓希, 金佳杰, 钱宏进, 刘铮, 李超卓, 董志成, 何宗谊, 余必胜. 检索增强生成系统中的可信度:一项调查, arXiv预印本 arXiv:2409.10102, 2024. URLhttps://arxiv.org/abs/2409.10102v1.

[1373] 周治华, 宋嘉阳, 姚坤鹏, 舒战, 和马雷. Isr-llm: 用于长时序序列任务规划的迭代自精炼大型语言模型. *IEEEInternationalConferenceonRobotics and Automation*, 2023.

[1374] 周治华, 李崇, 陈新怡, 王硕, 曹宇, 李志利, 王浩宇, 安荣桥, 石奇, 谭志兴, 韩旭, 石晓东, 刘志远, 和孙茂松. Llm×mapreduce: 使用大型语言模型简化长序列处理, arXiv preprint arXiv:2410.09342, 2024. URLhttps://arxiv.org/abs/2410.09342v1.

[1375] 周治建, 邱澳, 吴兆玄, 金成焕, Prakash Alok, Rus Daniela, 赵金华, Low B., 和 Liang P. Mem1: 学习协同记忆和推理以实现高效长时序智能体, arXiv preprint arXiv:2506.15841, 2025. URLhttps://arxiv.org/abs/2506.15841v1.

[1376] Andrew Zhu, Liam Dugan, and Christopher Callison-Burch. Redel: A toolkitfor llm-powered recursive multi-agent systems. 自然语言处理经验方法会议, 2024.

[1377] Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. Pose: Efficient context window extension of llms via positional skip-wise training. *International Conference on Learning Representations*, 2023.

[1378] Hongyin Zhu. Metaaid 2.5: A secure framework for developing metaverse applications via large language models. arXiv preprint, 2023.

[1379] Jason Zhu, Yanling Cui, Yuming Liu, Hao Sun, Xue Li, Markus Pelger, Liangjie Zhang, Tianqi Yan, Ruofei Zhang, and Huasha Zhao. Textgnn: Improving text encoder via graph neural network in sponsored search. *The Web Conference*, 2021.

[1380] Jiachen Zhu, Menghui Zhu, Renting Rui, Rong Shan, Congmin Zheng, Bo Chen, Yunjia Xi, Jianghao Lin, Weiwen Liu, Ruiming Tang, Yong Yu, and Weinan Zhang. Evolutionary perspectives on the evaluation of llm-based ai agents: A comprehensive survey, arXiv preprint arXiv:2506.11102, 2025. URL https://arxiv.org/abs/2506.11102v1.

[1381] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, Dengnian Chen, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Cong He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Ying Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Lijun Wu, Kai Zhang, Hui Deng, Jiaye Ge, Kaiming Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, arXiv preprint arXiv:2504.10479v3, 2025. URL https://arxiv.org/abs/2504.10479v3.

[1382] Mingwei Zhu, Leigang Sha, Yu Shu, Kangjia Zhao, Tiancheng Zhao, and Jianwei Yin. Benchmarking sequential visual input reasoning and prediction in multimodal large language models, arXiv preprint arXiv:2310.13473v1, 2023. URL https://arxiv.org/abs/2310.13473v1.

[1383] Rongzhi Zhu, Xiangyu Liu, Zequn Sun, Yiwei Wang, and Wei Hu. Mitigating lost-in-retrieval problems in retrieval augmented multi-hop question answering. 2025.

[1384] Rongzhi Zhu, Yi Liu, Zequn Sun, Yiwei Wang, and Wei Hu. When can large reasoning models save thinking? mechanistic analysis of behavioral divergence in reasoning. 2025.

[1385] Runchuan Zhu, Zinco Jiang, Jiang Wu, Zhipeng Ma, Jiahe Song, Fengshuo Bai, Dahua Lin, Lijun Wu, and Conghui He. Grait: Gradient-driven refusal-aware instruction tuning for effective hallucination mitigation. 2025.

[1386] Runchuan Zhu, Zhipeng Ma, Jiang Wu, Junyuan Gao, Jiaqi Wang, Dahua Lin, and Conghui He. Utilize the flow before stepping into the same river twice: Certainty represented knowledge flow for refusal-aware instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26157–26165, 2025.

[1387] Tongyao Zhu, Qian Liu, L. Pang, Zhengbao Jiang, Min-Yen Kan, and Min Lin. Beyond memorization: The challenge of random memory access in language models. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[1388] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and Jifeng Dai. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based

[1377] Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, 和 Sujian Li. Pose: 通过位置跳跃式训练高效扩展 llms 的上下文窗口. 国际学习表征会议, 2023.

[1378] Hongyin Zhu. Metaaid 2.5: 一种通过大型语言模型开发元宇宙应用的安全框架. arXiv 预印本, 2023.

[1379] Jason Zhu, Yanling Cui, Yuming Liu, Hao Sun, Xue Li, Markus Pelger, Liangjie Zhang, Tianqi Yan, Ruofei Zhang, 和 Huasha Zhao. Textgnn: 在赞助搜索中通过图神经网络改进文本编码器. 网络会议, 2021.

[1380] Jiachen Zhu, Menghui Zhu, Renting Rui, Rong Shan, Congmin Zheng, Bo Chen, Yunjia Xi, Jianghao Lin, Weiwen Liu, Ruiming Tang, Yong Yu, 和 Weinan Zhang. 基于llm的ai代理评估的进化视角：一项综合调查, arXiv 预印本 arXiv:2506.11102, 2025. URL https://arxiv.org/abs/2506.11102v1.

[1381] 金国珠, 王伟云, 陈哲, 刘兆阳, 叶胜龙, 郭立新, 段宇辰, 田浩, 苏伟杰, 邵杰, 高张伟, 崔尔飞, 曹越, 刘阳州, 王浩民, 许伟业, 李浩, 王嘉豪, 吕汉, 陈登年, 李松泽, 何一南, 姜檀, 罗嘉鹏, 王奕, 何丛, 石博天, 张兴成, 邵文琪, 何军军, 邢雄, 邱文文, 孙鹏, 焦鹏龙, 吴立军, 张凯, 邓会, 葛嘉叶, 陈凯明, 王黎明, 陶敏, 陆乐伟, 朱锡舟, 陆通, 林大华, 邱宇, 戴继峰, 王文海. Internvl3: 探索开源多模态模型的先进训练和测试时配方, arXiv preprint arXiv:2504.10479v3, 2025. URLhttps://arxiv.org/abs/2504.10479v3.

[1382] 朱明伟, 沙雷刚, 舒宇, 赵康嘉, 赵天成, 和 尹建伟. 多模态大语言模型中序列视觉输入推理和预测的基准测试, arXiv 预印本 arXiv:2310.13473v1, 2023. URLhttps://arxiv.org/abs/2310.13473v1.

[1383] 朱荣智，刘向宇，孙泽群，王依伟，和胡伟。缓解检索增强多跳问答中的检索丢失问题。2025。

[1384] 朱荣智，刘毅，孙泽坤，王怡伟，和胡伟。大型推理模型何时能节省思考？推理行为差异的机制分析。2025。

[1385] 朱润川, 姜子昂, 吴江, 马志鹏, 宋嘉禾, 白峰硕, 林大华, 吴立军, 何聪辉. Grait: 基于梯度驱动的拒绝感知指令微调以有效缓解幻觉. 2025.

[1386] 朱润川, 马志鹏, 吴江, 高俊元, 王嘉琪, 林大华, 和Conghui He. 利用流程避免重复入河：用于拒绝感知指令微调的确定性知识流表示。在 *AAAI*人工智能会议论文集，第39卷，第26157–26165页，2025年。

[1387] 朱通耀, 刘倩, Pang L., 姜正宝, Kan Min-Yen, 和 Lin Min. 超越记忆：语言模型中随机内存访问的挑战。 计算语言学协会年度会议, 2024年。

[1388] 朱锡舟, 陈云涛, 田浩, 陶晨欣, 苏伟杰, 杨晨宇, 黄高, 李斌, 陆磊伟, 王晓刚, 邱宇, 张赵翔, 和 戴继峰. 我的craftworld里有个鬼：通过基于文本的大型语言模型为开放世界环境设计通用智能体

knowledge and memory, arXiv preprint arXiv:2305.17144, 2023. URL https://arxiv.org/abs/2305.17144.

[1389] Yue Zhu, Hao Yu, Chen Wang, Zhuoran Liu, and Eun Kyung Lee. Towards efficient key-value cache management for prefix prefilling in llm inference, arXiv preprint arXiv:2505.21919, 2025. URL https://arxiv.org/abs/2505.21919v1.

[1390] Zhengqiu Zhu, Yong Zhao, Bin Chen, S. Qiu, Kai Xu, Quanjun Yin, Jin-Yu Huang, Zhong Liu, and Fei Wang. Conversational crowdsensing: A parallel intelligence powered novel sensing approach, arXiv preprint arXiv:2402.06654, 2024. URL https://arxiv.org/abs/2402.06654v1.

[1391] Zulun Zhu, Tiancheng Huang, Kai Wang, Junda Ye, Xinghe Chen, and Siqiang Luo. Graph-based approaches and functionalities in retrieval-augmented generation: A comprehensive survey, arXiv preprint arXiv:2504.10499, 2025. URL https://arxiv.org/abs/2504.10499v1.

[1392] Alex Zhuang, Ge Zhang, Tianyu Zheng, Xinrun Du, Junjie Wang, Weiming Ren, Stephen W. Huang, Jie Fu, Xiang Yue, and Wenhu Chen. Structlm: Towards building generalist models for structured knowledge grounding, arXiv preprint arXiv:2402.16671, 2024. URL https://arxiv.org/abs/2402.16671v7.

[1393] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. Toolqa: A dataset for llm question answering with external tools. *Neural Information Processing Systems*, 2023.

[1394] Zhixiong Zhuang, Maria-Irina Nicolae, Hui-Po Wang, and Mario Fritz. Proxyprompt: Securing system prompts against prompt extraction attacks, arXiv preprint arXiv:2505.11459, 2025. URL https://arxiv.org/abs/2505.11459v1.

[1395] Ziyuan Zhuang, Zhiyang Zhang, Sitao Cheng, Fangkai Yang, Jia Liu, Shujian Huang, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. Efficientrag: Efficient retriever for multi-hop question answering. In *In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3392–3411, 2024.

[1396] Chang Zong, Yuchen Yan, Weiming Lu, Eliot Huang, Jian Shao, and Y. Zhuang. Triad: A framework leveraging a multi-role llm-based agent to solve knowledge base question answering. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[1397] Yongshuo Zong, Ondrej Bohdal, and Timothy M. Hospedales. Vl-icl bench: The devil in the details of benchmarking multimodal in-context learning. arXiv preprint, 2024.

[1398] Yongshuo Zong, Ondrej Bohdal, and Timothy M. Hospedales. Vl-icl bench: The devil in the details of multimodal in-context learning. *International Conference on Learning Representations*, 2024.

[1399] Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Yankai Chen, Chunyu Miao, Hoang Nguyen, Yue Zhou, Weizhi Zhang, Liancheng Fang, Langzhou He, Yangning Li, Yuwei Cao, Dongyuan Li, Renhe Jiang, and Philip S. Yu. A survey on large language model based human-agent systems. arXiv preprint, 2025.

[1400] Tao Zou, Le Yu, Yifei Huang, Leilei Sun, and Bo Du. Pretraining language models with text-attributed heterogeneous graphs. *Conference on Empirical Methods in Natural Language Processing*, 2023.

[1401] Adam Zweiger, Jyothish Pari, Han Guo, Ekin Akyürek, Yoon Kim, and Pulkit Agrawal. Self-adapting language models, arXiv preprint arXiv:2506.10943, 2025. URL https://arxiv.org/abs/2506.10943v1.

知识与管理，arXiv preprint arXiv:2305.17144，2023。URLhttps://arxiv.org/abs/2305.17144。

[1389] Yue Zhu, Hao Yu, Chen Wang, Zhuoran Liu, 和 Eun Kyung Lee。面向高效键值缓存管理以支持前缀预填充的 llm 推理，arXiv preprint arXiv:2505.21919，2025。URLhttps://arxiv.org/abs/2505.21919v1。

[1390] Zhengqiu Zhu, Yong Zhao, Bin Chen, S. Qiu, Kai Xu, Quanjun Yin, Jin-Yu Huang, Zhong Liu, 和 Fei Wang。对话式众包感知：一种由并行智能驱动的创新感知方法，arXiv preprint arXiv:2402.06654，2024。URLhttps://arxiv.org/abs/2402.06654v1。

[1391] Zulun Zhu, Tiancheng Huang, Kai Wang, Junda Ye, Xinghe Chen, and Siqiang Luo. 图像检索增强生成中的基于图的方法和功能：一项综合调查，arXiv 预印本 arXiv:2504.10499，2025。URLhttps://arxiv.org/abs/2504.10499v1.

[1392] Alex Zhuang, Ge Zhang, Tianyu Zheng, Xinrun Du, Junjie Wang, Weiming Ren, Stephen W. Huang, Jie Fu, Xiang Yue, and Wenhu Chen. Structlm：面向结构化知识 grounding 的通用模型构建，arXiv 预印本 arXiv:2402.16671，2024。URL https://arxiv.org/abs/2402.16671v7.

[1393] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. Toolqa：用于 LLM 外部工具问答的数据集。 神经信息处理系统, 2023。

[1394] 庄志雄, 玛丽亚-伊琳娜·尼古拉耶, 王辉坡, 和 马里奥·弗里茨. Proxyprompt: 防止提示提取攻击的系统提示安全, arXiv 预印本 arXiv:2505.11459, 2025. URLhttps://arxiv.org/abs/2505.11459v1.

[1395] 庄子源, 张志阳, 程思涛, 杨方凯, 刘嘉, 黄树健, 林清伟, 桑瓦南·拉贾莫汉, 张冬梅, 和 张琪. Efficientrag: 多跳问答的高效检索器. 在 *2024* 年自然语言处理经验方法会议论文集, 页面 3392–3411, 2024.

[1396] 宗长, 闫宇晨, 陆伟明, 黄艾略特, 邵建, 和 庄宇. Triad: 一个利用多角色基于 LLM 的代理来解决知识库问答的框架. *2024* 年自然语言处理经验方法会议, 2024.

[1397] Yongshuo Zong、Ondrej Bohdal 和 Timothy M. Hospedales. Vl-icl bench: 评估多模态情境学习的细节中的魔鬼. arXiv 预印本, 2024.

[1398] Yongshuo Zong, Ondrej Bohdal, 和 Timothy M. Hospedales. Vl-icl bench: 多模态情境学习的细节之魔鬼. 学习表示国际会议, 2024.

[1399] Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Yankai Chen, Chunyu Miao, Hoang Nguyen, Yue Zhou, Weizhi Zhang, Liancheng Fang,LangzhouHe, YangningLi, YuweiCao,Dongyuan Li, Renhe Jiang, and Philip S. Yu. 基于大型语言模型的人机系统综述。arXiv 预印本，2025。

[1400] 陶祖, 刘乐, 黄一飞, 孙蕾蕾, 和杜波. 基于文本属性异构图预训练语言模型. 自然语言处理经验方法会议, 2023.

[1401] Adam Zweiger, Jyothish Pari, Han Guo, Ekin Akyürek, Yoon Kim, and Pulkit Agrawal. 自适应语言模型，arXiv 预印本 arXiv:2506.10943，2025. URL https://arxiv.org/abs/2506.10943v1.