

**Hybrid AI Agent Systems Architecture for Complex Processes**

Using Knowledge Graphs to enhance AI Agent planning, decision making and process execution capabilities

**Author**: Nico Wälti

**Master Thesis / Proposal**

Spring Semester 2025

**Master of Science FHNW in Business Information Systems**

**Supervisors:** Prof. Dr. Andreas Martin

# Declaration of Authenticity

I declare that all material presented in this paper is my own work or fully and specifically acknowledged, referred to wherever adapted from other sources. We understand that if at any time, it is shown that we have significantly misrepresented the material presented here, any degree or credits awarded to us based on that material may be revoked. We declare that all statements and information contained herein are true, correct, and accurate to the best of my knowledge and belief. This paper or part of it has not been published to date. It has thus not been made available to other interested parties or examination boards.

Olten, 01. June 2025

Ein Bild, das Schwarz, Dunkelheit enthält.

Automatisch generierte Beschreibung

Nico Wälti

# Abstract

Table of Content

[Declaration of Authenticity 1](#_Toc195005014)

[Abstract 2](#_Toc195005015)

[1 Introduction 5](#_Toc195005016)

[1.1 Background 5](#_Toc195005017)

[1.2 Problem Statement 7](#_Toc195005018)

[1.3 Thesis Statement 9](#_Toc195005019)

[1.4 Possible Research Questions 10](#_Toc195005020)

[1.4.1 Main Research Question: 10](#_Toc195005021)

[1.4.2 Sub Questions 10](#_Toc195005022)

[1.5 Thesis Structure 12](#_Toc195005023)

[1.6 Delineation and Limitations 13](#_Toc195005024)

[2 Literature Review Structure 14](#_Toc195005025)

[2.1 AI Agents: Concepts, Architectures, and Components 14](#_Toc195005026)

[2.1.1 Defining Agentic AI 14](#_Toc195005027)

[2.1.2 Evolution of Agent Architectures 19](#_Toc195005028)

[2.1.3 Core Components and Design Patterns in Agentic Systems 19](#_Toc195005029)

[2.1.4 Agent Memory: RAG, Vector DB, Graph DB, Graph RAG 20](#_Toc195005030)

[2.2 AI System Architectures 21](#_Toc195005031)

[2.2.1 Agent Architecture Overview 21](#_Toc195005032)

[2.2.2 GenAI Agents, Existing Frameworks, Limitations 21](#_Toc195005033)

[2.2.3 Cognitive Architectures 21](#_Toc195005034)

[2.2.4 Autonomous Systems, Planning 21](#_Toc195005035)

[2.2.5 Multi-Agent vs. Single-Agent Systems for Complex Processes 22](#_Toc195005036)

[2.2.6 Current State of Hybrid AI Agent Systems Architecture and Frameworks 22](#_Toc195005037)

[2.3 Information Architecture: Enabling Planning and Reasoning through Knowledge Representation 22](#_Toc195005038)

[2.3.1 Foundational Concepts: Knowledge Representation for Intelligent Agents 22](#_Toc195005039)

[2.3.2 Knowledge Graphs as the Backbone for Agent Planning and Reasoning 23](#_Toc195005040)

[2.3.3 Integrating LLMs with Knowledge Graph Representations (Graph RAG and Beyond) 24](#_Toc195005041)

[2.3.4 24](#_Toc195005042)

[2.3.4 State-of-the-Art and Challenges in KG-Enhanced Agent Architectures 24](#_Toc195005043)

[2.4 Focus on Use Case Preperation 25](#_Toc195005044)

[2.4.1 Autonomous creating a Literature Review 25](#_Toc195005045)

[2.5 Research Gap Analysis 25](#_Toc195005046)

[3 Research Methodology 26](#_Toc195005047)

[3.1 Introduction 26](#_Toc195005048)

[3.2 Research Paradigm and Strategy 26](#_Toc195005049)

[3.2.1 Research Philosophy 26](#_Toc195005050)

[3.2.2 Research Approach 26](#_Toc195005051)

[3.2.3 Research Strategy: Design Science Research (DSR) 26](#_Toc195005052)

[3.2.4 DSR Cycles 26](#_Toc195005053)

[3.2.5 Research Choice 27](#_Toc195005054)

[3.2.6 Time Horizon 27](#_Toc195005055)

[3.2.7 Research Techniques and Procedures 27](#_Toc195005056)

[3.3 Design Science Research (DSR) 27](#_Toc195005057)

[3.3.1 Design Science Research Cycles 28](#_Toc195005058)

[3.3.2 Design Science Research Phases 28](#_Toc195005059)

[3.4 Evaluation Methodology 28](#_Toc195005060)

[3.4.1 High-Level Assessment (Expert Evaluation) 28](#_Toc195005061)

[3.4.2 Low-Level Assessment (Accuracy & Process Compliance) 29](#_Toc195005062)

[3.4.3 Grading Frameworks 29](#_Toc195005063)

[3.5 Data Collection & Analysis 30](#_Toc195005064)

[3.5.1 Data Sources 30](#_Toc195005065)

[3.5.2 Evaluation Metrics 30](#_Toc195005066)

[3.5.3 Validation Strategy 30](#_Toc195005067)

[3.6 Research Timeline 30](#_Toc195005068)

[3.7 Literature Review 30](#_Toc195005069)

[3.8 Further Details 34](#_Toc195005070)

[4 Agent Architecture Design 37](#_Toc195005071)

[5 Information Architecture Design 38](#_Toc195005072)

[5.1 Detail Description 38](#_Toc195005073)

[5.2 Implementation Details 39](#_Toc195005074)

[5.3 Evaluation and Results 40](#_Toc195005075)

[6 Prototype Design 42](#_Toc195005076)

[7 Result Evaluation 43](#_Toc195005077)

[8 Discussion, Limitations and Future Work 44](#_Toc195005078)

[8.1 Current Limitations: 44](#_Toc195005079)

[8.2 Future Directions: 44](#_Toc195005080)

[8.3 Bonus Discussion: 44](#_Toc195005081)

[9 Conclusion 45](#_Toc195005082)

[10 References 46](#_Toc195005083)

[11 List of digital aids 48](#_Toc195005084)

[12 Appendix (If Necessary) 49](#_Toc195005085)

[12.1.1 Acceptance Criteria for Objectives 49](#_Toc195005086)

# Introduction

This research investigates how Generative AI and (Graph) Retrieval-Augmented Generation (GraphRAG) can be effectively combined into a hybrid AI system architecture to enhance planning, explainability, reasoning, and decision-making in complex processes, using the literature review process as a case study.

The increasing availability of large language models (LLMs) and other AI technologies has opened up exciting possibilities for automating complex tasks, including the writing of scientific texts. This research explores the development of a hybrid AI agent capable of autonomously producing high-quality research papers. This agent will leverage the strengths of different AI techniques, such as symbolic AI for planning, knowledge representation and reasoning, as well as large language models (LLM) for natural language understanding and generation. This approach aligns with the emerging concept of "agentic AI," which refers to AI models that can analyze data, set goals, and take action to achieve them. Agentic AI represents a significant advancement in AI capabilities, moving beyond passive tools to actively participating in complex processes. This research aims to contribute to this field by developing an agentic AI specifically designed for the challenging task of research paper writing.

* Kurzes Summary aus Literature Gap

## Background

*(Context of the Problem – not the Problem -> hint’s the problem, but no Problem Statement)*

Large Language Models have revolutionized how we process and generate text, but different architectural approaches have emerged to extend their capabilities. In this context, five architectures can be districted:

**Plain LLM:** A standalone large language model (like the early versions of ChatGPT series from OpenAI (OpenAI, 2022)) that generates responses based solely on its pre-trained knowledge (and possibly fine-tuning). It operates as a text predictor without external tools or context beyond the prompt. Such models are powerful generalists but suffer from fixed knowledge (training data cut-off) and may confidently produce incorrect information (hallucinations) when faced with gaps. In literature reviews, a plain LLM can summarize known information but might omit recent or specific studies and invent plausible-sounding details if it “doesn’t know” something (Silva, 2023).

**LLM Agent:** An LLM agent is an LLM embedded in a loop that allows it to **plan actions and interact with external tools or the environment (Shen et al., 2023; Yao et al., 2023)**. Instead of a one-pass answer, the model iteratively decides on actions (e.g. search the web, retrieve a file, run code) and uses the results to inform subsequent steps. This paradigm (exemplified by the ReAct framework (Yao et al., 2023) or systems like AutoGPT (AutoGPT, 2024) lets the model break down tasks and handle multi-step problems. In a literature review, an LLM agent could autonomously search for papers, read or summarize them, and synthesize findings across sources. The agent’s ability to **reason and act in an interleaved manner** often yields more reliable and interpretable results than a single-pass LLM (Yao et al., 2023).

**LLM with RAG (Retrieval-Augmented Generation):** Retrieval-Augmented Generation combines an LLM with an external knowledge retrieval step. Before generating a response, the system fetches relevant documents or facts from a database or search index based on the query (Silva, 2023). The LLM then conditions its answer on this retrieved context. RAG thus **grounds the LLM’s output in up-to-date, specific information**. This approach has proven effective at reducing hallucinations and overcoming the stale knowledge of static training data. For literature reviews, RAG is extremely valuable – the LLM can pull in actual study abstracts, statistics, and quotes from papers, ensuring the review is based on evidence. In fact, RAG was introduced to address knowledge-intensive tasks by having the LLM “open-book” consult a corpus rather than rely only on parametric memory (Lewis et al., 2021).

**Agent with RAG:** This combines the two ideas above – an **LLM-based agent that uses retrieval as one of its tools**. In this architecture, the LLM not only plans actions but can specifically query document databases or search engines during its reasoning process. It essentially brings RAG into a multi-step agent loop. For example, an agent might receive a high-level prompt (“Survey recent work on X”), then **proactively issue search queries, retrieve relevant excerpts, and compile a structured review**. This architecture can handle more complex research workflows: the agent can decide when to retrieve more information and what to do with it (summarize, compare, etc.), rather than relying on a single retrieval step. By iterating between retrieval and generation, an agent with RAG can fill information gaps dynamically and verify facts as it composes the literature review.

**Agent with Graph-based RAG:** This is an advanced variant of the RAG-enabled agent that leverages a **graph-structured knowledge base** for planning or retrieval. In graph-based RAG, relationships between pieces of information are explicitly represented in a knowledge graph (KG) – for instance, a graph of papers and their citations or key concepts. The agent can traverse or query this graph to guide its understanding and action sequence (Larson, 2024). This architecture is promising for literature reviews because academic knowledge is inherently connected (papers reference prior work, concepts relate hierarchically). An agent that uses a graph can plan a **more structured exploration** of the literature: e.g. find all papers on subtopic A, then see which other subtopics they cite (graph neighbors), ensuring comprehensive coverage. Moreover, the graph provides a form of memory and reasoning scaffold – the LLM can use it to “connect the dots” between disparate information pieces​. Recent research (GraphRAG by Microsoft) showed that using LLMs to build and navigate a knowledge graph yields substantial improvements on complex Q&A tasks, outperforming standard RAG when answers require linking information across documents. For our purposes, a graph-based RAG agent could, for example, maintain a graph of concepts and sources while reviewing literature, leading to a review that is both wide-ranging and deeply interconnected (Larson, 2024).

Each of these architectures brings unique strengths for conducting literature reviews, and understanding their differences is crucial. Next, we present a comparative analysis across critical dimensions relevant to performing literature reviews, followed by in-depth analyses of each architecture’s capabilities and limitations.

## Problem Statement

**Despite the impressive capabilities of modern LLMs and their extensions, current architectures fall short when tasked with complex, multi-step processes like conducting a thorough literature review in research or business settings.** Plain LLMs often produce confident answers that cannot be trusted (hallucinations ~40% on truth benchmarks (Lin et al., 2022), and they lack access to the latest or proprietary information (Silva, 2023). Retrieval-augmented models ground answers in data, but they typically operate in a single-turn fashion and may struggle to synthesize information spread across many documents or to plan an extensive review. LLM-based agents introduce reasoning and tool use, yet **they can drift or miss information** without a guiding structure – they have no built-in memory of the knowledge domain, and thus might explore some threads deeply while neglecting others (coverage imbalance). Moreover, autonomous agents raise concerns about reliability and safety (prone to prompt injection attacks (Liu et al., 2024) if not carefully managed, and can get stuck in loops or go off-topic).

In complex research or business analysis, we need an AI system that can **comprehensively gather information, accurately synthesize it, and transparently justify its findings**, all while respecting constraints like data privacy and user control. No single existing architecture perfectly satisfies this:

* **A plain LLM** is too unreliable and opaque for high-stakes analysis. It may miss recent developments entirely, or output plausible-sounding falsehoods (Silva, 2023). For a literature review, this could mean citing non-existent studies or missing critical papers – unacceptable in academia or data-critical business reports.
* **A RAG system** ensures each claim is backed by sources (AWS, n.d.), but it has **no inherent planning ability**: it won’t decide by itself which subtopics to delve into or how to organize the review. It answers the query given, and if the query is broad (“summarize 10 years of research on X”), it may produce a generic summary based on the top few hits, potentially overlooking important diversity. It also typically has a fixed retrieval cutoff (say top 5 documents), risking incomplete evidence if the literature is extensive.
* **An LLM agent** (like AutoGPT or ReAct) brings reasoning but, without structured guidance, can be **inefficient or inconsistent**. It might search the web multiple times with similar queries or not know when to stop searching vs. start writing. Different runs might yield different results (non-repeatable) if the agent’s chain-of-thought diverges. Critically, if the agent isn’t grounded by retrieval in each step, it could start to hallucinate intermediate conclusions which skew the whole process.
* Even an **agent with RAG** has limitations. It retrieves information as needed and can iterate, but it typically treats the knowledge base as a flat search space. It doesn’t inherently know the **global structure of the knowledge**. It may not realize that “to fully cover topic X, there are A, B, C sub-areas” unless it stumbles upon them or the user enumerated them. Essentially, its world model is formed on-the-fly and can be myopic, depending on the luck of search results. There’s also a challenge in explainability: while it can cite sources for final answers, the internal decision of why search this or why not look into that may not be easily interpretable without a framework.

**Using graph databases as a foundation for agent planning is a promising approach to bridge these gaps.** A knowledge graph can encode relationships between concepts, papers, findings, and even user-specific business data in a structured form. By giving the agent access to this graph:

* The agent can systematically explore the research landscape. The graph acts as a map: ensuring that if topic X has 5 sub-nodes (subtopics or key papers), the agent will “visit” each of them instead of getting stuck on just one. It won’t forget to cover a branch because the graph makes it explicit. Essentially, the agent’s plan can be derived from traversing the graph (e.g., do a depth-first search to cover all relevant nodes). This addresses the **coverage** issue – the process becomes more **repeatable** and thorough, guided by the graph structure rather than chance.
* The graph provides a form of **memory and state** that the agent can refer to at any time, independent of the token window. No matter how complex the conversation or analysis gets, the agent can query the graph for an overview or a specific connection, anchoring its reasoning. This reduces the risk of the agent losing context or contradicting itself over a long session.
* Graph-based retrieval can **uncover non-obvious connections** (like linking two concepts via a shared factor) that a keyword search might miss (Godsey, 2024). This means the agent could discover insights (e.g., “Method A from field Y was also applied in field Z”) that current RAG might not surface if no single document states it plainly. For a literature review, such cross-cutting insights are valuable.
* Planning with a graph also enhances **explainability**: the agent’s path in the graph can be shown as the reasoning trace. For instance, “Topic X -> key paper by Smith -> follow its citation to Jones’s paper -> compare outcomes.” This is intuitive to humans (it mirrors how we navigate references). It holds the agent accountable to follow logical links, which fosters trust.
* In terms of **tool integration**, a graph can act as a filter for the retriever: the agent might use the graph to get a list of relevant documents, then only retrieve those from the text database. This ensures high **precision** in retrieval, as seen with the LinkedIn KG-RAG (which improved answer quality by a large margin) (Xu et al., 2024).
* It can also protect **privacy**: the graph might contain metadata that flags certain info as sensitive, and the agent can be instructed to avoid or anonymize those parts in its output, achieving compliance more systematically than ad-hoc prompt-based filtering.
* From a development perspective, a graph-based approach provides a **framework to incorporate human knowledge or rules** into the AI’s operation. For example, a domain expert could curate part of the knowledge graph (say, ensure seminal works are linked in), and the agent will naturally include those because they’re prominent nodes. This synergy of human-curated structure and AI-driven fill-in of details is powerful and could be part of the thesis exploration.
* Furthermore, the Database can be updated with new Knowledge with a few nodes and therefore provide a grounding to the AI Agent with a “Training Data Set of 1” rather than thousands or millions of training data required to teach an Agent about new Information.

**In summary, current architectures struggle with the scale, complexity, and reliability that real-world literature reviews demand.** They either lack memory (plain LLM), or lack planning (simple RAG), or lack a global view of knowledge (even agents with sequential RAG). This leads to issues like missing important pieces of information, inability to justify why a piece of info was included or omitted, and inconsistent results on repeated runs. A graph-based RAG agent aims to overcome these by giving the agent a scaffold of the knowledge domain up front. Preliminary evidence (e.g., GraphRAG, ESCARGOT) indicates such systems can dramatically improve question answering and reliability (Larson, 2024; Matsumoto et al., 2025).

## Thesis Statement

The thesis statement derives from the identified research problem and functions as the working hypothesis for this study. The findings of this thesis will be utilized to either confirm or refute the thesis statement in the final chapter.

*It is possible to design a hybrid agentic architecture, combining LLM- and knowledge graph-based planning, that can autonomously conduct complex processes, such as multi-step research tasks*

* + *Beschreiben was research tasks sind => wie deep research.*

## Possible Research Questions

The research questions originate from the identified problem and the thesis statement. Building upon the primary research questions, additional research questions have been formulated and structured in alignment with the phases of the design science research approach. The table below presents an overview of the research questions addressed in this thesis alongside their corresponding research objectives.

### Main Research Question:

How can a hybrid agentic architecture, combining large language models (LLMs) and knowledge graph-based planning, be designed to autonomously perform complex processes such as literature reviews?

What information elements and system architectural components are essential to design a hybrid agentic architecture, combining LLM- and knowledge graph-based planning, capable of autonomously conducting complex, multi-step research tasks?

### **Sub Questions**

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Sub Question** | **ID** | **Objective** |
| **1** | **Chapter 2: Literature Review**  **SQ1:** What are the key architectural elements and design patterns of contemporary AI agents? | 1.1 | Provide a comprehensive overview of components and subsystems used in existing agentic systems (Chapter 2.1) |
| 1.2 | Analyze and synthesize state-of-the-art architectures of AI agents (Chapter 2.2) |
| 1.3 | Analyze and synthesize existing approaches and information architectures used in AI planning systems, with a focus on how knowledge representation, reasoning mechanisms, and data structures support autonomous multi-step decision-making. (Chapter 2.3) |
| 1.4 | Analyze and synthesize existing approaches for autonomous Literature Review creation by AI. (Chapter 2.4) |

|  |  |  |  |
| --- | --- | --- | --- |
| **2** | **Chapter 4: Agent Architecture Design**  **SQ 2:** What is the optimal system architecture for a hybrid AI agent that integrates a knowledge graph (for planning and reasoning) with a Large Language Model (for language understanding and generation), specifically for the task of autonomously generating a comprehensive and methodologically sound literature review? | 2.1 | Compare different architectures against each other (with different scaled content scopes). |
|  | 2.2 | (D1) Design an Agent Architecture that supports autonomous agents that is able to perform complex processes. |
|  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **3** | **Chapter 5: Information Architecture Design**  **SQ 3:** What is the optimal information architecture (schema design, data representation, and querying mechanisms) for the knowledge graph component of the hybrid AI agent, to effectively support:   * Representation of diverse knowledge types (concepts, relationships, facts, arguments). * Efficient retrieval of relevant information for planning and knowledge-based reasoning. * Reasoning about relationships, contradictions, and research gaps. * Incremental updates and maintenance of the knowledge base? | 3.1 | (D2) Design an information architecture that supports autonomous agent planning in complex processes. |

|  |  |  |  |
| --- | --- | --- | --- |
| **4** | **Chapter 7: Results and Evaluation** | 5.1 | Select/adapt a grading framework or an Agent paper generation Benchmark to evaluate AI vs. human-generated reviews and assess logical coherence of literature review and expert evaluation. |
|  | **SQ 4: Performance and Quality Benchmarks:** How does the quality and efficiency of a literature review generated by the proposed hybrid AI agent system compare to:   * A literature review produced by a human expert (e.g., a researcher or advanced student)? * A literature review generated by a baseline AI system (e.g., an LLM-only approach, a deep research system)? |
|  | 5.2 | Create high and low level Testcases, to assess the correct information processing capabilities and high-level connection identification capabilities (considering DB scaling levels). |

## Thesis Structure

This thesis is organized into seven chapters, following the five Design Science Research (DSR) phases as outlined by Dresch et al. (2015): Awareness, Suggestion, Development, Evaluation, and Conclusion. Each chapter systematically builds upon the previous one, ensuring a coherent progression from problem identification to solution development and assessment.



Figure 1: Thesis Map

## ****Delineation and Limitations****

**Scope:** This research focuses specifically on the creation of a structured suggestion for an information architecture to support AI agent planning in complex processes. The deliverable does not include the development or implementation of the code or system, which will be used for the proof of concept. Additionally, the deliverable excludes any generated literature reviews that may be used to test or demonstrate the proposed information architecture.

# Literature Review

Was mache ich, wenn ich ein spannendes Paper finde, das einen gewissen Teil meiner Literatur Recherche zu teilen bereits abdeckt? Lese ich die Papers von dort – wie weit darf ich mich orientieren?

This chapter aims to provide the necessary background from the current body of knowledge concerning AI agents, the system architectures they operate within, and the information structures that enable their complex reasoning and planning capabilities, particularly in the context of automating literature reviews. The initial sections focus on the state-of-the-art regarding AI agents (Section 2.1) and broader AI system architectures (Section 2.2), including cognitive and multi-agent approaches. Subsequently, the chapter explores information architecture (Section 2.3), emphasizing knowledge representation techniques like knowledge graphs and their synergy with Large Language Models (LLMs). The specific application context and challenges of automating literature reviews are then examined (Section 2.4). Finally, this chapter culminates in a research gap analysis (Section 2.5), identifying the specific contribution this thesis aims to make.



Figure 2 - Chapter Map Literature Review

## AI Agents: Concepts, Architectures, and Components

This chapter focuses on AI agents, exploring the fundamental concepts, architectures, and components that underpin their functionality. The chapter commences with a foundational definition of agentic AI in Section 2.1.1, clarifying its core attributes and differentiating it from non-agentic systems. Section 2.1.2 then examines the evolution of agent architectures, highlighting the significant impact of Large Language Models (LLMs). Subsequently, Section 2.1.3 dissects the essential components and common design patterns that constitute agentic systems. The chapter concludes with Section 2.1.4, which investigates various approaches to agent memory, a vital aspect for sustained operation and learning. This overview serves as a basis for the subsequent exploration of broader AI system architectures and information structures.

### Defining Agentic AI

Agentic AI refers to systems that autonomously operate within a defined environment, embodying characteristics like proactivity, reactivity, and the capability to interact with surrounding elements. Agency in AI emphasizes the autonomy of systems to make decisions based on their perceptions rather than relying solely on external instructions. Specifically, agentic systems are designed to assess their environment, respond to stimuli accordingly, and execute actions that align with predefined goals (Svensson & Keller, 2024). This capability is particularly crucial in domains where response timing and environmental adaptation significantly enhance performance outcomes, such as in collaborative human-AI relationships (Li et al., 2024).

The term AI agent was not defined by a single publication but rather evolved over time through various contributions in artificial intelligence research. One of the most influential sources in shaping the concept was written by Russell and Norvig (1995), which introduced a systematic framework for understanding agents as entities that perceive their environment and act rationally to achieve goals. It formalized the agent concept through the agent function, agent program, and the PEAS framework. Complementing this, **Wooldridge and Jennings (1995)** provided a comprehensive theoretical foundation in their paper, defining key agent characteristics such as autonomy, social ability, reactivity, and proactiveness, and distinguishing between different types of agents in the context of multi-agent systems. Together, these works significantly shaped the modern understanding and application of AI agents.

A critical distinction between agentic AI and passive AI models lies in their operational frameworks. Goal-oriented AI agents are endowed with clear objectives and the ability to choose actions based on their understanding of the environment, contrasting with foundational models like base LLMs that primarily focus on language prediction and generation without an inherent capacity for task completion or proactive engagement (Li et al., 2024). While LLMs excel in generating coherent and contextually relevant text, they lack the active decision-making processes and functional goals that characterize agentic systems.

The definition of an AI agent varies significantly across different sources, reflecting diverse perspectives on their core functionalities and applications. These definitions range from foundational concepts of perceiving environments and acting to achieve goals, to more specialized views emphasizing human interaction, collaborative roles in specific domains like healthcare or IT, cognitive capabilities, and considerations like user privacy or predictability.

|  |  |  |
| --- | --- | --- |
| **Definition Source** | **Definition of AI Agent** | **Concepts used to define AI Agents (Combined from Definition & Snippet)** |
| (Anthropic, 2024) | AI Agents are systems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks. |  |
| Stratis Tsirtsis, Manuel Gomez-Rodriguez, Tobias Gerstenberg | A computational entity that can perform tasks and collaborate with humans, where the accountability of actions, especially in unforeseen circumstances, is evaluated through cognitive processes of responsibility judgments. | Computational entity, task performance, human-AI collaboration, accountability, responsibility judgments, cognitive processes, action expectations, counterfactual considerations, contributions to outcomes, interdependent actions, generative model of agent behavior, unexpected actions. |
| Yvonne A. Farah, Michael C. Dorneich | Entities that work alongside humans in teams (Human-Autonomy Teams) toward common goals and engage in teamwork processes including planning and coordination. | Entities working with humans, Human-Autonomy Teams (HATs), common/shared goals, teamwork processes, teaming capabilities, technological skills, goal specification, planning, mission analysis, coordination, monitoring, backup behaviors, support levels (passive, reactive, proactive), interdependence, effective communication, initiative taking. |
| Carolin Wienrich, Astrid Carolus, David Roth-Isigkeit, Andreas Hotho | Adaptive systems that deliver advice and interact with users while considering their individual characteristics. | Adaptive systems, complexity, advisory role, interaction with users, consideration of individual characteristics, explainable AI (XAI), human-centered design, attribution of human characteristics, trust, understanding, workload. |
| Raja Sengupta, Renée Sieber | A computational entity that operates autonomously to perform tasks by perceiving its environment and taking actions based on predefined algorithms, potentially specialized for geographic contexts. | Computational entity, autonomous operation, task performance, environment perception, action based on algorithms, AI research principles, explicit incorporation of geographic context, geographic data models, diversity in applications (e.g., geospatial agents, artificial life agents, software agents). |
| Peter R. Lewis, Ştefan Sarkadi | An entity functioning based on abstracted definitions of intelligence, potentially lacking certain qualities inherent to human cognition, designed to perform tasks delegated from humans. | Intelligent agents, abstraction, cognitive features (potentially incomplete compared to humans), embodied AI technologies, performing human-delegated tasks, potential imbalance of insight/understanding, reflection (often missing), agent-based modelling (ABM), multi-agent systems (MAS). |
| James Chao, Wiktor Piotrowski, Roni Stern, Héctor J. Ortiz-Peña, Mitch Manzanares, Shiwali Mohan, Douglas S. Lange | An agent that autonomously operates within real-world environments, relying on automated planners to optimize actions toward goals while adapting to changing conditions (novelties). | Autonomous agents, real-world environment operation, automated planners, optimal actions, goal achievement, objective function optimization, adaptation to novelties/changing conditions, environment dynamics, schedulers, execution engines, continuous time-space coordination, concurrent actions. |
| Alexis Yim, Annie Peng Cui, Michael Walsh | Entities characterized by features (like cuteness) that evoke emotional connections, enhancing consumer trust, attachment, and influencing behavior. | Emotional connection evocation, cuteness (baby schema, whimsical), trustworthiness, attachment, consumer behavior influence, satisfaction, commitment, purchase intention influence, popularity. |
| Sebastian Köhler | Robots and computer programs capable of acting and deciding independently (autonomous action/decision-making), sometimes under supervision. | Robots, computer programs, autonomous action, independent decision-making, supervised agency, potential for harm, responsibility assignment, collaborative agency (as a contrasted concept), use as instruments. |
| Aaron M. Garvey, TaeWoo Kim, Adam Duhachek | Entities perceived as having weaker intentions compared to human agents, which affects consumer responses to offers; anthropomorphism can modify this perception. | Perceived intentions (weaker than human), influence on consumer responses (to offers), anthropomorphism, interaction with consumers, administering offers, lack of inferred selfish/benevolent intentions, potential mitigation of blame/enhancement of credit. |
| Ali Ladak, Steve Loughnan, Matti Wilks | Entities perceived by humans as possessing moral agency (ability to do right/wrong) and moral patienthood (capacity to be the target of right/wrong actions), based on cognitive and social capacities. | Perceived entities, moral agency, moral patienthood, cognitive capacities, social capacities, psychological perceptions, mind perception, potential overlap with human morality attributions, potential unique factors (e.g., sci-fi fan identity), chatbots. |
| John E. Laird | Systems integrating real-time decision-making, planning, natural language understanding, metacognition, theory of mind, mental imagery, and various learning forms, often within a cognitive architecture like Soar. | Cognitive architecture components, real-time decision-making, planning, natural language understanding, metacognition, theory of mind, mental imagery, multiple forms of learning (reinforcement, episodic, semantic), problem-solving, adaptation, knowledge encoding/application, symbol structures, non-symbolic reasoning (spatial visual system). |
| Stephanie Milani, Arthur Juliani, Ida Momennejad, Raluca Georgescu, Jaroslaw Rzepecki, Alison Shaw, Gavin Costello, Fei Fang, Sam Devlin, Katja Hofmann | Systems designed to generate behavior that mimics human-like actions, such as navigation in video games, assessed via comparison and crowd-sourcing. | Human-like behavior generation, navigation behavior (video games), comparison with baseline AI/humans, crowd-sourced assessments, Turing Test passing capability, goal-directed navigation, reinforcement learning application, human-centered computing (HCI) aspect. |
| Chathura Gamage, Matthew Stephenson, Vimukthini Pinto, Jochen Renz | An entity designed to interact with an environment (e.g., a game) to achieve specific goals, employing strategies that can sometimes be deceived by complex task structures or environments. | Interaction with environment, goal achievement, strategy employment, gameplay (e.g., Angry Birds), susceptibility to deception, task complexity handling, planning vs. reactive shots, stability, solvability, evaluation bias exploitation, reward structure sensitivity. |
| Daniel B. Shank, Mallory North, Carson Arnold, Patrick Gamez | Perceived entities whose moral characteristics (virtues/vices) are assessed based on behaviors, often judged differently than humans due to perceptions of mind (experiential and agentic). | Perceived entities, attribution of virtuous/unvirtuous characteristics, assessment based on behavior, comparison to human judgments, mind perception (experiential mind: emotions, sensations; agentic mind: intentions, actions), behavior-to-character attributions, virtue ethics domains. |
| Scott Schanke, Gordon Burtch, Gautam Ray | A system, often voice-based and conversational, that engages with users, featuring personalized and autonomous capabilities affecting user trust. | Conversational interaction, voice-based agents, autonomous nature, personalization (e.g., voice cloning), consumer trust influence, disclosure of autonomy, homophily perception. |
| Andrea Roli, Johannes Jaeger, Stuart Kauffman | A system capable of performing tasks autonomously using algorithms, physical characteristics, and environmental interactions, distinct from biological organisms in its ability to handle novel affordances. | Autonomous task performance, algorithmic operation, leveraging physical characteristics/environment, situational reasoning, goal selection, dealing with ambiguity (limited compared to organisms), affordance identification/exploitation (limited, cannot leverage new ones algorithmically), distinction from organismic agency, limits on AGI. |
| Kun Wang, Xi Yang, Shuo Yang, Xian Du, Ruijing Shi, Wendong Bai, Yu Wang | A system utilizing machine learning techniques to extract and analyze features from medical images (e.g., ultrasound) to assist in diagnostic processes. | Machine learning application, feature extraction (medical images), data analysis, diagnostic assistance (e.g., breast cancer detection), predictive analysis, multimodal data integration (e.g., ultrasound types), improving diagnostic efficacy. |
| Daniel Müller, Tshilidzi Marwala | Systems that make decisions, potentially based on relative net utility (aligning with limited reward perception like humans) rather than nominal utility (infinite reward perception). | Decision-making systems, utility processing (relative net utility vs. nominal utility), comparison to human behavioral economics, risk aversion modeling, potential for market efficiency improvement, handling subjective utility. |
| Filip Thiessen, Thierry Tondu, Véronique Verhoeven, Guy Hubens, Gunther Steenackers, Wiebren Tjalma | An automated system capable of enhancing medical procedures (e.g., perforator detection) through advanced data analysis and decision-making algorithms, improving reproducibility and efficiency. | Automation, data analysis, decision-making algorithms, enhancing medical procedures (e.g., microsurgical reconstruction mapping), improving accuracy/reproducibility, workflow enhancement, decreasing human input time, potential use with imaging techniques (CTA, DIRT). |
| O. V. Kubryak, Sergey V. Kovalchuk, Nadezhda G. Bagdasaryan | Intelligent systems capable of performing complex tasks with predictive ability and substantial computing power in uncertain environments, evaluated based on cognitive behaviors and context. | Intelligent systems, complex task performance, predictive ability, computing power ("brute force"), uncertainty management, environmental context adaptability, cognitive behavioral characteristics assessment, anticipatory ability, "common sense" application, adequacy of solutions. |
| Silvana Hinsen, Peter Hofmann, Jan Jöhnk, Nils Urbach | Systems that enable interactions with humans, impacting various industries, requiring purposeful design based on interaction types and human experiences. | Enabling human interactions, impacting industries/daily lives, purposeful design requirement, consideration of human experiences/perceptions, interaction characteristics/dimensions, interaction types based on agent characteristics, use cases analysis, practitioner insights. |
| Marlene Blaß, Henner Gimpel, Philip Karnebogen | Components within a larger framework (e.g., health care services) that utilize AI technologies to interact with data and support service delivery, classified within a taxonomy. | Components of larger systems, utilization of AI technologies, interaction with data, support for service delivery (e.g., health care), role within AI framework/taxonomy, impact on outcomes (e.g., health impact), design characteristics, classification into archetypes. |
| Haruno Suzuki, Jingwen Zhang, Diane Dagyong Kim, Kenji Sagae, Holli A. DeVon, Yoshimi Fukuoka | An entity, often a chatbot, perceived as either human or artificial based on the humanness of its messages and interaction characteristics, influencing user perception and interaction. | Chatbot entity, identity perception (human vs. artificial), influence of message humanness, interaction characteristics, human-like communication style, effectiveness of messages, user attitude influence, establishment of correct perceptions. |

Table 1: AI Agent Definitions

As these examples illustrate, the definitions of AI agents diverge significantly based on the specific facet being emphasized. Key dimensions of variation include the degree of required autonomy and cognitive capability (from basic reactivity to complex reasoning), the nature and importance of human interaction, the specific application domain (such as healthcare, IT, or personal assistance), and operational considerations like privacy, predictability, or the underlying technology (like LLMs directing processes). This highlights that the concept of an "AI Agent" is often tailored to the context in which it is being discussed or developed.

Despite their nature as technical tools, the frameworks used to define and describe Artificial Intelligence (AI) agents are notably diverse. These definitional approaches often originate from a broad spectrum of fields, encompassing both technical and non-technical disciplinary perspectives. Conversely, influential industry actors, including large technology firms like Microsoft and Anthropic, typically emphasize technical attributes, such as agent capabilities and system architecture, in their characterizations. Krishnan (2025) highlights that definitions of AI agents are varying significantly, depending on their specific capabilities and architecture. Unlike conventional AI systems, which typically operate within predefined constraints and necessitate explicit instructions for task execution, AI agents demonstrate significant autonomy in goal-oriented behavior. This architectural distinction, as conceptualized by Microsoft (Ray, 2024), positions agents as superstructural layers interacting with foundational language models; they engage in observation, data acquisition, model input provision, and collaborative action plan generation. This is supported similarly by Anthropic, which categorize AI Workflows and AI Agents, depending on their degree of freedom and summarize both categories as AI Agent systems. such a configuration empowers agents with advanced functionalities, including the decomposition of complex problems into manageable sub-tasks, reasoning over available information, strategic tool utilization, learning from feedback mechanisms, and maintaining contextual coherence across interactions.

In the following Table the used concepts are summarized and counted to better see the focus points of the definitions.

|  |  |  |
| --- | --- | --- |
| **Concept Category / Theme** | **Frequency (# References)** | **Key Included Concepts** |
| Human-AI Interaction / Collaboration / User Focus | 11 | Collaboration, user interaction, HCI, human-centeredness, user perception, conversational interaction, human-AI teams. |
| Perception (Environment & Social/User) | 9 | Environment perception/interaction, user perception, perceived intentions/identity, mind perception, social capacities. |
| Task Performance / Goal Orientation | 8 | Task execution, goal achievement/specification, problem-solving, gameplay, supporting service delivery, objective functions. |
| Cognition / Intelligence / Mind Concepts | 8 | Cognitive processes/capacities/behaviors, intelligence, reasoning, metacognition, theory of mind, knowledge representation, cognitive architecture. |
| Decision-Making / Planning / Strategy | 7 | Decision-making, planning, strategy, automated planners, action selection. |
| Ethics / Responsibility / Trust / Morality | 7 | Responsibility, accountability, trust, trustworthiness, ethics, moral agency/patienthood, blame/credit. |
| Autonomy / Independence | 6 | Autonomy, independent action/decision-making, self-direction, supervised agency. |
| Learning / Adaptation / Adaptability | 6 | Learning (various forms), adaptation, adaptability, handling novelty/change, adaptive systems. |
| Human-like Qualities / Anthropomorphism | 6 | Human-like behavior/communication, anthropomorphism, cuteness, emotional connection, message humanness. |
| Data Processing / Analysis / Prediction | 5 | Data analysis, feature extraction, prediction, utility processing. |
| Embodiment / Robotics | 4 | Robots, embodied AI, physical characteristics/interaction. |
| Context / Environment | 4 | Environmental context, real-world operation, specific domains (geospatial, healthcare, games), uncertainty. |
| Healthcare / Medical Applications | 4 | Diagnostic assistance, medical image analysis, healthcare chatbots, health service components. |
| Teamwork / Coordination | 3 | Teamwork processes, coordination, multi-agent systems (MAS). |
| Automation | 2 | Automation, automating processes, replacing human input. |
| Explainability / Transparency | 1 | Explainable AI (XAI). |

Table 2: List of Concepts used for Definitions

The Table systematically breaks down the components of AI agent definitions found in the literature. It organizes these components into conceptual categories, shows their relative frequency based on references, and provides illustrative keywords and concepts for each category, ranging from an often-used Human-AI Interaction concept to lesser used Explainability concept.

Soll ich versuchen eine eigene Definition aus den häufigsten Konzepten zu kreieren oder eine auszuwählen, die die Basis für meine Arbeit macht?

### Evolution of Agent Architectures

The architectural evolution of AI agents, particularly leveraging LLMs, has progressed from relatively static models to dynamic frameworks that integrate tools and feedback loops. Standalone LLMs are limited by their fixed knowledge base and inability to act autonomously; they merely generate text outputs based on learned patterns without context-specific adaptability (Li et al., 2024). This limitation necessitated the development of architectures enabling interactions with external resources, such as API calls or prompting systems, to introduce flexibility and responsiveness in agent behavior (Kotseruba & Tsotsos, 2018, Yu et al., 2024).

The definition of "LLM Agent" architecture emphasizes a continuous loop of reasoning, action selection, and observation. In such architectures, agents utilize LLMs to formulate responses while concurrently evaluating the consequences of their actions in a feedback-driven manner. Influential frameworks incorporate mechanisms for reasoning and decision-making within agentic processes, establishing a foundation for subsequent advancements, including enhanced understanding and operational effectiveness in complex environments (Li et al., 2024).

The integration of Retrieval-Augmented Generation (RAG) represents a substantial enhancement in agent grounding. By accessing external, contemporary, or proprietary information before response generation, agents equipped with RAG can provide more accurate and contextually relevant outputs. The "Agent with RAG" architecture illustrates a scenario wherein retrieval is a selectable tool within the agent's operational loop, contributing to a significant increase in reliability and effectiveness in diverse applications (Yuan et al., 2022; , (Kobuki et al., 2023).

Models that incorporate knowledge graphs, such as "Agent with Graph-based RAG," represent a notable progression in sophistication. These architectures leverage structured knowledge to facilitate enhanced performance, allowing agents to navigate complex datasets and infer relationships effectively. This has implications for various applications, including strategic decision-making and collaborative tasks within multi-agent frameworks (Chen et al., 2017).

### Core Components and Design Patterns in Agentic Systems

Central to constructing capable AI agents are the functional modules defining their operational capabilities. Planning capabilities embedded within agent architectures include task decomposition, goal management, and strategic thinking, which help agents effectively manage complex tasks and varied objectives (Chohan et al., 2021). Reasoning mechanisms, such as logical deduction, chain-of-thought patterns, and causal inference, enable agents to interpret information and make informed decisions, enhancing their effectiveness in varied contexts (Kovač et al., 2024).

Agents also require efficient memory systems to manage both short-term contexts and long-term knowledge. Short-term context management involves understanding conversation histories and immediate observations, often constrained by token limits in LLMs. For long-term memory storage, technologies such as vector stores, databases, and graph structures enable agents to recall interactions and acquired knowledge, enriching the learning process and facilitating more informed decision-making (Lee et al., 2020; , Ruan et al., 2022).

A crucial aspect of agent functionality is the ability to select, parameterize, and invoke external tools, such as APIs or search engines. Architectural considerations must ensure that agents can trigger these tools effectively and integrate their outputs into the operational framework cohesively (Zarzà et al., 2023). For agency in complex environments, agents must adeptly perceive and interact with both digital and physical surroundings, allowing them to navigate multifaceted scenarios effectively (Kobuki et al., 2023).

Among the foundational design patterns, the Reason-Act (ReAct) pattern interleaves thought processes with actions, enabling agents to alternate between reasoning and execution dynamically. Reflection or self-critique patterns further allow agents to evaluate past actions to enhance future performance, emphasizing continuous improvement (Lewis & Sarkadi, 2023). Other prevalent patterns in agent design include Plan-and-Execute frameworks and collaborative mechanisms for multi-agent cooperation, which establish structured approaches to achieving shared goals while optimizing resource allocation and operational efficiency (Vila et al., 2022).

In conclusion, the landscape of AI agents has evolved significantly, with advancements in architecture and core components that enable agents to exhibit higher levels of autonomy, adaptability, and efficacy. The integration of LLMs and structured methodologies underscores the growing sophistication of these systems, reinforcing the importance of constant innovation in the field of artificial intelligence.

### **Agent Memory: RAG, Vector DB, Graph DB, Graph RAG**

* Detailed Analysis of RAG: How does it work? What are its limitations (e.g., hallucination, retrieval accuracy)? This deep dive into RAG is essential as it's a core component.
* Comparison of Vector DBs and Graph DBs: When is each more appropriate for knowledge representation in AI, specifically for literature review tasks? This comparison is key to justifying the use of a graph database.
* Hybrid Approaches Combining Vector and Graph Databases: Explore the potential of combining these database types for more sophisticated knowledge representation and reasoning. For example, Neo4j has started to introduce Vector Indexes. This is directly relevant to the "hybrid" aspect of the thesis.
* Semantic Search and its Relationship to RAG and these DBs: Define and analyze how semantic search is facilitated by these technologies and how it benefits literature retrieval.

## ****AI System Architectures****

### ****Agent Architecture Overview****

* Frage: Soll ich die Frameworks/Technologien auch bereits vergleichen (Stärken / Schwächen) und auf Nützlichkeit / Relevanz beurteilen? Evtl. bereits entscheiden, welche Technologien/Komponenten für die Design-Phase in Frage kommen, resp. für den Vergleich verwendet werden können.
* Workflows, Multi Agent Systems,
  + - <https://www.anthropic.com/research/building-effective-agents> - This provides a good starting point for understanding agent architectures and workflows.

### ****GenAI Agents, Existing Frameworks, Limitations****

* Survey of Existing Frameworks: LangChain, AutoGen, Hugging Face Agents, etc. Compare their features, capabilities, and limitations, specifically in the context of research tasks or literature review.
* Challenges in Agent Coordination and Communication: This is a key limitation in multi-agent systems, particularly relevant to your focus on complex processes.
* Evaluation Metrics for GenAI Agents: How do we assess their performance, reliability, and safety? This is important for later comparisons.
* Trends and Best Practices for Implementations: What are the current best practices for building and deploying GenAI agents?

### ****Cognitive Architectures****

* Different Types of Cognitive Architectures: SOAR, ACT-R, LIDA, etc. Analyze their principles, strengths, weaknesses, and suitability for your use case (integrating diverse AI components for literature review). Include a clear comparison table/matrix.
* Review ACT-R, Soar, CLARION, focusing on how procedural and declarative memory, chunking, and production rules model human-like cognitive processes, and how these could be leveraged for literature review.
* The Role of Memory (working, episodic, semantic) in Cognitive Architectures: How do these memory systems influence agent behavior, and how can they be used to improve the agent's ability to understand and synthesize research?
* Incorporating GenAI within Cognitive Architectures: Explore how GenAI models (like LLMs) can be integrated into these architectures to enhance their capabilities. This is a key point for your "hybrid" approach.

### ****Autonomous Systems, Planning****

* + - Different Approaches to Planning in AI: Classical planning, hierarchical task networks (HTNs), reinforcement learning-based planning, etc. Explain these approaches and their relevance to literature review.
    - Challenges in Real-world Planning: Uncertainty, partial observability, dynamic environments. Discuss how these challenges manifest in the context of literature review.
    - Integrating Planning with GenAI: How can GenAI models be used for more flexible and adaptive planning in the literature review process?
    - Explainable Planning: How can the agent explain the decisions it took?

### **Multi-Agent vs. Single-Agent Systems for Complex Processes**

* + - This section directly addresses the multi-agent aspect of your thesis.
    - Define different types of sub-agents and their potential roles in an autonomous literature review system (e.g., a "retrieval agent," a "summarization agent," a "planning agent," a "synthesis agent").
    - Compare centralized vs. decentralized control mechanisms for coordinating these sub-agents, drawing on existing multi-agent system literature. Discuss the trade-offs.
    - Discuss common conflict resolution strategies in multi-agent systems (e.g., negotiation, voting, priority-based systems) and how they could be applied to literature review.
    - Present models (even conceptual or borrowed from other papers) of how multi-agent coordination could work for literature review, specifically addressing the challenges of this task.

### ****Current State of Hybrid AI Agent Systems Architecture**** and Frameworks

* + This section is crucial for providing a comprehensive overview of the existing landscape.
  + Summary of recent papers/projects showcasing hybrid systems (specifically those relevant to complex tasks or, ideally, literature review).
  + Clear articulation of the limitations and challenges identified in the literature. This sets up the problem statement.
  + Discussion of the gap between theory (what should be possible with hybrid systems) and practice (what's actually been achieved).
  + Brief case studies (even if they're just summaries of papers) of both successes and failures. This provides concrete examples.

## **Information Architecture: Enabling Planning and Reasoning through Knowledge Representation**

### **Foundational Concepts: Knowledge Representation for Intelligent Agents**

**Topics:** - defining KR and its importance specifically for agentic systems performing complex tasks.

- Defining Knowledge Representation in the context of AI agents: Why do agents need structured knowledge beyond raw data or LLM parametric memory?

- Goals of KR for complex processes (like literature reviews): Enabling planning, multi-step reasoning, knowledge synthesis, retrieval, explainability, consistency, and handling domain-specific constraints.

- Overview of KR Paradigms: Symbolic (Logic, Rules, Ontologies, **Knowledge Graphs**): Strengths in explicit structure, reasoning, explainability.

- Sub-symbolic/Distributed (Embeddings, Vector DBs): Strengths in semantic similarity, handling unstructured data.

- Justify the need for a hybrid approach: Leveraging KGs for structure/reasoning and LLMs/vectors for semantics/flexibility.

\* Identify specific knowledge types required for the literature review use case: Bibliographic metadata (papers, authors, venues), conceptual relationships (topics, subtopics, keywords), research artifacts (methodologies, datasets, findings, claims, arguments), provenance (sources), and process knowledge (steps in a lit review). (Directly relevant to SQ3a).

### **Knowledge Graphs as the Backbone for Agent Planning and Reasoning**

Deep dive into KGs, explaining why they are suitable and how they facilitate the core agent capabilities needed.

**- Why Knowledge Graphs for Literature Review & Complex Processes?** Advantages: Representing interconnectedness, explicit semantics, traversability for exploration, structured querying, integrating heterogeneous information. Compare briefly against limitations of relational DBs or pure vector DBs for this structured reasoning task.

- **Core KG Concepts & Design Principles:** Nodes, Edges, Properties, Schema/Ontology. Discuss design considerations for representing the knowledge types identified in 2.3.1. Importance of schema for consistency and queryability.

**Enabling Planning via KGs:** Representing process workflows, task dependencies, and goals within the graph. \* Using graph traversal algorithms (e.g., pathfinding, neighbourhood exploration) to guide the agent's planning and exploration of the literature space. \* Example: Planning which sub-topics to explore based on graph structure, identifying prerequisite knowledge.

-**Enabling Reasoning via KGs:** Deductive reasoning: Applying rules or querying specific graph patterns (e.g., "Find all papers that cite X and use methodology Y"). \* Inductive reasoning: Identifying emerging themes or patterns by analyzing clusters or connectivity (e.g., "Concepts A and B are frequently discussed together"). \* Abductive reasoning: Finding potential explanations or connections via pathfinding between seemingly unrelated entities. \* Facilitating Complex Queries: Multi-hop queries to uncover indirect relationships, identifying supporting/contradicting evidence, pinpointing research gaps (nodes with specific missing connections). (Connects to SQ3b, SQ3c).

### **Integrating LLMs with Knowledge Graph Representations (Graph RAG and Beyond)**

Explicitly bridge the gap between the structured KG and the generative LLM component of the hybrid system.

- How KGs ground LLMs (reducing hallucination, providing context, structured memory) and how LLMs enhance KGs (natural language interface, summarizing graph query results, populating KG from text).

- **Graph-Augmented Generation (Graph RAG) Mechanisms:** \* Detailed explanation of how KG queries retrieve targeted, structured context for LLM prompts. Comparison to standard Vector RAG (precision vs. recall trade-offs, handling complex relationships). \* Techniques: Using KG entities/relationships to construct prompts, summarizing graph query results into natural language context for the LLM.

- **Using LLMs for KG Interaction:** \* Natural Language to Graph Query (Text-to-Cypher/SPARQL). \* Using LLMs to help populate, enrich, or validate the KG (briefly, as this leans towards implementation).

2.3.4 **Managing Knowledge Dynamics and Quality in the Information Architecture**

-**Knowledge Acquisition & Integration:** Strategies for populating the KG (automated extraction, LLM-based extraction, curation). Challenges in maintaining consistency and quality during ingestion. .

- **Knowledge Updating & Maintenance:** Strategies for incorporating new research findings, handling schema evolution, ensuring timeliness.

- **Representing Knowledge Quality & Provenance:** Modeling source reliability, evidence strength, certainty levels within the graph schema (e.g., properties on nodes/edges). \* Tracking data provenance (origin of information).

- **Handling Conflicting Information & Ambiguity:** \* Schema-level approaches (e.g., specific relationship types for contradiction).

- Query-level detection.

- Strategies for representing and reasoning over conflicting evidence (e.g., argumentation frameworks represented in the graph).

- **Searchability & Scalability:** Indexing strategies (keyword, vector, graph-native), query optimization techniques relevant to KG performance.

### **State-of-the-Art and Challenges in KG-Enhanced Agent Architectures**

- Review of recent relevant systems/frameworks combining KGs and LLMs for reasoning, planning, or knowledge-intensive tasks (cite specific examples like GraphRAG, relevant academic papers on KG+LLM agents).

Focus specifically on approaches relevant to multi-step processes, research, or synthesis.

- Identify key limitations and open challenges in the current state-of-the-art (e.g., robust KG construction from text, complex temporal/causal reasoning on KGs, seamless planning-reasoning-action loops, evaluation difficulties for complex tasks).

- Clearly articulate how your proposed architecture aims to address some of these challenges.

## **Focus on Use Case Preperation**

### **Autonomous creating a Literature Review**

* + - Defining the Scope and Research Questions: How can the agent help in formulating clear and focused research questions for the literature review?
    - Identifying Relevant Sources: Strategies for searching academic databases, evaluating source credibility, and filtering irrelevant results.
    - Information Extraction and Summarization: Techniques for extracting key information (e.g., methods, results, conclusions) from research papers and generating concise summaries.
    - Synthesizing Information and Identifying Gaps: How can the agent connect different ideas from multiple papers and identify areas where further research is needed? This is a key aspect of a good literature review.
    - Structuring and Writing the Literature Review: Outlining, paragraph organization, citation management, and generating coherent text.
    - Evaluation of good Literature
      * Criteria for Evaluating Research Papers: Methodology, results, discussion, impact, novelty, etc. How does the agent assess the quality of research papers?
      * Metrics for Assessing the Quality of a Literature Review: Completeness, coherence, critical analysis, etc. How will you evaluate the output of your agent?
      * Bias Detection in Literature: How can the agent help in identifying potential biases in research studies?
    - Specific Challenges of Automating Literature Reviews with AI:
      * Understanding Complex Research Methodologies: Can AI truly grasp the nuances of different research designs and their limitations?
      * Dealing with Paywalled Content: Strategies for accessing and processing information from behind paywalls (ethically and practically).
      * Handling Different Citation Styles: Ensuring consistency and accuracy in referencing.
    - Searching and comparing (feature based comparison /not quality assessment) different AI-Literature (Deep Research) Tools.

## Research Gap Analysis

Gemäss Andreas am Schluss ein Kapitel und ein Verweis zum Gap in der Introduction.

# Research Methodology

Always use the word “because” when you say you choose a design – why do you choose /Do this?

## Introduction

This chapter details the research methodology adopted to address this question. The research employs a combination of the Research Onion model (Saunders et al., 2023) and Design Science Research (DSR) methodology (Dresch et al., 2014; Hevner et al., 2004).

## Research Paradigm and Strategy

### Research Philosophy

This study adopts a **pragmatic** research philosophy, balancing theoretical understanding with practical application (Creswell & Creswell, 2018). This is appropriate because the research aims to both understand the capabilities of hybrid AI systems and to create a working artifact. Pragmatism allows for the integration of both qualitative and quantitative methods and for iterative refinement of the artifact based on empirical findings. The goal is to produce knowledge that is both theoretically sound and practically useful for automating complex tasks like literature reviews.

### Research Approach

The thesis follows an **inductive** approach, leveraging empirical data generated from the development and evaluation of the hybrid AI agent to refine and validate the proposed architecture. While elements of deductive reasoning are incorporated (e.g., testing the AI agent's performance against established literature review methodologies), the primary focus is on **developing** new frameworks and insights through iterative experimentation. Specifically, the inductive process involves: (1) observing the limitations of existing AI approaches for literature reviews; (2) designing and implementing a novel hybrid AI agent system; (3) evaluating the agent's performance; and (4) iteratively refining the system's architecture and information representation based on the evaluation results.

### Research Strategy: Design Science Research (DSR)

This study employs Design Science Research (DSR), a problem-solving paradigm focused on creating and evaluating innovative artifacts to address real-world problems (Dresch et al., 2014; Hevner et al., 2004). DSR is appropriate because the core of this research is the creation of a novel hybrid AI agent system for automating literature reviews. This is a practical problem with significant real-world implications, and DSR provides a structured framework for designing, building, and evaluating the artifact.

### DSR Cycles

The DSR process is structured around three interconnected cycles:

**Relevance Cycle:** The relevance cycle ensures the AI agent addresses the identified need for improved efficiency, accuracy, and comprehensiveness in literature reviews, a challenge highlighted by the limitations of existing LLM-only and standard RAG approaches (Silva, 2023; Lewis et al., 2021). This cycle connects the research to the practical application environment by ensuring the artifact is designed to solve a relevant problem.

**Rigor Cycle:** The rigor cycle grounds the artifact's design in established theories of knowledge graphs (Angles & Gutierrez, 2008), retrieval-augmented generation (Lewis et al., 2021), and AI planning (Ghallab et al., 2004), ensuring a theoretically sound foundation. This cycle ensures the research draws upon existing knowledge and contributes to the knowledge base.

**Design Cycle:** The design cycle involves iterative development, testing, and refinement of the AI agent, incorporating feedback from expert evaluations (Section 3.4) and performance metrics (Section 3.5). This iterative process allows for continuous improvement of the artifact based on empirical data. This is the core cycle of building and evaluating the artifact.

### Research Choice

A multi-method qualitative and quantitative approach is adopted, employing a convergent parallel design (Creswell & Plano Clark, 2018). This means that qualitative and quantitative data are collected and analyzed concurrently, and then the findings are integrated to provide a more complete understanding.

**Qualitative:** Expert evaluations of AI-generated literature reviews and (potentially) interviews with domain experts. Qualitative data will provide insights into the quality, coherence, and reasoning of the AI-generated reviews, capturing nuances that quantitative metrics might miss.

**Quantitative:** Accuracy metrics (precision, recall, F1-score), evaluation benchmarks (e.g., comparing against human-generated reviews, as documented by several studies (Peoples et al., 2023; Singh et al., 2017)), to assess the performance of the AI agent. Quantitative data will provide objective measures of the accuracy, completeness, and efficiency of the AI agent.

### Time Horizon

The research follows a **cross-sectional** time horizon, focusing on the available data within a specific timeframe (March 2025 - June 2025). However, the artifact development will incorporate an iterative approach, allowing for punctual updates based on new insights, as new relevant papers are currently released on short periodic intervals.

### Research Techniques and Procedures

The study employs the following data collection and evaluation techniques:

* **Systematic Literature Review** – Establishing the foundational understanding of AI planning, RAG-based retrieval, and knowledge graphs.
* **Expert Evaluations** – Assessing the quality of AI-generated literature reviews.
* **Performance Metrics** – Analyzing accuracy, coherence, and reliability of AI outputs.
* **Comparative Benchmarking** – Evaluating AI agent performance against results by students.

## Design Science Research (DSR)

DSR serves as the guiding framework for this study. It consists of three iterative and interrelated cycles:

### Design Science Research Cycles

* **Relevance Cycle:** Identifies the problem space and ensures that the AI agent and information architecture address real-world challenges in literature review automation.
* **Rigor Cycle:** Integrates theoretical foundations from knowledge graphs, retrieval-augmented generation (RAG), and AI planning.
* **Design Cycle:** Focuses on iterative build-and-evaluate activities to refine the artifact, incorporating expert feedback and empirical testing.

### Design Science Research Phases

1. **Awareness Phase:**
   * Identifying gaps in current AI planning methodologies for literature reviews.
   * Reviewing limitations of existing AI architectures (e.g., LLM-only, RAG, standard agent frameworks).
   * If the prof-of-concept of the Hybrid AI System is already deliver tested consistent results, the tool will be used for a automated broadening and deepening of the literature review.
2. **Suggestion Phase:**
   * Designing a hybrid AI agent that integrates **Graph-based RAG for structured planning**.
   * Defining the **information architecture** to support AI decision-making.
3. **Development Phase:**
   * Developing a hybrid agent system – either from scratch or with a pre-evaluated framework.
   * Implementing the graph database to enhance AI-driven retrieval.
   * Developing evaluation mechanisms to measure AI effectiveness.
4. **Evaluation Phase:**
   * Compare Outputs from Hybrid Agent and LLM-based Agents or literature generating tools.
   * Conducting expert reviews of AI-generated literature outputs.
   * Analyzing accuracy, coherence, and methodology adherence.
   * Comparing AI-generated and human-generated reviews.

## Evaluation Methodology

### High-Level Assessment (Expert Evaluation)

Experts (professors, researchers) will evaluate AI-generated literature reviews based on:

* **Coherence and logical flow**
* **Coverage and completeness**
* **Citation accuracy and factual correctness**
* **Compliance with research methodologies**

The evaluations will follow grading frameworks based on academic standards.

* **Black-Box Testing:** Evaluating Papers without knowing, which Result is AI generated, and which is created by a student
* **White-Box Testing:** Analyzing AI decision pathways, information retrieval sequences, and processing logic.

### Low-Level Assessment (Accuracy & Process Compliance)

This assessment ensures that the AI agent:

* Uses **verifiable citations** in literature synthesis.
* **Follows structured research methodologies** in generating content.
* Implements **AI-assisted quality-checking** mechanisms for factual validation.

### Grading Frameworks

Academic grading frameworks will be adapted to ensure objective evaluation, covering:

* Logical coherence
* Accuracy of references
* Depth and breadth of literature coverage

## Data Collection & Analysis

### Data Sources

* **Graph Database:** Structured knowledge base integrating research papers, citations, and topic hierarchies.
* **Academic Repositories:** Google Scholar, PubMed, ArXiv, etc.
* **Evaluation Data:** AI-generated literature reviews vs. human-generated baselines.

### Evaluation Metrics

* **Accuracy:** Percentage of correct references and citations.
* **Completeness:** Extent of literature coverage.
* **Coherence:** Logical structure and argumentation.
* **Efficiency:** Time reduction in literature review processes.

### Validation Strategy

AI-generated reviews will be cross-verified through:

* **Expert assessments**
* **Cited source verification**
* **Automated quality control checks**

## Research Timeline

The research will generally adhere to the planned timeline; however, it is important to consider that an iterative approach is adopted, with the following phases serving as focal points:

1. **March 2025 - June 2025:** Awareness & Suggestion Phase – Reviewing literature, defining architecture.
2. **July 2025 - October 2025:** Development Phase – Implementing the knowledge graph and AI agent.
3. **October 2025 - December 2025:** Evaluation Phase – Conducting expert assessments, analyzing AI-generated reviews.

**Further Project Planning:**

In beginning of September 2025, contacts must be established to professors, that might support the evaluation as an expert.

As a guidance deadline, 9. November 2025 is set to deliver the final version of the proof-of-concept system, so the test literature reviews can be created. The experts will be given 3 Weeks’ time for the evaluation.

## Literature Review

This section explains how the literature review was (or will be?) conducted. Two methods will systematically be conducted for finding the broadest spectrum of relevant papers.

**Method 1: Searching for Keywords in Databases and in semantic-based searches**

This is the most traditional and fundamental approach to finding relevant literature. Here's a more detailed explanation:

* **Databases:** The image doesn't specify which databases, but this is a critical detail. Common academic databases include:
  + **General/Multidisciplinary:**
    - Web of Science
    - Scopus
    - Google Scholar
  + **Subject-Specific:**

The choice of database(s) is crucial and should be justified in the full methodology. A researcher would likely use multiple databases to ensure comprehensive coverage.

* **Keywords:** These are the search terms the researcher uses. Developing effective keywords is a skill in itself. It involves:
  + **Identifying key concepts:** What are the core ideas of the research topic?
  + **Using synonyms and related terms:** "Artificial intelligence" might also require searching for "machine learning," "deep learning," etc. Researchers use Boolean operators (AND, OR, NOT) to combine terms.
  + **Using truncation and wildcards:** comput would find "computer," "computers," "computation," "computational," etc. wom?n would find "woman" and "women."
  + **Refining searches:** It's an iterative process. Initial searches often yield too many (or too few) results, requiring adjustments to the keywords and search strategy. Researchers might use filters (e.g., publication date, language, study type) to narrow down results.
* **Process:** The researcher would typically:
  + Define their research question clearly.
  + Identify key concepts and develop initial keywords.
  + Choose relevant databases.
  + Conduct initial searches.
  + Evaluate the results (titles, abstracts) for relevance.
  + Refine keywords and search strategies as needed.
  + Download and manage relevant articles.

**Key Words and Search Terms for Literature Review:**

**General Concepts**

* Hybrid AI Agent Systems
* Generative AI (GenAI)
* Large Language Models (LLMs)
* Agentic AI
* Multi-agent systems
* AI-driven research automation
* Cognitive architectures
* Autonomous AI planning
* AI planing
* AI decision-making
* Symbolic and Subsymbolic AI
* AI reasoning and explainability
* AI Access Rights management (RBAC)

**Knowledge Representation & Retrieval**

* Knowledge graphs (KG)
* Retrieval-Augmented Generation (RAG)
* Graph-based RAG (GraphRAG)
* Vector databases vs. graph databases
* Semantic search in AI
* Graph Search / Multi Hop Querying
* Sub Graph Creation
* Information architecture for AI planning
* Ontologies in AI
* Structured knowledge representation
* Reasoning on knowledge graphs / Graph DB
* AI-driven information synthesis
* Provenance and trust in knowledge representation
* Cleaning / remodeling a GraphDB

**AI Agent Architectures & Planning**

* AI agent planning and execution
* Multi-phase AI processes
* Explainable AI (XAI)
* AI workflow automation
* Planning under uncertainty
* Cognitive modeling in AI
* Knowledge-based AI agents
* Graph-structured AI planning
* Multi-agent coordination
* AI-driven task decomposition
* Tool Use / Tool Selection
* MCP Server / Services
* Cognitive Processes in AI

**Autonomous Literature Review & Research Automation**

* AI-assisted literature review
* Automated knowledge discovery
* AI-generated research synthesis
* AI-supported citation analysis
* Information retrieval for literature reviews
* AI-driven academic writing
* Research question formulation via AI
* AI-generated bibliometric analysis
* AI and research integrity
* Evaluating AI-generated literature reviews
* Literature review quality assessment
* Literature Evaluating Framework

**Benchmarking & Evaluation of AI Agents**

* AI performance benchmarks
* AI reliability metrics
* Evaluation of reasoning capabilities
* Human-AI comparative research quality
* AI research credibility assessment
* AI hallucination mitigation
* AI-driven research validation
* Bias detection in AI research
* AI citation accuracy metrics

**Ethical & Practical Considerations in AI Research**

* AI in academic integrity
* AI in research automation
* AI transparency in knowledge generation / reconstruction
* AI Research Gap
* Explainability and accountability in AI research

**Method 2: Connected Papers: Using Tools like "Lit Maps" and "Connected Papers"**

This method represents a more modern, visual, and network-based approach to literature discovery. It leverages the citation relationships between papers.

* **Connected Papers / Lit Maps:** These are web-based tools (Connected Papers is a specific tool, "Lit Maps" likely refers to a general category of literature mapping tools) that visualize the connections between research papers. They rely on the fact that papers cite each other.
* **How it works:**
  1. **Seed Paper:** You start with a single paper that you know is highly relevant to your topic. This is your "seed paper."
  2. **Citation Network:** The tool analyzes the papers that cite your seed paper (forward citations) and the papers that your seed paper cites (backward citations). This creates a network of related papers.
  3. **Visualization:** The tool presents this network visually, often as a graph. Papers that are closely connected (i.e., frequently cited together) are clustered together. This helps you quickly identify key papers and research areas.
  4. **Similarity Metrics:** These tools often use algorithms to calculate the "similarity" between papers based on their citation patterns. This helps identify papers that are topically related, even if they don't directly cite each other.
  5. **Prior and derivative works:** These tools are also able to show what are prior and what are the derivative works.
* **Benefits:**
  1. **Discovering hidden connections:** You can find relevant papers you might have missed using keyword searches alone.
  2. **Identifying influential papers:** Highly cited papers often appear prominently in the network.
  3. **Exploring different sub-areas:** Clusters in the graph can represent different sub-topics within your broader research area.
  4. **Faster exploration:** It can be a more efficient way to explore the literature than manually sifting through search results.
* **Difference from simple citation searching:** Tools do exist to search using citations. Method 2 represents the usage of visual tools.

## Further Details

**Participants (Expert Review):**

* **Recruitment:** How will you recruit experts (professors, PhD students, master students)? Specify your inclusion/exclusion criteria.
* **Sample Size:** How many experts will you involve? Justify your sample size (consider statistical power if you're aiming for quantitative comparisons).
* **Ethical Considerations:** Address any ethical considerations, such as informed consent and anonymity.

**Procedure (Expert Review):**

* **Task:** Describe the specific task experts will perform. Will they compare AI-generated reviews to human-generated reviews? Will they evaluate specific aspects of the reviews (e.g., coherence, completeness, accuracy)?
* **Rating Scale:** Describe the rating scale used by experts (e.g., Likert scale, qualitative feedback). Provide clear instructions and examples for the experts.
* **Blinding:** Will the experts be blinded to whether a review is AI-generated or human-generated (double-blind, single-blind, or not blinded)? Justify your choice.

**Grading Framework Evaluation:**

* **Framework Selection:** Clearly identify the grading framework(s) you will use. Justify your choice based on its relevance to literature reviews and its established validity/reliability. Cite the source of the framework.
* **Application:** Explain how you will apply the framework to both AI-generated and human-generated reviews. Will this be done manually or through an AI system? If using an AI, describe its implementation.
* **Metrics:** Specify the metrics you will derive from the grading framework (e.g., overall score, scores on specific dimensions).

**Accuracy Evaluation:**

* **Metrics:** Define your accuracy metrics. Examples include:
  + **Precision:** The proportion of retrieved information that is relevant.
  + **Recall:** The proportion of relevant information that is retrieved.
  + **F1-Score:** The harmonic mean of precision and recall.
  + **Factuality:** A measure of how well the generated text aligns with established facts (this might require manual verification or comparison to a trusted knowledge source).
  + **Citation Accuracy:** How accurately are sources cited?
  + **Claim Verification:** Are the claims made in the review supported by the cited evidence?
* **Procedure:** Describe how you will measure these metrics. Will you use automated tools, manual analysis, or a combination?
* **Ground Truth:** How will you establish a "ground truth" for comparison (e.g., a manually curated set of relevant papers and their relationships)?

# Agent Architecture Design

# Information Architecture Design

•**Overview**

•Present the overarching system design, showing how the cognitive architecture component and the LLM interact.

•**Cognitive Architecture Component**

•Describe how ACT-R modules handle topic selection by accessing declarative memory to identify research gaps

•Explain how procedural memory and chunking guide the literature review process, enabling the system to reason about relevance and credibility

•Show how production rules support argument construction, ensuring internal logical consistency and proper scientific methodology

•**LLM Component**

•Detail how the LLM uses the structured outline from the cognitive architecture to generate coherent and stylistically refined text for each section of the paper.

•Address methods of prompt engineering, fine-tuning, and iterative refinement to maintain fluency and factual accuracy.

•**Integration Mechanism**

•Explain the translation of symbolic representations into LLM-understandable prompts and the feedback loop where the LLM’s outputs are evaluated and refined by

1. **Introduction:**
   * **Purpose:** State the purpose of the Information Architecture (to support AI agent planning for complex processes, specifically literature review generation).
   * **Target Audience:** Who is this artifact for? (e.g., AI researchers, developers, information architects).
   * **Scope:** Clearly define the scope of the artifact. What is included, and what is not included?

## **Detail Description**

1. **Conceptual Model (Detailed):**
   * **Diagram:** Include a clear, well-labeled diagram of your conceptual model (entity-relationship diagram, UML class diagram, or similar). This is essential.
   * **Entity Definitions:** Provide detailed definitions of each entity (concept) in your model.
   * **Relationship Definitions:** Provide detailed definitions of each relationship type.
   * **Attribute Definitions:** Provide detailed definitions of all attributes.
   * **Examples:** Give concrete examples of each entity, relationship, and attribute.
2. **Graph Database Implementation (Detailed):**
   * **Database Choice:** Briefly reiterate your choice of graph database.
   * **Schema Diagram:** Include a diagram of your graph database schema (node types, relationship types, properties).
   * **Schema Definition (Code):** Provide the actual code used to create the schema in your chosen database (e.g., Cypher CREATE statements). This is crucial for reproducibility.
   * **Data Loading Procedures:** Describe in detail how data is loaded into the database. Include scripts or code snippets if applicable.
   * **Query Examples (More):** Provide a more extensive set of example queries that demonstrate the capabilities of the Information Architecture. These should go beyond the basic examples in the Method chapter. Show how complex queries can be constructed to support different aspects of the literature review process.
3. **Usage Guidelines:**
   * **How to Use:** Provide clear instructions on how to use the Information Architecture. This should include:
     + How to connect to the database.
     + How to execute queries.
     + How to interpret the results.
     + How to extend the schema (if applicable).
   * **Limitations:** Be honest about the limitations of your artifact. What are its current weaknesses? What are potential areas for future improvement?
4. **Example Application (Literature Review):**
   * **Walkthrough:** Provide a step-by-step walkthrough of how the Information Architecture can be used to support a specific part of the literature review process (e.g., identifying relevant papers, synthesizing findings, identifying research gaps).
   * **Illustrative Queries:** Show the specific queries that would be used at each step.
   * **Expected Results:** Show examples of the expected results from these queries.

**Key Considerations:**

* **Clarity and Precision:** Use clear, unambiguous language throughout. Avoid jargon.
* **Reproducibility:** Provide enough detail so that another researcher could replicate your work.
* **Justification:** Justify all your design choices. Explain why you made the decisions you did.
* **Diagrams:** Use diagrams liberally to illustrate your concepts and models.
* **Code:** Include relevant code snippets (e.g., Cypher queries, schema definitions).
* **Realism:** Use realistic examples and data whenever possible.
* **Ethical:** Address, that no personal data is used, that all data is handled correctly and in a secure way

## Implementation Details

•**Data Representation (Subq. 2 & 3):**

•Discuss using knowledge graphs, ontologies, or structured databases for representing literature and domain knowledge.

•Explain how chunks represent arguments, facts, and hypotheses.

•**Algorithms and Procedures (Subq. 1-4):**

•Provide pseudocode and flowcharts showing the iterative cycle of topic selection, literature review, argument skeleton formation, and final text generation.

•**Software and Tools (Subq. 1-4):**

•Mention frameworks (ACT-R environment, LLM APIs, Python libraries), and database solutions for knowledge storage (e.g., Neo4j for graph databases).

## Evaluation and Results

•**Evaluation Metrics (Subq. 5):**

•Define metrics for coherence, relevance, novelty, factual accuracy, and stylistic quality.

•Discuss both automated evaluation (e.g., perplexity, retrieval-augmented fact checks) and human expert assessments.

•**Experimental Setup (Subq. 5):**

•Describe the dataset of research papers and baselines used for comparison.

•Outline experimental parameters, such as sample size, domains tested, and the complexity of selected topics.

•**Results (Subq. 5):**

•Present quantitative metrics comparing the hybrid system to LLM-only baselines.

•Provide qualitative examples showing enhanced logical structure and context-awareness in hybrid-generated texts.

•**Discussion (Subq. 3 & 5):**

•Interpret results, linking them back to the research question and subquestions.

•Highlight improvements in logical coherence and fact consistency due to the cognitive architecture’s role.

# Prototype Design

# Result Evaluation

# Discussion, Limitations and Future Work

## Current Limitations:

•Acknowledge challenges in fully automating comprehensive knowledge acquisition and the complexity of modeling advanced scientific reasoning.

•Address difficulties in generating truly novel, creative hypotheses and the ethical considerations of automated research generation.

## Future Directions:

•Suggest integrating advanced reasoning models, reinforcement learning approaches for novelty, and improved ontologies or semantic web technologies for richer knowledge representation.

•Discuss developing mechanisms for explainability and bias mitigation in the hybrid approach.

## Bonus Discussion:

Should we reconsider how a high-quality research paper is constructed, given that AI-driven applied research provides access to more computational power and a greater volume of information? Instead of adhering strictly to a traditional literature review process, could the research process be transformed into a more experimental or testing-oriented approach? This shift would enable a structured validation of AI-generated findings, offering not only greater academic value but also a more rigorous means of evaluating AI-generated papers.

Furthermore, an additional enhancement could involve the automatic scoring of citations. By integrating predefined weighting criteria—such as the recency of a paper, the number of citations, the scale of data samples, and the rigor of its experimental design—it would be possible to assess individual citations with far greater precision than is typically achieved in conventional research. This approach could provide a more nuanced and systematic evaluation of sources, improving the overall reliability and relevance of AI-assisted research.

# Conclusion

# References

Anthropic. (2024, December 19). Building effective agents. *Building Effective Agents*. https://www.anthropic.com/research/building-effective-agents

AutoGPT. (2024, September 25). Everything You Need to Know About the AutoGPT Platform. *Everything You Need to Know About the AutoGPT Platform*. https://autogpt.net/everything-you-need-to-know-about-the-autogpt-platform/

AWS. (n.d.). *Was ist RAG? – Retrieval Augmented Generation erklärt*. Amazon Web Services, Inc. Retrieved 4 March 2025, from https://aws.amazon.com/de/what-is/retrieval-augmented-generation/

Dresch, A., Lacerda, D., & Antunes Júnior, J. A. V. (2014). Design Science Research: A Method for Science and Technology Advancement. In *Design Science Research: A Method for Science and Technology Advancement*. https://doi.org/10.1007/978-3-319-07374-3

Godsey, B. (2024, September 18). How Knowledge Graph RAG Boosts LLM Results. *DataStax*. https://www.datastax.com/blog/how-knowledge-graph-rag-boosts-llm-results

Hevner, A., R, A., March, S., T, S., Park, Park, J., Ram, & Sudha. (2004). Design Science in Information Systems Research. *Management Information Systems Quarterly*, *28*, 75.

Krishnan, N. (2025). *AI Agents: Evolution, Architecture, and Real-World Applications* (arXiv:2503.12687; Version 1). arXiv. https://doi.org/10.48550/arXiv.2503.12687

Larson, J. (2024, February 13). GraphRAG: A new approach for discovery using complex information. *Microsoft Research Blog*. https://www.microsoft.com/en-us/research/blog/graphrag-unlocking-llm-discovery-on-narrative-private-data/

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* (arXiv:2005.11401). arXiv. https://doi.org/10.48550/arXiv.2005.11401

Lin, S., Hilton, J., & Evans, O. (2022). *TruthfulQA: Measuring How Models Mimic Human Falsehoods* (arXiv:2109.07958). arXiv. https://doi.org/10.48550/arXiv.2109.07958

Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., Zhang, T., Liu, Y., Wang, H., Zheng, Y., & Liu, Y. (2024). *Prompt Injection attack against LLM-integrated Applications* (arXiv:2306.05499). arXiv. https://doi.org/10.48550/arXiv.2306.05499

Matsumoto, N., Choi, H., Moran, J., Hernandez, M. E., Venkatesan, M., Li, X., Chang, J.-H., Wang, P., & Moore, J. H. (2025). ESCARGOT: An AI agent leveraging large language models, dynamic graph of thoughts, and biomedical knowledge graphs for enhanced reasoning. *Bioinformatics (Oxford, England)*, *41*(2), btaf031. https://doi.org/10.1093/bioinformatics/btaf031

OpenAI. (2022, November 30). *Introducing ChatGPT*. Introducing ChatGPT. https://openai.com/index/chatgpt/

Peoples, N., Østbye, T., & Yan, L. L. (2023). Burden of proof: Combating inaccurate citation in biomedical literature. *BMJ*, *383*, e076441. https://doi.org/10.1136/bmj-2023-076441

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., & Zhuang, Y. (2023, November 2). *HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face*. Thirty-seventh Conference on Neural Information Processing Systems. https://openreview.net/forum?id=yHdTscY6Ci

Silva, M. (2023, October 18). Retrieval augmented generation: Keeping LLMs relevant and current - Stack Overflow. *Retrieval Augmented Generation: Keeping LLMs Relevant and Current*. https://stackoverflow.blog/2023/10/18/retrieval-augmented-generation-keeping-llms-relevant-and-current/

Singh, M., Niranjan, A., Gupta, D., Bakshi, N. A., Mukherjee, A., & Goyal, P. (2017). *Citation sentence reuse behavior of scientists: A case study on massive bibliographic text dataset of computer science* (arXiv:1705.02499). arXiv. https://doi.org/10.48550/arXiv.1705.02499

Xu, Z., Cruz, M. J., Guevara, M., Wang, T., Deshpande, M., Wang, X., & Li, Z. (2024). Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2905–2909. https://doi.org/10.1145/3626772.3661370

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). *ReAct: Synergizing Reasoning and Acting in Language Models* (arXiv:2210.03629). arXiv. https://doi.org/10.48550/arXiv.2210.03629

# List of digital aids

The following list of tools have been used to support individual areas of the scientific work.

|  |  |  |
| --- | --- | --- |
| **Aid** | **Usage** | **Affected Areas** |
| Zotero | Preparation of in-text citations and creation of bibliography | Entire work |
| ChatGPT | Summarizing tasks and rephrasing of various sentences | Entire work |
| Gemini | Idea Generation and brainstorming | Entire work |
| Cline / Anthropic | Code Generation for Prof-of-concept system | Code |
| Maybe: My Own PoC Artifact | Literature Review Generation | Literature Review |
| Sci-Space | Literature searching | Literature Review |
| Connected Paper | Literature searching | Literature Review |
| LitMaps | Literature searching | Literature Review |
|  |  |  |

# Appendix (If Necessary)