Multi-Modal and Multi-Agent Systems Meet Rationality: A Survey

Bowen Jiang[1, 2], Yangxinyu Xie[1, 2], Xiaomeng Wang[1], Weijie J. Su[1]
Camillo J. Taylor[1], Tanwi Mallick[2]
University of Pennsylvania[1] Argonne National Laboratory[2]
Philadelphia, PA, 19104, USA Lemont, IL, 60439, USA
{bwjiang@seas, xinyux@wharton, cjtaylor@seas}.upenn.edu, tmallick@anl.gov

Abstract

Rationality is the quality of being guided by reason, characterized by logical thinking and decision-making that align with evidence and logical rules. This quality is essential for effective problem-solving, as it ensures that solutions are well-founded and systematically derived. Despite the advancements of large language models (LLMs) in generating human-like text with remarkable accuracy, they present biases inherited from the training data, inconsistency across different contexts, and difficulty understanding complex scenarios involving multiple layers of context. Therefore, recent research attempts to leverage the strength of multiple agents working collaboratively with various types of data and tools for enhanced consistency and reliability. To that end, this paper aims to understand whether multi-modal and multi-agent systems are advancing toward rationality by surveying the state-of-the-art works, identifying advancements over single-agent and single-modal systems in terms of rationality, and discussing open problems and future directions. We maintain an open repository at https://github.com/bowen-upenn/MMMA_Rationality.

1 Introduction

Large language models (LLMs) have demonstrated promising results across a broad spectrum of tasks, particularly in exhibiting capabilities that plausibly mimic human-like reasoning (Wei et al., 2022; Yao et al., 2024; Besta et al., 2024; Shinn et al., 2024; Bubeck et al., 2023; Valmeekam et al., 2023; Prasad et al., 2023). These models leverage the richness of human language to abstract concepts, elaborate thinking process, comprehend complex user queries, and develop plans and solutions in decision-making scenarios. Despite these advances, recent research has revealed that even state-of-the-art LLMs exhibit various forms of irrational behaviors, such as the framing effect, certainty effect, overweighting bias, and conjunction fallacy (Binz and Schulz, 2023; Echterhoff et al., 2024; Mukherjee and Chang, 2024; Macmillan-Scott and Musolesi, 2024; Wang et al., 2024a; Suri et al., 2024). These biases significantly challenge the utility of LLMs in natural language processing research. For example, LLM-based evaluators, a popular choice for automated assessments for text generation, display cognitive biases against certain responses irrespective of their actual quality or relevance (Stureborg et al., 2024; Koo et al., 2023). Irrationality and hallucinations (Bang et al., 2023; Guerreiro et al., 2023; Huang et al., 2023) also undermine the practical deployment of LLMs in critical sectors like healthcare, finance, and legal services (He et al., 2023; Li et al., 2023i; Kang and Liu, 2023; Cheong et al., 2024), where reliability and consistency are paramount. The emerging concern about the factual accuracy and trustworthiness of LLMs highlighting an urgent need to develop better agents or agent systems (Nakajima, 2023; Gravitas, 2023) with rational reasoning processes.

One possible reason for the LLMs' irrational behaviors, as suggested by Bubeck et al. (2023) and Sun (2024), is the *autoregressive* nature of existing language models. This architecture doesn't allow for an "internal scratchpad" beyond these models' inner parametric representations of knowledge, causing them to fail to reason rationally when faced with problems that require more complex and iterative procedures. Thus, an important question emerges: How can we design an LLM-based agent capable of rational decision-making that can overcome these biases and inconsistencies?

Recent advancements in multi-modal and multi-agent frameworks offer a promising direction to address this challenge, which leverage the expertise of different agents acting together towards a collective goal. Multi-modal foundation models (Awadalla et al., 2023; Liu et al., 2023b; Wang et al., 2023d; OpenAI, 2023; Reid et al., 2024) enhance reasoning by grounding decisions in a broader sensory context, akin to how human brains integrate rich sensory inputs to form a more holistic base of knowledge. Meanwhile, multi-agent systems introduce mechanisms such as consensus, debate, and self-consistency (Du et al., 2023; Liang et al., 2023; Talebirad and Nadiri, 2023; Madaan et al., 2024; Cohen et al., 2023; Shinn et al., 2024; Mohtashami et al., 2023), which allow for more refined and reliable output through collaborative interaction among multiple instances. Each agent is specialized in different domains and offers its unique perspectives, simulating the dynamics of discussion in human societies. Multi-agent systems can also incorporate multi-modal agents and agents specialized in querying external knowledge sources or tools (Lewis et al., 2020; Schick et al., 2024; Tang et al., 2023; Pan et al., 2024) to overcome hallucinations, ensuring that their results are more robust, deterministic, and trustworthy, thus significantly improving the quality of the generated responses towards rationality.

This survey provides a unique lens to interpret the underlying motivations behind current multimodal and/or multi-agent systems. Drawing from cognitive science, we first delineate four fundamental requirements for rational thinking in Section 2, and then Section 4 discusses how research fields within the multi-modality and multi-agents literature are progressing towards rationality by inherently improving these criteria. We posit that such advancements are bridging the gap between the performance of these systems and the expectations for a rational thinker, in contrast to traditional single-agent language-only models. Furthermore, Section 5 highlights the lack of sufficient evaluation metrics and benchmarks in the existing literature to adequately measure the rationality of LLMs or agent systems. We hope this survey can inspire further research at the intersection between agent systems and cognitive science.

2 Defining Rationality

A rational agent should avoid reaching contradictory conclusions in decision making processes, respecting the physical and factual reality of the world in which it operates. Thus, drawing on foundational works in rational decision-making (Tversky and Kahneman, 1988; Hastie and Dawes, 2009; Eisenführ et al., 2010), this section adopts an axiomatic approach to define rationality, presenting four substantive axioms that we expect a rational agent or agent systems to fulfill:

Grounding The decision of a rational agent is grounded on the physical and factual reality. For example, a flight booking agent must accurately retrieve available airports without fabricating nonexistent ones, and a video generation agent should adhere to the laws of physics in a world model.

Naturally, in order to make a sound decision, the agent must be able to integrate sufficient and accurate information from different sources and modalities grounded in reality without hallucination. While this requirement is generally not explicitly stated in the cognitive science literature when defining rationality, it is implicitly implied, as most humans have access to physical reality through multiple sensory signals.

Orderability of Preferences When comparing alternatives in a decision scenario, a rational agent can rank the options based on the current state and ultimately select the most preferred one based on the expected outcomes. This orderability consists of several key principles, including comparability, transitivity closure, solvability, etc. with detailed defined in Appexdix A. The orderability of preferences ensures the agent can make consistent and logical choices when faced with multiple alternatives. LLM-based evaluations heavily rely on this property, as discussed in Appendix B.

Independence from irrelevant context The agent's preference should not be influenced by information irrelevant to the decision problem at hand. LLMs have been shown to exhibit irrational behavior when presented with irrelevant context (Shi et al., 2023; Wu et al., 2024; Liu et al., 2024d; Yoran et al., 2023), leading to confusion and suboptimal decisions. To ensure rationality, an agent must be able to identify and disregard irrelevant information, focusing solely on the factors that directly impact the decision-making processes.

Invariance The preference of a rational agent remains invariant across equivalent representations of the decision problem, regardless of specific wordings or modalities.

3 Scope

Unlike existing surveys (Han et al., 2024; Guo et al., 2024; Xie et al., 2024a; Durante et al., 2024; Cui et al., 2024; Xu et al., 2024b; Zhang et al., 2024a;

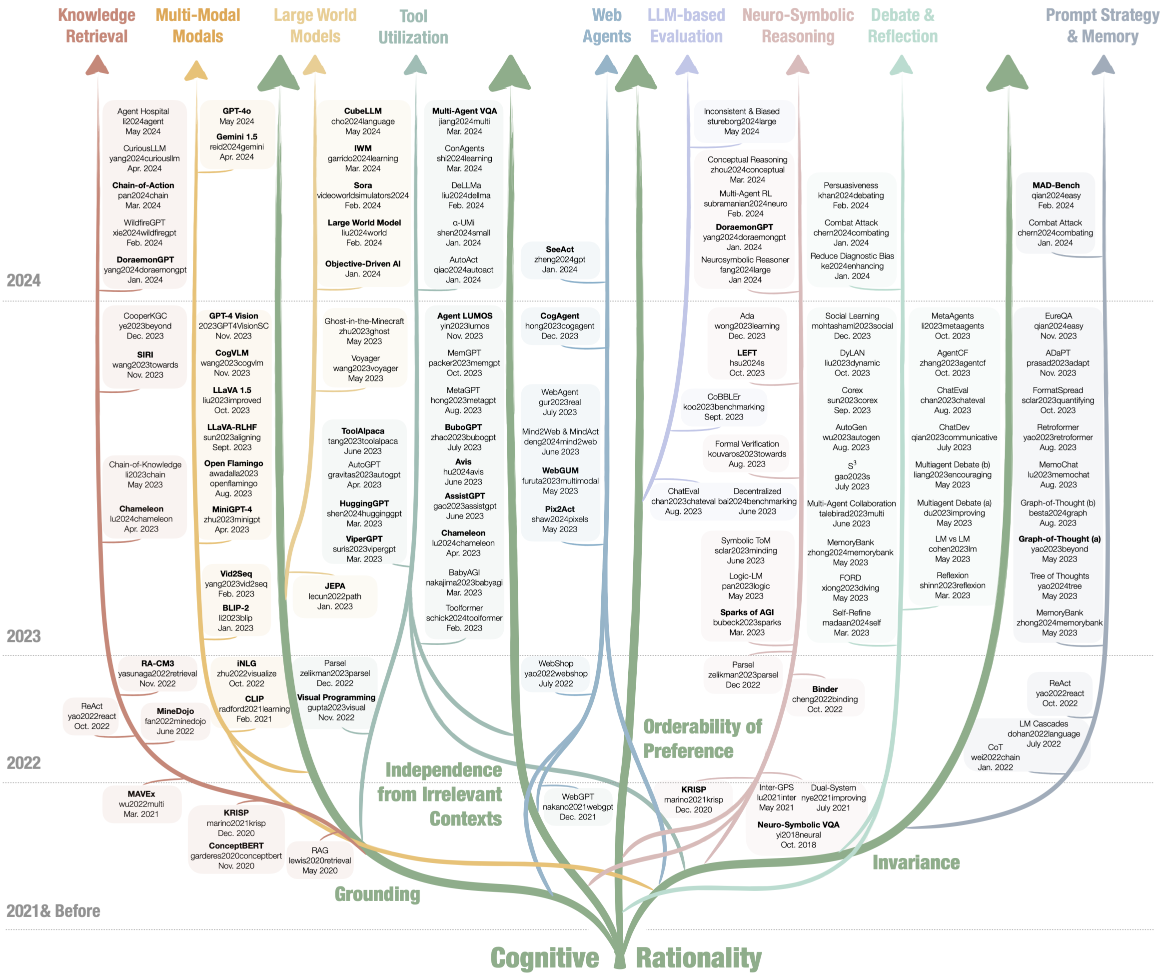

Figure 1: The evolutionary tree of multi-agent and/or multi-modal systems related to the four axioms of rationality. Many proposed approaches strive to address multiple axioms simultaneously. **Bold** fonts are used to mark works that involve multi-modalities. This tree also includes foundational works to provide a clearer reference of time.

Cheng et al., 2024; Li et al., 2024a) that focus on the components, structures, agent profiling, planning, communications, memories, and applications of multi-modal and/or multi-agent systems, **this survey is the first to specifically examine the increasingly important relations between rationality and these multi-modal and multi-agent systems**, exploring how they contribute to enhancing the rationality in decision-making processes.

We emphasize that *rationality*, by definition, is not equivalent to *reasoning* (Khardon and Roth, 1997; Huang and Chang, 2022; Zhang et al., 2024a; Qiao et al., 2022), although deeply intertwined. Rationality involves making logically consistent decisions grounded with reality, while reasoning refers to the cognitive process of drawing logical inferences and conclusions from available information, as illustrated in the following thought experiment:

> *Consider an environment where the input space and the output decision space are finite. A lookup table with consistent mapping from input to output is inherently rational, while no reasoning is necessarily present in the mapping.*

Despite this example, it is still crucial to acknowledge that reasoning typically plays a vital role in ensuring rationality, especially in complex and dynamic real-world scenarios where a simple lookup table is insufficient. Agents must possess the ability to reason through novel situations, adapt to changing circumstances, make plans, and achieve rational decisions based on incomplete or uncertain information.

Rationality is also different from Theory of Mind (ToM) (Apperly and Butterfill, 2009; Nye et al., 2021; Oguntola et al., 2021; Hagendorff, 2023; Li et al., 2023c; Sclar et al., 2023b; Kosinski, 2023) in machine psychology. ToM refers to the model's ability to understand that others' mental states, beliefs, desires, emotions, and intentions may be different from its own.

4 Towards Rationality through Multi-Modal and Multi-Agent Systems

This section surveys recent advancements in multi-modal and multi-agent systems, categorized by their fields as depicted in Figure 1.

Each category of research in Figure 1, such as knowledge retrieval or neuro-symbolic reasoning, addresses one or more fundamental requirements for rational thinking. These rationality requirements are typically *intertwined*; therefore, an approach that enhances one aspect of rationality often inherently improves others simultaneously. Meanwhile, the overall goal of current multi-agent system in achieving rationality can usually be distilled into two key concepts: *deliberation* and *abstraction*. Deliberation encourages slower reasoning process such as brainstorming and reflection, while abstraction refers to boiling down the problem its into logical essence like calling APIs of tools or incorporating neuro-symbolic reasoning agents.

Most existing studies do not explicitly base their frameworks on rationality in their original writings. Our analysis aims to reinterpret these works through the lens of our four axioms of rationality, offering a novel perspective that bridges existing methodologies with rational principles.

Besides, *dual-process theories* (Sun, 2001; Evans, 2003; Kahneman, 2011; Sun, 2024; Nye et al., 2021) in human cognition suggest that a single LLM, in analogy, might engage primarily in *"System 1"* thinking that is fast and automatic, but prone to biases. In contrast, *"System 2"* thinking is slow and rule-based (Kahneman, 2011), which is more reliable but requires more cognitive effort. This section will also discuss how multi-modal and multi-agent systems align with these theories.

4.1 Towards Grounding & Invariance through Multi-Modal Models

Multi-modal approaches aim to improve the information grounding across various channels, such as language and vision. By incorporating multi-modal agents, multi-agent systems can greatly expand their capabilities, enabling a richer, more accurate, and contextually aware interpretation of environment.

Multi-Modal Foundation Models Grounding an agent solely based on textual language can be challenging, as information can be represented much more efficiently through other sensory modes. For example, as a picture is worth a thousand words, recent advances in large vision-language pretraining have enabled LLMs with robust language comprehension capabilities to finally perceive the visual world. Multi-modal foundation models, including but not limited to CLIP (Radford et al., 2021), VLBERT and ViLBERT (Su et al., 2019; Lu et al., 2019), BLIP-2 (Li et al., 2023e), (Open) Flamingo (Alayrac et al., 2022; Awadalla et al., 2023), LLaVA (Liu et al., 2024c, 2023b), CogVLM (Wang et al., 2023d), MiniGPT-4 (Zhu et al., 2023a), GPT-4 Vision (OpenAI, 2023) and GPT-4o (OpenAI, 2024), and Gemini 1.5 Pro (Reid et al., 2024) serve as the cornerstones for multi-modal agent systems to ground knowledge in vision and beyond.

Alignment with Dual-Process Theories Generating output texts from input images in these models requires only a single inference pass, which is quick and straightforward. This process bypasses the need for iterative reasoning or reflective steps, aligning single-model inference closely with the System 1 process of fast and automatic thinking.

The adaptation of Reinforcement Learning from Human Feedback (RLHF) (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022), a technique popularized in language-only models, in LLaVA-RLHF (Sun et al., 2023b) demonstrates promising advancements in reducing hallucination from cross-modal misalignment. Furthermore, visual instruction-tuning (Liu et al., 2024c; Dai et al., 2024; Bai et al., 2023; Wang et al., 2023c) enables advanced foundation models like LLaVA (Liu et al., 2024c, 2023b), GPT-4 Vision (OpenAI, 2023), and Gemini 1.5 Pro (Reid et al., 2024) to engage in more detailed instruction following, multi-round human-agent interactions, and collaborations with other agents, performing deliberate multimodal reasoning, tool using (Schick et al., 2024), and strategic thinking to understand more engineered or context-dependent queries. This opens the possibility of subsequent research on the System 2 process in multi-modal foundation models.

Information Grounding Multi-modalities help enhance the functionality of agent systems through more diverse information grounding. For example, Chain-of-Action (Pan et al., 2024) advances the single-modal Search-in-the-Chain (Xu et al., 2023) by supporting multi-modal data retrieval for faithful question answering. DoraemonGPT (Yang et al., 2024) decomposes complex tasks into simpler ones toward understanding dynamic scenes, where multi-modal understanding is necessary for spatial-temporal videos analysis. RA-CM3 (Yasunaga et al., 2022) augments baseline retrieval-augmented LLMs with raw multi-modal documents that include both images and texts, assuming that these two modalities can contextualize each other and make the documents more informative, leading to better generator performance. The multi-modal capabilities also allow HuggingGPT (Shen et al., 2024b), Agent LUMOS (Yin et al., 2023), ToolAlpaca (Tang et al., 2023), and AssistGPT (Gao et al., 2023b) to expand the scope of tasks they can address, including cooperation among specialized agents or tools capable of handling different information modalities.

Web agents are another example of how multi-modal agents surpass language-only ones. In agents like Pix2Act (Shaw et al., 2024), WebGUM (Furuta et al., 2023), CogAgent(Hong et al., 2023b), and SeeAct (Zheng et al., 2024a), web navigation is grounded on graphical user interface (GUI) rather than solely on HTML texts (Shen et al., 2024a; Yao et al., 2022a; Deng et al., 2024; Gur et al., 2023). This method of visual grounding offers higher information density compared to HTML codes that are usually lengthy, noisy, and sometimes even incomplete (Zheng et al., 2024a). Supporting the importance of vision, ablation studies in WebGUM (Furuta et al., 2023) also reports 5.5% success rate improvement on the MiniWoB++ dataset (Shi et al., 2017; Liu et al., 2018) by simply adding the image modality.

Large world models is an emerging and promising direction to reduce multimodal hallucinations. The notion is also mentioned in “Objective-driven AI" (LeCun, 2024), where agents have behavior driven by fulfilling objectives, i.e., drives, and they understand how the world works with common sense knowledge, beyond an auto-regressive generation. LeCun (2024) proposes the urgency for agents to learn to reason beyond feed-forward, i.e., the System 1 subconscious computation, and start making System 2 reasoning and planning on complicated actions to satisfy objectives with a grounding on world models. For example, Ghost-in-the-Minecraft (Zhu et al., 2023b) and Voyager (Wang et al., 2023a) have agents living in a well-defined game-world environment. JEPA (LeCun, 2022) creates a recurrent world model in an abstract representation space. Large World Model (LWM) (Liu et al., 2024b) and Sora (Brooks et al., 2024) develop insights from both textual knowledge and the world through video sequences. They both advance toward general-purpose simulators of the world, but still lack reliable physical engines for guaranteed grounding in real-world dynamics.

Invariance Across Modalities Achieving representation invariance across modalities is another critical facet of rationality. Given adequate information grounding, agents should make consistent decisions across different modalities that share equivalent underlying logic.

Multi-modal foundation models are particularly adept at promoting invariance by processing multi-modal data in an unified representation. Specifically, their large-scale cross-modal pretraining stage seamlessly tokenizes both vision and language inputs into a joint hidden embedding space, learning cross-modal correlations through a data-driven approach. In other words, image tokens are simply regarded as a foreign language (Wang et al., 2022a). Moreover, the cross-modal validation inherent in multi-modal foundation models allows for reconciliation of data from different modalities, closing their distance in the hidden embedding space (Radford et al., 2021).

The concept of invariance is the cornerstone of Visual Question Answering (VQA) agents (Chen et al., 2022; Jiang et al., 2024; Wang et al., 2023e; Yi et al., 2018; Wang et al., 2022a; Bao et al., 2022; Zhao and Xu, 2023). On one hand, these agents must grasp the invariant semantics of any open-ended questions posed about images, maintaining consistency despite variations in wording, syntax, or language. On the other hand, within a multi-agent VQA system, visual agents can provide crucial verification and support for language-based reasoning (Wang et al., 2023e; Jiang et al., 2024; Zhao and Xu, 2023), while language queries can direct the attention of visual agents, based on a shared and invariant underlying knowledge across vision and language domains.

4.2 Towards Grounding through Knowledge Retrieval

Bounded Rationality (March and Simon, 1958; Selten, 1990) is a concept tailored to cognitively limited agents, suggesting that decision-making is limited by the resources available at hand, and any deviations from the optimal are primarily due to insufficient computational capacity and bounded working memory. In terms of LLMs, the parametric nature of their existing architecture (Vaswani et al., 2017) fundamentally limits how much information they can hold. As a result, in the face of uncertainty, LLMs often hallucinate (Bang et al., 2023; Guerreiro et al., 2023; Huang et al., 2023), generating outputs that are not supported by the factual reality of the environment. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) marks a significant milestone in addressing such an inherent limitation of LLMs. Broadly speaking, RAG refers to any mechanism that provides external knowledge to the input context of an LLM and helps it deliver responses with up-to-date, factual, and grounded information.

A multi-modal and/or multi-agent system can include planning agents in its framework, which is akin to the System 2 process that can determine how and where to retrieve external knowledge, and what specific information to acquire. Additionally, the system can have summarizing agents that utilize retrieved knowledge to enrich the system's language outputs with better factuality.

There are multiple works that construct large-scale knowledge graphs (KGs) (Hogan et al., 2021) from real-world sources to effectively expand their working memory. Specifically, compared to language-only models, MAVEx (Wu et al., 2022) improves system's scores by 9.5% compared to an image-only baseline through the integration of knowledge from ConceptNet (Speer et al., 2017) and Wikipedia (Wikipedia contributors, 2004). It also improves the scores by 8.3% by using the image modality for cross-modal validations with an oracle. Thanks to the external knowledge base, ReAct (Yao et al., 2022b) reduces false positive rates from hallucination by 8.0% compared to CoT (Wei et al., 2022). CuriousLLM (Yang and Zhu, 2024) presents ablation studies showing the effectiveness of KGs on improving reasoning within the search process. MineDojo (Fan et al., 2022) observes that internet-scale multi-modal knowledge allows models to significantly outperform all creative task baselines. Equipped with world knowledge, RA-CM3 (Yasunaga et al., 2022) can finally generate faithful images from captions compared to CM3 (Aghajanyan et al., 2022) and Stable Diffusion (Rombach et al., 2022). CooperKGC (Ye et al., 2023) enables multi-agent collaborations, leveraging knowledge bases of different experts. It finds that the incorporation of KGs improves F1 scores by 10.0-33.6% across different backgrounds, and adding more collaboration rounds also enhance performance by about 10.0-30.0%. DoraemonGPT (Yang et al., 2024) supports knowledge tools to assist the understanding of specialized video contents. SIRI (Wang et al., 2023e) builds a multi-view knowledge base to increase the explainability of visual question answering. Grounding agents in external knowledge base also promotes more factual rationales and fewer hallucinations, especially in scientific and medical domains, exemplified by Chameleon (Lu et al., 2024), Chain-of-Knowledge (Li et al., 2023g), WildfireGPT (Xie et al., 2024b), and Agent Hospital (Li et al., 2024b). Chain-of-Knowledge (Li et al., 2023g) even discovers that integrating multiple knowledge sources enhances performance by 2.1% compared to using a single source in its experiments.

4.3 Towards Grounding & Invariance & Independence from Irrelevant Contexts through Tool Utilization

Information Grounding Enabling agents to use tools also expands their bounded working memories, akin to retrieving external knowledge. Toolformer (Schick et al., 2024) opens a new era that allows LLMs to use external tools via API calls following predefined syntax, effectively extending their capabilities beyond their intrinsic limitations and enforcing consistent and predictable outputs.

A multi-agent system can coordinate agents understanding when and which tool to use, which modality of information the tool should expect, how to call the corresponding API, and how to incorporate outputs from the API calls, which anchors subsequent reasoning processes with more accurate information beyond their parametric memory. For example, VisProg (Gupta and Kembhavi, 2023), ViperGPT (Surís et al., 2023), and Parsel (Zelikman et al., 2023) generate Python programs to reliably execute subroutines. Gupta and Kembhavi (2023); Surís et al. (2023) also invoke off-the-shelf models for multimodal assistance. Foundation models are not specifically trained for object detection or segmentation, so BuboGPT (Zhao et al., 2023) and Multi-Agent VQA (Jiang et al., 2024) call SAM (Kirillov et al., 2023; Ren et al., 2024) as the tool. Jiang et al. (2024) finds 8.8% of accuracy improvements compared to a single agent. Besides, BabyAGI (Nakajima, 2023), Chamelon (Lu et al., 2024), AssistGPT (Gao et al., 2023b), Avis (Hu et al., 2024), ToolAlpaca (Tang et al., 2023), MetaGPT (Hong et al., 2023a), Agent LUMOS (Yin et al., 2023), AutoAct (Qiao et al., 2024), α-UMi (Shen et al., 2024a), and ConAgents (Shi et al., 2024) harness compositional reasoning to enable generalized multi-agent systems with planning and modular tool-using capabilities in real-world scenarios.

Abstraction that Boils Down to Logical Essence In most cases, tools require translating natural language queries into API calls with predefined syntax. Once the planning agent has determined the APIs and their input arguments, the original queries that may contain irrelevant contexts become invisible to the tools, and the tools will ignore any variance in the original queries as long as they share the equivalent underlying logic. Take Multi-Agent VQA (Jiang et al., 2024) as an example. It has an LLM which provides only relevant object names rather than the whole visual question to the Grounded SAM (Ren et al., 2024) component of the system acting as an object-detector. Similarly, the image editing tools in VisProg (Gupta and Kembhavi, 2023) only receive a fixed set of arguments translated from user queries to perform deterministic code executions. SeeAct (Zheng et al., 2024a) as a Web agent explores vision-language models, ranking models, and a bounding box annotation tool to improve Web elements grounding from lengthy and noisy HTML codes. Consequently, using tools in a multi-agent system enhances the invariance and independence from irrelevant contexts, ensuring that their operations are streamlined and focused solely on necessary information.

4.4 Towards Orderability of Preferences & Invariance & Independence from Irrelevant Context through Neuro-Symbolic Reasoning

Alignment with Dual-Process Theories Neural-symbolic reasoning is another promising approach to achieving consistent ordering of preferences and invariance by combining the strengths of languages and symbolic logic in a multi-agent system. Nye et al. (2021) bridge neural-symbolic reasoning with the dual-process theories. It posits that LLMs are predominantly single systems operating in a mode akin to System 1. To address this, it introduces a multi-agent system that includes System 2-like agents to break down complex tasks into symbolic programs, filter candidate responses generated the previous LLM, and compare them with a minimal world model to check consistency. Such a slower interaction between LLMs and neuro-symbolic modules enhances the coherence and explainablity of the overall reasoning process.

Coherent Orderability of Preference A multi-agent system incorporating symbolic modules can not only understand language queries but also solve them with a level of consistency, providing a faithful and transparent reasoning process based on well-defined rules that adhere to logical principles, which is unachievable by LLMs alone. Logic-LM (Pan et al., 2023) combines problem formulating, symbolic reasoning, and result interpreting agents, where the symbolic reasoner empowers LLMs with deterministic symbolic solvers to perform inference, ensuring a correct answer is consistently chosen. Its multi-agent framework also encourages self-refinement that modifies logical formulation errors using error messages from the symbolic reasoner as the feedback. Besides, SymbolicToM (Sclar et al., 2023b) and KRISP (Marino et al., 2021) construct explicit symbolic graphs and answer questions by retrieving nodes in the graph. Binder (Cheng et al., 2022), Parsel (Zelikman et al., 2023), LEFT (Hsu et al., 2024), and Fang et al. (2024) decompose tasks into planning, parsing, and execution, where the symbolic reasoning agents can help maintain a coherent order of preferences among symbolic options in the system outputs. By skipping the symbolic module, Parsel (Zelikman et al., 2023) observes a substantial performance drop of 19.5%. LEFT (Hsu et al., 2024) also outperforms end-to-end baselines without symbolic programs by 3.85% on average across multiple experiments. In more explicit scenarios, logical modules can directly compare the order of options represented as variables—such as "left" or "right" in relational logic (Hsu et al., 2024)—rather than relying on a single LLM to generate responses indeterministically within the natural language space.

Abstraction that Boils Down to Logical Essence Neuro-Symbolic modules typically expect standardized input formats (Zelikman et al., 2023; Pan et al., 2023; Sclar et al., 2023b; Hsu et al., 2024; Fang et al., 2024; Yang et al., 2024; Subramanian et al., 2024), analogous to making API calls with external tools. This layer of abstraction enhances the independence from irrelevant contexts and maintains the invariance of LLMs when handling natural language queries. The only relevant factor is the parsed inputs into the predetermined neuro-symbolic programs. For instance, Ada (Wong et al., 2023) introduces symbolic operators to abstract actions, ensuring that lower-level planning models are not compromised by irrelevant information in the queries and observations. Without the symbolic action library, a single LLM would frequently fail at grounding objects or obeying environmental conditions, resulting in a significant accuracy gap of approximately 59.0-89.0%.

4.5 Towards Orderability of Preferences & Invariance through Reflection, Debate, and Memory

Alignment with Dual-Process Theories Due to the probabilistic outputs of LLMs, which resemble the rapid, non-iterative nature of human System 1 cognition, ensuring preference orderability and invariance is challenging. In contrast, algorithms that enable self-reflection and multi-agent systems that promote debate and consensus can help align outputs more closely with the deliberate and logical decision-making typical of System 2 processes, thus enhancing rational reasoning in agents.

Memory is one of the most fundamental cognitive processes that lead to reasoning, creativity, learning, and even self-consciousness in humans (Solso and Kagan, 1979; Craik and Lockhart, 1972; Leydesdorff and Hodgkin, 2017; Johnson-Laird, 1983; Laird, 2019; Sun, 2001). In other words, even an artificial general intelligence (AGI) that lacks memory and forgets every past conversation after each query would find it hard to develop a coherent and rational decision-making.

Deliberation that Slows Down the "Thinking" Process A narrow definition of the agent memory includes historical information within the same conversation (Zhang et al., 2024b), like the multi-round self-reflection prompting strategies that encourage agents to critically evaluate their previous responses (Shinn et al., 2024; Madaan et al., 2024; Wang et al., 2022b; Zhong et al., 2024; Lu et al., 2023). Historical conservations, usually from different or opposite perspectives that corrects initial errors, help produce a more error-prone and consistent final decision. Yao et al. (2024, 2023b); Besta et al. (2024) even support backtracking to further slow down the "thinking" process.

A broader definition of agent memory expands to historical information across multi-tasks and multi-agents (Zhang et al., 2024b). Corex (Sun et al., 2023a) finds that orchestrating multiple agents to work together yields better complex reasoning results, exceeding strong single-agent baselines (Wang et al., 2022b) by an average of 1.1-10.6%. Retroformer (Yao et al., 2023a) equips the single-agent Reflexion (Shinn et al., 2024) algorithm with an additional LLM to generate verbal reinforcement cues and assist its self-improvement, enhancing accuracy by 1.0-20.9%. ChatEval (Chan et al., 2023) introduces a multi-agent debate framework to mimic human annotators collaborating in robust answer evaluations. Its multi-agent approach achieves greater alignment with human preferences compared to single-agent evaluations, enhancing accuracy by 6.2% for GPT-3.5 and 2.5% for GPT-4, and an increase of 16.3% and 10.0% in average Spearman and Kendall-Tau correlations (Zhong et al., 2022) with human judgements in GPT-4. MetaAgents (Li et al., 2023j) effectively coordinates agents within task-oriented social contexts to achieve consistent behavior patterns, and the implementation of agent reflection in this system leads to a 21.0% improvement in success rates.

LM vs LM (Cohen et al., 2023), FORD (Xiong et al., 2023), Multi-Agent Debate (Liang et al., 2023; Du et al., 2023), DyLAN (Liu et al., 2023d), and Khan et al. (2024) highlight the profound impact of multi-agent collaboration through cross-examination and debates. These studies demonstrate substantial improvements in performance when multiple agents are orchestrated to work in collaboration. Specifically, LM vs LM (Cohen et al., 2023) illustrates how its multi-agent framework improves F1 scores by an average of 15.7% compared to the single-agent baseline (Yoshikawa and Okazaki, 2023). FORD (Xiong et al., 2023) reports an accuracy increase up to 4.9% compared to a single LLM. Liang et al. (2023) indicates significant improvements in accuracy — 17.0% for translation tasks and 16.0% for reasoning tasks — by employing a multi-agent strategy, effectively bridging the performance gap between GPT-3.5 and GPT-4 by harnessing multi-agents. Du et al. (2023) finds that multi-agent debates not only enhance reasoning performance by 8.0-14.8%, but more importantly, increase factual accuracy by 7.2-15.9%. DyLAN (Liu et al., 2023d) observes 3.5-4.1% in accuracy improvements over single-agent execution. Multi-agent debating in Khan et al. (2024) also leads to more truthful answers, boosting single-agent baselines by 28.0%. Multi-Agent Collaboration (Talebirad and Nadiri, 2023), ChatDev (Qian et al., 2023), AgentCF (Zhang et al., 2023), AutoGen (Wu et al., 2023), Social Learning (Mohtashami et al., 2023), S^3 (Gao et al., 2023a), Ke et al. (2024), and Chern et al. (2024) continue to push the frontier of a multi-agent system's applications beyond daily conversation to a versatile set of real-world task completions.

LLMs are also sensitive to prompt perturbations due to token bias and noises (Sclar et al., 2023a). One of the most worrying examples are adversarial attacks (Gehman et al., 2020; Ganguli et al., 2022; Du et al., 2022; Wei et al., 2024; Perez et al., 2022; Zou et al., 2023) through malicious prompt engineering, also named the Red Team Task. Chern et al. (2024) introduce a multi-agent debating approach with agents having harmless, neutral, or harmful agent intentions. It finds that multi-round multi-agent debate surpasses the self-reflection of a single agent, thus improving the invariance.

These collaborative approaches, in summary, allow each agent in a system to compare and rank its preference on choices from its own or from other agents through critical judgments. It helps enable the system to discern and output the most dominant decision as a consensus, thereby improving the orderability of preference. At the same time, through such a slow and critical thinking process, errors in initial responses or input prompts are more likely to be detected and corrected. Accumulated experience from past error planning contributes to a self-evolving process within the multi-agent system (Zhang et al., 2024b), resulting in a final response or a consensus that is less sensitive to specific wording or token bias, moving the response towards better consistency and invariance.

5 Evaluating Rationality of Agents

The amount of studies for testing rationality in multi-modal and multi-agent systems remains scant, despite the growing interest in the field. While there are numerous reasoning benchmarks available, including commonsense reasoning (Talmor et al., 2019), logical reasoning (Liu et al., 2021, 2023a), multi-hop reasoning (Yang et al., 2018), mathematical reasoning (Hendrycks et al., 2021), structured data reasoning (Chen et al., 2020), conceptual reasoning (Zhou et al., 2024), and general capabilities through multi-agent evaluations (Rasheed et al., 2024; Ma et al., 2024; Wang et al., 2024b; Abdelnabi et al., 2023).

However, they do not directly measure rationality. Many of these benchmarks fail to prove whether reasoning is actually used in solving the tasks, leaving no guarantee that these tasks will be solved consistently when generalized to other representations or domains. Issues such as data contamination (Magar and Schwartz, 2022; Dong et al., 2024; Sainz et al., 2023; Jacovi et al., 2023) further compound the problem, as some benchmarks may inadvertently include the training data of these LLMs, leading to inflated performance scores. Hence, even though solid reasoning will imply rationality, existing approaches fall short in making the logic click. In this section, we point to several existing ingredients that can constitute the bread-and-butter of future generations of evaluation approaches for rationality.

Adapting Cognitive Psychology Experiments Recent works propose adapting vignette-based experiments borrowed from cognitive psychology to test whether LLMs are susceptible to cognitive biases and fallacies. For instance, Binz and Schulz (2023) tested GPT-3 on the conjunction fallacy, finding that they exhibit human-like biases. However, many of these approaches are informal and subjective, failing to scale in a way that allows for drawing statistically significant conclusions. Moreover, LLMs may be subject to cognitive biases not existent in humans, such as the hypothetical "algorithmic bias" proposed by Bender et al. (2021), which could lead to unintended consequences in decision-making tasks. Further research is needed to uncover and characterize these potential biases.

Testing Grounding against Hallucination Information grounding is usually evaluated by the level of hallucination (Bang et al., 2023; Guerreiro et al., 2023; Huang et al., 2023), which can impact the rationality of agent systems. Multiple evaluation benchmarks targeting language-only dialogue have been proposed, such as BEGIN (Dziri et al., 2022b), HaluEval (Li et al., 2023f), DialFact (Gupta et al., 2021), FaithDial (Dziri et al., 2022a), AIS (Rashkin et al., 2023), and others (Zheng et al., 2023b; Das et al., 2023; Cao et al., 2021). In contrast, *benchmarks on multi-agent frameworks beyond language dialogue or those involving multi-modalities are very limited.* Liu et al. (2024a) moves beyond conversation to code generation; EureQA (Li et al., 2023a) focuses on reasoning chains; and TofuEval (Tang et al., 2024) evaluates hallucination in multi-domain summarization. Object hallucination (Rohrbach et al., 2018; Biten et al., 2022), POPE (Li et al., 2023h), and LLaVA-RLHF (Sun et al., 2023b) are the few examples evaluating multi-modal hallucination. The community needs more hallucination benchmarks to quantitatively evaluate the extent to which multi-modal and multi-agents reduce hallucinations in comparison with baselines.

Testing the Orderability of Preference There are almost no benchmarks for evaluating whether LLMs or agents have a consistent preference in the selection of available options. The Multiple Choice Problem (MCP) serves as a common testing ground. Zheng et al. (2023a) shows that LLMs are susceptible to changes in the positioning of options. Since the underlying logic remains the same, it also makes LLMs fail to pass the property of invariance. There are tons of MCP benchmarks (PaperswithcodeMCQA), but solely on the accuracy of selections, overlooking the consistency of preference. However, Robinson et al. (2023) highlights that the Proportion of Plurality Agreement (PPA) offers a measure of order invariance that does not depend on the model's ability to perform a task, suggesting a promising direction. We also need evaluations into vision and other modalities.

Testing the Principle of Invariance The realization of language can be independent of its meaning (Frege, 1892; Wittgenstein, 1953; Heineman, 2023). Recent studies within the realm of data contamination investigate whether LLMs can generate consistent responses across different, yet inherently equivalent, framings of the same task. These studies introduce perturbations to the original task descriptions to assess whether LLMs' responses will change significantly. Perturbation techniques include modifying instruction templates (Weber et al., 2023), paraphrasing task descriptions (Yang et al., 2023; Ohmer et al., 2024), or altering the order of in-context learning exemplars (Lu et al., 2021; Pecher et al., 2024). Specifically, approaches involve some versions of paraphrasing or permutation, such as changing the instruction templates (Weber et al., 2023), rewording task descriptions (Yang et al., 2023; Ohmer et al., 2024; Wang et al., 2024b), translating the prompts into a different language (Ohmer et al., 2023, 2024; Xu et al., 2024a) and then back to the original language (Yang et al., 2023), and making subtle changes to entities in task descriptions without affecting the logical structure, like altering names of the characters, numerical values in math problems, or locations of the events (Wang et al., 2024b). Permutation also includes reordering in-context learning examples (Lu et al., 2021; Pecher et al., 2024) and, in the case of multiple-choice questions, rearranging the options (Zong et al., 2023; Zheng et al., 2023a).

It is crucial to recognize that these perturbations are superficial; the altered task descriptions remain syntactically and semantically equivalent to their originals, although linguistic expressions or narratives may vary substantially. This observation highlights the need for developing methods that go beyond surface-level perturbations to effectively evaluate the robustness and invaraince of LLMs across diverse problem framings and modalities.

Testing Independence from Irrelevant Context We need more expansive perturbations to evaluate the independence from irrelevant context. Studies such as Shi et al. (2023), Wu et al. (2024), Liu et al. (2024d), and Yoran et al. (2023) have explored the phenomenon of "lost-in-context" by introducing random or misleading sentences into original problem statements. While earlier benchmarks like those from Weston et al. (2015), Sinha et al. (2019), Clark et al. (2020), and Webson and Pavlick (2021) have included irrelevant content, they have been predominantly limited to language modalities and single-agent systems. Recent benchmarks such as MileBench (Song et al., 2024), Mementos (Wang et al., 2024c), Seed-bench-2 (Li et al., 2023b), and DEMON (Li et al., 2023d) begin to evaluate multi-modal agents in long context or image sequences, where accurately responding to a specific question requires isolating only the relevant information from the long context window.

6 Open Problems and Future Directions

Inherent Rationality It is important to understand that integrating most of these agents or modules with LLMs still does not *inherently* make LLMs more rational. **Current methods are neither sufficient nor necessary, but they serves as instrumental tools that bridges the gap between an LLM's response and rationality.** These approaches enables multi-agent systems, which are black boxes from the user's perspective, to more closely mimic rational thinking in their output responses. However, despite these more rational responses elicited from multi-modal and multi-agent systems, the challenge of how to effectively close the loop and bake these enhanced outputs back into the LLMs (Zhao et al., 2024), beyond mere finetuning, remains an open topic. In other words, can we leverage these more rational outputs to inherently enhance a single foundation model's rationality in its initial responses in future applications?

More Comprehensive Evaluation on Rationality Section 4 thoroughly compares multi-modal and multi-agent systems over their LLM-based single-agent baselines. However, the choices of evaluation metrics are important (Schaeffer et al., 2024); these examples predominantly focus on the accuracy of their final performance, ignoring the most interesting intermediate reasoning steps and the concept of rationality. Section 5 furthermore acknowledges that while there have been some efforts to assess the rationality of agent systems, the field still lacks comprehensive and rigorous evaluation metrics. Moreover, **most existing benchmarks on rationality provide limited comparisons between multi-agent frameworks and single-agent baselines**, thus failing to fully elucidate the advantages multi-agent frameworks can offer.

Future research should prioritize the development of more robust and scalable methods for evaluating rationality, taking into account unique challenges and biases posed by agents. **A promising direction is to create benchmarks specifically tailored to assess rationality, going beyond existing ones on accuracy.** These new benchmarks should avoid data contamination and emphasize tasks that demand consistent reasoning across diverse representations and domains. There is a need for more rigorous and large-scale studies on the principles of invariance and orderability of preference, together with their applications to testing rationality in agent systems. This would involve developing more sophisticated perturbation methods that probe the consistency of reasoning at a deeper level, as well as designing experiments that yield statistically significant results.

Encouraging More Multi-Modal Agents in Multi-Agent Systems Research into the integration of multi-modality within multi-agent systems would be promising. Fields such as multi-agent debate, collaboration, and neuro-symbolic reasoning, as shown in Figure 1, currently under-utilize the potential of multi-modal sensory inputs. We believe that expanding the role of multi-modalities, including but not limited to vision, sounds, and structured data could significantly enhance the capabilities and rationality of multi-agent systems.

7 Conclusions

This survey builds connections between multi-modal and multi-agent systems with rationality, guided by dual-process theories and the four axioms we expect a rational agent or agent systems should satisfy: *information grounding, orderability of preference, independence from irrelevant context, and invariance across logically equivalent representations*. Our findings suggest that the grounding can usually be enhanced by multi-modalities, world models, knowledge retrieval, and tool utilization. The remaining three axioms are typically intertwined, and we sometimes describe their collective characteristics informally using terms such as coherence, consistency, and trustworthiness. These axioms are simultaneously improved by achievements in multi-modalities, tool utilization, neuro-symbolic reasoning, self-reflection, and multi-agent collaborations. These fields of research, by either slowing down the "thinking" process or boiling down real-world problems to logical essence, mimic the *"System 2"* thinking in human cognition, thereby enhancing the rationality of multi-agent systems in decision-making scenarios, compared to single-agent language-only baselines that resemble the *"System 1"* process.

Collaboration between the AI research community and cognitive psychologists could be particularly fruitful. We need better evaluation benchmarks on the rationality of agents, more exploration to mitigate cognitive biases in multi-modal and multi-agent systems, and deeper understanding of how these biases arise and how they can be mitigated, ultimately enhancing rationality in decision-making processes.

8 Limitations

The fields of multi-modal and multi-agent systems are rapidly evolving. Despite our best efforts, it is inherently impossible to encompass all related works within the scope of this survey. Our discussion also possesses limited mention of the reasoning capabilities, theory of mind in machine psychology, and cognitive architectures, all of which lie beyond the scope of this survey but are crucial for a deeper understanding of LLMs and agent systems. Furthermore, the concept of rationality in human cognitive science may encompass more principles and axioms than those defined in our survey.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2023. Llm-deliberation: Evaluating llms with interactive multi-agent negotiation games. *arXiv preprint arXiv:2309.17234*.

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, et al. 2022. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Ian A Apperly and Stephen A Butterfill. 2009. Do humans have two systems to track beliefs and belief-like states? *Psychological review*, 116(4):953.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2024. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. 2022. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1381–1390.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2021. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. *arXiv preprint arXiv:2109.09784*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *arXiv preprint arXiv:2304.00723*.

Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, et al. 2024. Exploring large language model based intelligent agents: Definitions, methods, and prospects. *arXiv preprint arXiv:2401.03428*.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, et al. 2022. Binding language models in symbolic languages. *arXiv preprint arXiv:2210.02875*.

Inyoung Cheong, King Xia, KJ Feng, Quan Ze Chen, and Amy X Zhang. 2024. (a) i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. *arXiv preprint arXiv:2402.01864*.

Steffi Chern, Zhen Fan, and Andy Liu. 2024. Combating adversarial attacks with multi-agent debate. *arXiv preprint arXiv:2401.05998*.

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into automatic evaluation using large language models. *arXiv preprint arXiv:2310.05657*.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. *arXiv preprint arXiv:2002.05867*.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*.

Fergus IM Craik and Robert S Lockhart. 1972. Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, 11(6):671–684.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Souvik Das, Sougata Saha, and Rohini K Srihari. 2023. Diving deep into modes of fact hallucinations in dialogue systems. *arXiv preprint arXiv:2301.04449*.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2024. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. *arXiv preprint arXiv:2402.15938*.

Wei Du, Yichun Zhao, Boqun Li, Gongshen Liu, and Shilin Wang. 2022. Ppt: Backdoor attacks on pre-trained models via poisoned prompt tuning. In *IJCAI*, pages 680–686.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. 2024. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M Ponti, and Siva Reddy. 2022a. Faithdial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022b. Evaluating attribution in dialogue systems: The begin benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in high-stakes decision-making with llms. *arXiv preprint arXiv:2403.00811*.

Franz Eisenführ, Martin Weber, and Thomas Langer. 2010. *Rational decision making*. Springer.

Jonathan St BT Evans. 2003. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10):454–459.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. 2022. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35:18343–18362.

Meng Fang, Shilong Deng, Yudi Zhang, Zijing Shi, Ling Chen, Mykola Pechenizkiy, and Jun Wang. 2024. Large language models are neurosymbolic reasoners. *arXiv preprint arXiv:2401.09334*.

Gottlob Frege. 1892. über sinn und bedeutung ["on sense and reference"]. *Zeitschrift für Philosophie und philosophische Kritik*, 100(1):25–50.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Hiroki Furuta, Ofir Nachum, Kuang-Huei Lee, Yutaka Matsuo, Shixiang Shane Gu, and Izzeddin Gur. 2023. Multimodal web navigation with instruction-finetuned foundation models. *arXiv preprint arXiv:2305.11854*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023a. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*.

Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. 2023b. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *arXiv preprint arXiv:2306.08640*.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023c. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Significant Gravitas. 2023. Autogpt. *Python. https://github.com/Significant-Gravitas/ Auto-GPT*.

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. Dialfact: A benchmark for fact-checking in dialogue. *arXiv preprint arXiv:2110.08222*.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962.

Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. 2023. A real-world webagent with planning, long context understanding, and program synthesis. *arXiv preprint arXiv:2307.12856*.

Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*.

Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. 2024. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*.

Reid Hastie and Robyn M Dawes. 2009. *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage Publications.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*.

David Heineman. 2023. Rethinking reasoning evaluation with theories of intelligence.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023a. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. 2023b. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*.

Joy Hsu, Jiayuan Mao, Josh Tenenbaum, and Jiajun Wu. 2024. What's left? concept grounding with logic-enhanced foundation models. *Advances in Neural Information Processing Systems*, 36.

Ziniu Hu, Ahmet Iscen, Chen Sun, Kai-Wei Chang, Yizhou Sun, David Ross, Cordelia Schmid, and Alireza Fathi. 2024. Avis: Autonomous visual information seeking with large language model agent. *Advances in Neural Information Processing Systems*, 36.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. *arXiv preprint arXiv:2305.10160*.

Bowen Jiang, Zhijun Zhuang, Shreyas S Shivakumar, Dan Roth, and Camillo J Taylor. 2024. Multi-agent vqa: Exploring multi-agent foundation models in zero-shot visual question answering. *arXiv preprint arXiv:2403.14783*.

Philip Nicholas Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. 6. Harvard University Press.

Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of large language models in finance: An empirical examination of hallucination. *arXiv preprint arXiv:2311.15548*.

Yu He Ke, Rui Yang, Sui An Lie, Taylor Xin Yi Lim, Hairil Rizal Abdullah, Daniel Shu Wei Ting, and Nan Liu. 2024. Enhancing diagnostic accuracy through multi-agent conversations: Using large language models to mitigate cognitive bias. *arXiv preprint arXiv:2401.14589*.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*.

Roni Khardon and Dan Roth. 1997. Learning to reason. *Journal of the ACM (JACM)*, 44(5):697–725.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *arXiv:2304.02643*.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*.

Michal Kosinski. 2023. Evaluating large language models in theory of mind tasks. *arXiv e-prints*, pages arXiv–2302.

John E Laird. 2019. *The Soar cognitive architecture*. MIT press.

Yann LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1).

Yann LeCun. 2024. Objective-driven ai: Towards ai systems that can learn, remember, reason, plan, have common sense, yet are steerable and safe. University of Washington, Department of Electrical & Computer Engineering. Slide presentation retrieved from University of Washington.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Selma Leydesdorff and Katharine Hodgkin. 2017. *Memory cultures: Memory, subjectivity and recognition*. Routledge.

Bangzheng Li, Ben Zhou, Fei Wang, Xingyu Fu, Dan Roth, and Muhao Chen. 2023a. Deceiving semantic shortcuts on reasoning chains: How far can models go without hallucination? *arXiv preprint arXiv:2311.09702*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2023b. Seedbench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. 2024a. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214.

Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. 2023c. Theory of mind for multi-agent collaboration via large language models. *arXiv preprint arXiv:2310.10701*.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. 2023d. Finetuning multimodal llms to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024b. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023e. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023f. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq Joty, and Soujanya Poria. 2023g. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *arXiv preprint arXiv:2305.13269*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023h. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023i. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 374–382.

Yuan Li, Yixuan Zhang, and Lichao Sun. 2023j. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *arXiv preprint arXiv:2310.06500*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Evan Zheran Liu, Kelvin Guu, Panupong Pasupat, Tianlin Shi, and Percy Liang. 2018. Reinforcement learning on web interfaces using workflow-guided exploration. *arXiv preprint arXiv:1802.08802*.

Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, and Li Zhang. 2024a. Exploring and evaluating hallucinations in llm-powered code generation. *arXiv preprint arXiv:2404.00971*.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023a. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024b. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3622–3628.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024d. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023c. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023d. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. 2023. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *arXiv preprint arXiv:2308.08239*.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2024. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. 2024. Agentboard: An analytical evaluation board of multi-turn llm agents. *arXiv preprint arXiv:2401.13178*.

Olivia Macmillan-Scott and Mirco Musolesi. 2024. (ir) rationality and cognitive biases in large language models. *arXiv preprint arXiv:2402.09193*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. *arXiv preprint arXiv:2203.08242*.

James G March and Herbert A Simon. 1958. Organizations. *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121.

Amirkeivan Mohtashami, Florian Hartmann, Sian Gooding, Lukas Zilka, Matt Sharifi, et al. 2023. Social learning: Towards collaborative learning with large language models. *arXiv preprint arXiv:2312.11441*.

Anirban Mukherjee and Hannah Hanwen Chang. 2024. Heuristic reasoning in ai: Instrumental use and mimetic absorption. *arXiv preprint arXiv:2403.09404*.

Yohei Nakajima. 2023. Babyagi. *Python. https://github.com/yoheinakajima/babyagi*.

Maxwell Nye, Michael Tessler, Josh Tenenbaum, and Brenden M Lake. 2021. Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. *Advances in Neural Information Processing Systems*, 34:25192–25204.

Ini Oguntola, Dana Hughes, and Katia Sycara. 2021. Deep interpretable models of theory of mind. In *2021 30th IEEE international conference on robot & human interactive communication (RO-MAN)*, pages 657–664. IEEE.

Xenia Ohmer, Elia Bruni, and Dieuwke Hupkes. 2023. Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses. *CoRR, abs/2305.11662*.

Xenia Ohmer, Elia Bruni, and Dieuwke Hupkes. 2024. From form (s) to meaning: Probing the semantic depths of language models using multisense consistency. *arXiv preprint arXiv:2404.12145*.

OpenAI. 2023. Gpt-4v(ision) system card.

OpenAI. 2024. Gpt-4o. Software available from OpenAI. Accessed: 2024-05-20.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*.

Zhenyu Pan, Haozheng Luo, Manling Li, and Han Liu. 2024. Chain-of-action: Faithful and multimodal question answering through large language models. *arXiv preprint arXiv:2403.17359*.

PaperswithcodeMCQA. Multiple choice qa. https://paperswithcode.com/task/multiple-choice-qa/latest. Accessed: 2024-05-28.

Branislav Pecher, Ivan Srba, and Maria Bielikova. 2024. On sensitivity of learning with limited labelled data to the effects of randomness: Impact of interactions and systematic choices. *arXiv preprint arXiv:2402.12817*.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. 2023. Adapt: As-needed decomposition and planning with language models. *arXiv preprint arXiv:2311.05772*.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*.

Shuofei Qiao, Ningyu Zhang, Runnan Fang, Yujie Luo, Wangchunshu Zhou, Yuchen Eleanor Jiang, Chengfei Lv, and Huajun Chen. 2024. Autoact: Automatic agent learning from scratch via self-planning. *arXiv preprint arXiv:2401.05268*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Zeeshan Rasheed, Muhammad Waseem, Kari Systä, and Pekka Abrahamsson. 2024. Large language model evaluation via multi ai agents: Preliminary results. *arXiv preprint arXiv:2404.01023*.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Raquel B Robinson, Karin Johansson, James Collin Fey, Elena Márquez Segura, Jon Back, Annika Waern, Sarah Lynne Bowman, and Katherine Isbister. 2023. Leveraging large language models for multiple choice question answering. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–5.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. *arXiv preprint arXiv:2310.18018*.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2024. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023a. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023b. Minding language models'(lack of) theory of mind: A plug-and-play multi-character belief tracker. *arXiv preprint arXiv:2306.00924*.

Reinhard Selten. 1990. Bounded rationality. *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, 146(4):649–658.

Peter Shaw, Mandar Joshi, James Cohan, Jonathan Berant, Panupong Pasupat, Hexiang Hu, Urvashi Khandelwal, Kenton Lee, and Kristina N Toutanova. 2024. From pixels to ui actions: Learning to follow instructions via graphical user interfaces. *Advances in Neural Information Processing Systems*, 36.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233.

Weizhou Shen, Chenliang Li, Hongzhan Chen, Ming Yan, Xiaojun Quan, Hehong Chen, Ji Zhang, and Fei Huang. 2024a. Small llms are weak tool learners: A multi-llm agent. *arXiv preprint arXiv:2401.07324*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2024b. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. 2017. World of bits: An open-domain platform for web-based agents. In *International Conference on Machine Learning*, pages 3135–3144. PMLR.

Zhengliang Shi, Shen Gao, Xiuyi Chen, Lingyong Yan, Haibo Shi, Dawei Yin, Zhumin Chen, Pengjie Ren, Suzan Verberne, and Zhaochun Ren. 2024. Learning to use tools via cooperative and interactive agents. *arXiv preprint arXiv:2403.03031*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L Hamilton. 2019. Clutrr: A diagnostic benchmark for inductive reasoning from text. *arXiv preprint arXiv:1908.06177*.

Robert L Solso and Jerome Kagan. 1979. *Cognitive psychology*. Houghton Mifflin Harcourt P.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. 2024. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Chitra Subramanian, Miao Liu, Naweed Khan, Jonathan Lenchner, Aporva Amarnath, Sarathkrishna Swaminathan, Ryan Riegel, and Alexander Gray. 2024. A neuro-symbolic approach to multi-agent rl for interpretability and probabilistic decision making. *arXiv preprint arXiv:2402.13440*.

Qiushi Sun, Zhangyue Yin, Xiang Li, Zhiyong Wu, Xipeng Qiu, and Lingpeng Kong. 2023a. Corex: Pushing the boundaries of complex reasoning through multi-model collaboration. *arXiv preprint arXiv:2310.00280*.

Ron Sun. 2001. *Duality of the mind: A bottom-up approach toward cognition*. Psychology Press.

Ron Sun. 2024. Can a cognitive architecture fundamentally enhance llms? or vice versa? *arXiv preprint arXiv:2401.10444*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023b. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Gaurav Suri, Lily R Slater, Ali Ziaee, and Morgan Nguyen. 2024. Do large language models show decision heuristics similar to humans? a case study using gpt-3.5. *Journal of Experimental Psychology: General*.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898.

Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Liyan Tang, Igor Shalyminov, Amy Wing-mei Wong, Jon Burnsky, Jake W Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, et al. 2024. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. *arXiv preprint arXiv:2402.13249*.

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, and Le Sun. 2023. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*.

Amos Tversky and Daniel Kahneman. 1988. Rational choice and the framing of decisions. *Decision making: Descriptive, normative, and prescriptive interactions*, pages 167–192.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models-a critical investigation. *Advances in Neural Information Processing Systems*, 36:75993–76005.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023b. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023c. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*.

Pengda Wang, Zilin Xiao, Hanjie Chen, and Frederick L Oswald. 2024a. Will the real linda please stand up... to large language models? examining the representativeness heuristic in llms. *arXiv preprint arXiv:2404.01461*.

Siyuan Wang, Zhuohan Long, Zhihao Fan, Zhongyu Wei, and Xuanjing Huang. 2024b. Benchmark self-evolving: A multi-agent framework for dynamic llm evaluation. *arXiv preprint arXiv:2402.11443*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023d. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022a. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. 2024c. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Zeqing Wang, Wentao Wan, Runmeng Chen, Qiqing Lao, Minjie Lang, and Keze Wang. 2023e. Towards top-down reasoning: An explainable multi-agent approach for visual question answering. *arXiv preprint arXiv:2311.17331*.

Lucas Weber, Elia Bruni, and Dieuwke Hupkes. 2023. Mind the instructions: a holistic evaluation of consistency and interactions in prompt-based learning. *arXiv preprint arXiv:2310.13486*.

Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2024. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Wikipedia contributors. 2004. Plagiarism — Wikipedia, the free encyclopedia. [Online; accessed 22-July-2004].

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Blackwell.

Lionel Wong, Jiayuan Mao, Pratyusha Sharma, Zachary S Siegel, Jiahai Feng, Noa Korneev, Joshua B Tenenbaum, and Jacob Andreas. 2023. Learning adaptive planning representations with natural language guidance. *arXiv preprint arXiv:2312.08566*.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-modal answer validation for knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2712–2721.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? *arXiv preprint arXiv:2404.03302*.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024a. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*.

Yangxinyu Xie, Tanwi Mallick, Joshua David Bergerson, John K Hutchison, Duane R Verner, Jordan Branham, M Ross Alexander, Robert B Ross, Yan Feng, Leslie-Anne Levy, et al. 2024b. Wildfiregpt: Tailored large language model for wildfire analysis. *arXiv preprint arXiv:2402.07877*.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Diving into the inter-consistency of large language models: An insightful analysis through debate. *arXiv preprint arXiv:2305.11595*.

Shaoyang Xu, Weilong Dong, Zishan Guo, Xinwei Wu, and Deyi Xiong. 2024a. Exploring multilingual human value concepts in large language models: Is value alignment consistent, transferable and controllable across languages? *arXiv preprint arXiv:2402.18120*.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-seng Chua. 2023. Search-in-the-chain: Towards the accurate, credible and traceable content generation for complex knowledge-intensive tasks. *arXiv preprint arXiv:2304.14732*.

Xinrun Xu, Yuxin Wang, Chaoyi Xu, Ziluo Ding, Jiechuan Jiang, Zhiming Ding, and Börje F Karlsson. 2024b. A survey on game playing agents and large models: Methods, applications, and challenges. *arXiv preprint arXiv:2403.10249*.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E Gonzalez, and Ion Stoica. 2023. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. 2024. Doraemongpt: Toward understanding dynamic scenes with large language models. *arXiv preprint arXiv:2401.08392*.

Zukang Yang and Zixuan Zhu. 2024. Curious-llm: Elevating multi-document qa with reasoning-infused knowledge graph prompting. *arXiv preprint arXiv:2404.09077*.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022a. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022b. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Murthy, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, et al. 2023a. Retroformer: Retrospective large language agents with policy gradient optimization. *arXiv preprint arXiv:2308.02151*.

Yao Yao, Zuchao Li, and Hai Zhao. 2023b. Beyond chain-of-thought, effective graph-of-thought reasoning in large language models. *arXiv preprint arXiv:2305.16582*.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2022. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*.

Hongbin Ye, Honghao Gui, Aijia Zhang, Tong Liu, Wei Hua, and Weiqiang Jia. 2023. Beyond isolation: Multi-agent synergy for improving knowledge graph construction. *arXiv preprint arXiv:2312.03022*.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31.

Da Yin, Faeze Brahman, Abhilasha Ravichander, Khyathi Chandu, Kai-Wei Chang, Yejin Choi, and Bill Yuchen Lin. 2023. Lumos: Learning agents with unified data, modular design, and open-source llms. *arXiv preprint arXiv:2311.05657*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Hiyori Yoshikawa and Naoaki Okazaki. 2023. Selective-lama: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2017–2028.

Eric Zelikman, Qian Huang, Gabriel Poesia, Noah Goodman, and Nick Haber. 2023. Parsel: Algorithmic reasoning with language models by composing decompositions. *Advances in Neural Information Processing Systems*, 36:31466–31523.

Junjie Zhang, Yupeng Hou, Ruobing Xie, Wenqi Sun, Julian McAuley, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023. Agentcf: Collaborative learning with autonomous language agents for recommender systems. *arXiv preprint arXiv:2310.09233*.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024a. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024b. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*.

Shu Zhao and Huijuan Xu. 2023. Less is more: Toward zero-shot local scene graph generation via foundation models. *arXiv preprint arXiv:2310.01356*.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2023. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*.

Zhonghan Zhao, Ke Ma, Wenhao Chai, Xuan Wang, Kewei Chen, Dongxu Guo, Yanting Zhang, Hongwei Wang, and Gaoang Wang. 2024. Do we really need a complex agent system? distill embodied agent into a single model. *arXiv preprint arXiv:2404.04619*.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024a. Gpt-4v (ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023a. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023b. Why does chatgpt fall short in providing truthful answers. *ArXiv preprint, abs/2304.10513*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

Ben Zhou, Hongming Zhang, Sihao Chen, Dian Yu, Hongwei Wang, Baolin Peng, Dan Roth, and Dong Yu. 2024. Conceptual and unbiased reasoning in language models. *arXiv preprint arXiv:2404.00205*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, et al. 2023b. Ghost in the minecraft: Generally capable agents for open-world enviroments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*.

Yongshuo Zong, Tingyang Yu, Bingchen Zhao, Ruchika Chavhan, and Timothy Hospedales. 2023. Fool your (vision and) language model with embarrassingly simple permutations. *arXiv preprint arXiv:2310.01651*.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Orderability of Preferences.

Comparability When faced with any two alternatives A and B, the agent should have at least a weak preference, i.e., $A \succeq B$ or $B \succeq A$. This means that the agent can compare any pair of alternatives and determine which one is preferred or if they are equally preferred.

Transitivity If the agent prefers A to B and B to C, then the agent must prefer A to C. This ensures that the agent's preferences are consistent and logical across multiple comparisons.

Closure If A and B are in the alternative set S, then any probabilistic combination of A and B (denoted as ApB) should also be in S. This principle ensures that the set of alternatives is closed under probability mixtures.

Distribution of probabilities across alternatives If A and B are in S, then the probability mixture of (ApB) and B, denoted as [(ApB)qB], should be indifferent to the probability mixture of A and B, denoted as (ApqB). This principle ensures consistency in the agent's preferences when dealing with probability mixtures of alternatives.

Solvability When faced with three alternatives A, B, and C, with the preference order $A \succeq B \succeq C$, there should be some probabilistic way of combining A and C such that the agent is indifferent between choosing B or this combination. In other words, the agent should be able to find a solution to the decision problem by making trade-offs between alternatives.

One consequence of the orderability is the concept of **dominance**: If alternative A is better than alternative B in terms of one attribute and at least as good in terms of all other attributes, the dominant option A should be chosen. An example of a fallacy that violates dominance is the sunk cost fallacy, where an agent continues to invest in a suboptimal alternative due to past investments, despite the availability of better options based on future outcomes.

B The Orderability of Preferences Matters for LLM-based Evaluations

This section talks about LLM-based evaluation rather than evaluating the rationality of LLMs discussed in Section 5. Recent research underscores a critical need for more rational LLM-based evaluation methods, particularly for assessing open-ended language responses. CoBBLEr (Koo et al., 2023) provides a cognitive bias benchmark for evaluating LLMs as evaluators, revealing a preference for their own outputs over those from other LLMs. Stureborg et al. (2024) argues that LLMs are biased evaluators towards more familiar tokens and previous predictions, and exhibit strong self-inconsistency in the score distribution. Luo et al. (2023); Shen et al. (2023); Gao et al. (2023c); Wang et al. (2023b); Chen et al. (2023); Chiang and Lee (2023); Zheng et al. (2024b); Fu et al. (2023); Liu et al. (2023c) also point out the problem with a single LLM as the evaluator, with concerns over factual and rating inconsistencies, a high dependency on prompt design, a low correlation with human evaluations, and struggles with the comparison. As a result, having a coherent orderability of preferences aligned with human preference becomes increasingly important.

Multi-agent systems might be a possible remedy. By involving multiple evaluative agents from diverse perspectives, it becomes possible to achieve a more balanced and consistent orderability of preferences. For instance, ChatEval (Chan et al., 2023) posits that a multi-agent debate evaluation usually offers judgments that are better aligned with human annotators compared to single-agent ones. Bai et al. (2024) also finds decentralized methods yield fairer evaluation results.