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

Bowen Jiang [1] **Yangxinyu Xie** [1 2] **Xiaomeng Wang** [1] **Weijie J. Su** [1] **Camillo J. Taylor** [1] **Tanwi Mallick** [2]

Abstract

Rationality is 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. 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 survey aims to define some axioms of rationality, understand whether multi-modal and multi-agent systems are advancing toward rationality, identify their advancements over single-agent, language-only baselines, and discuss open problems and future directions.

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., 2023; 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 & Schulz, 2023; Echterhoff et al., 2024; Mukherjee & Chang, 2024; Macmillan-Scott & Musolesi, 2024; Wang et al., 2024; 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., 2023d; Kang & 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., 2023a; Wang et al., 2023c; 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 & Nadiri, 2023; Madaan et al., 2024; Cohen et al., 2023; Shinn et al., 2023; 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

[1]University of Pennsylvania, Philadelphia, PA, 19104, USA [2]Argonne National Laboratory, Lemont, IL, 60439, USA. Correspondence to: Bowen Jiang <bwjiang@seas.upenn.edu>, Yangxinyu Xie <xinyux@wharton.upenn.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

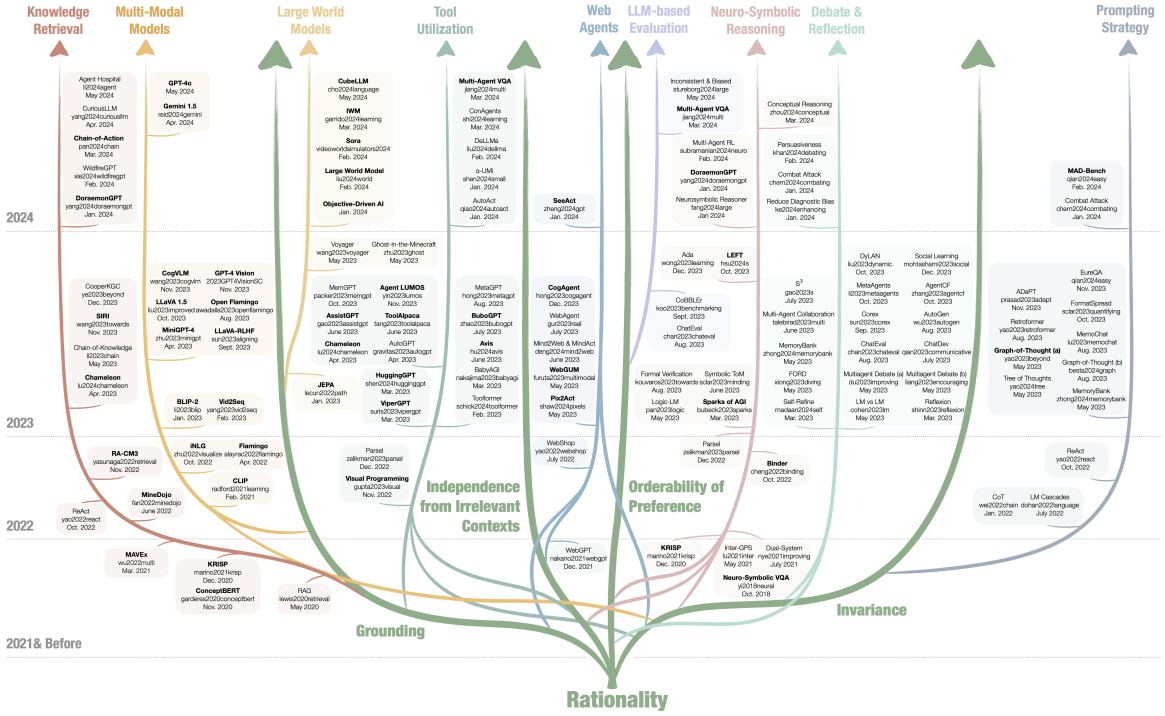

Figure 1. The evolutionary tree of multi-agent and/or multi-modal systems related to the four axioms of rationality. Many proposed approaches strive to address multiple axioms simultaneously. **Bold** fonts are used to mark works that involve multi-modalities. This tree also includes foundational works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

perspective, 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 multi-modal and/or multi-agent systems. Drawing from cognitive science, we first delineate four fundamental requirements for rational thinking. We then discuss 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. 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. Therefore, drawing on foundational works in rational decision-making (Tversky & Kahneman, 1988; Hastie & 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. 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 details in Appendix 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., 2024c), 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., 2024; Zhang et al., 2024a; 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 relationship between rationality and these multi-modal and multi-agent systems**, exploring how they contribute to enhancing rationality in decision making. We emphasize that *rationality*, by definition, is not equivalent to *reasoning* or *Theory of Mind*, although they are deeply intertwined. We leave explanations to Appendix C.

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, 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 into its 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.

4.1. Towards Grounding through Multi-Modal Models

Multi-modal approaches aim to improve information grounding across various channels, such as language and vision. By incorporating multi-modal models (Radford et al., 2021; Alayrac et al., 2022; Awadalla et al., 2023; Liu et al., 2024a; 2023a; Wang et al., 2023c; Zhu et al., 2023a; OpenAI, 2023; 2024; Reid et al., 2024), multi-agent systems can greatly expand their capabilities, enabling a richer, more accurate, and contextually aware interpretation of environment. 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. We leave more discussions to Appendix D.

4.2. Towards Grounding through Knowledge Retrieval

The existing transformer architecture (Vaswani et al., 2017) fundamentally limits how much information LLMs 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.

A multi-agent system can include planning agents in its framework, which determine how and where to retrieve external knowledge, and what specific information to acquire. External knowledge source could be a knowledge graph (Gardères et al., 2020; Hogan et al., 2021), a database (Lu et al., 2024; Xie et al., 2024b), and more. Additionally, the system can have summarizing agents that utilize retrieved knowledge to enrich the system's language outputs with better factuality. For example, 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). We provide a detailed survey of how multi-agent systems surpass single-agent baselines in Appendix E.

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

Similar to knowledge retrieval, 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 understand 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 & Kembhavi, 2023) generates Python programs to reliably execute subroutines. We provide more examples in Appendix F.

In most cases, utilizing 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 context 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. This improves the invariance property from noisy queries and independence of irrelevant context. Examples are shown in Appendix F.

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

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 and logical principles, which is unachievable by LLMs alone. Logic-LM (Pan et al., 2023), for example, combines problem formulating, symbolic reasoning, and summarizing agents, where the symbolic reasoner empowers LLMs with deterministic symbolic solvers to perform inference, ensuring a correct answer is consistently chosen. These modules typically expect standardized input formats, enhancing invariance and independence similar to API calls of tool usage. More examples are included in Appendix G.

4.5. Towards Orderability of Preferences & Invariance through Reflection, Debate, and Prompt Strategies

Single agents with self-reflection prompting (Shinn et al., 2023) and multi-agent systems that promote debate and consensus can help align outputs more closely with deliberate and logical decision-making, thus enhancing rational reasoning. For instance, Corex (Sun et al., 2023) 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%. More similar results are discussed in Appendix H. 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 invariance.

5. Open Problems and Future Directions

This survey builds connections between multi-modal and multi-agent systems with rationality, guided by 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 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, which could be improved by achievements in multi-modalities, tool utilization, neuro-symbolic reasoning, self-reflection, and multi-agent collaborations.

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 bridge the gap between an LLM's response and rationality.** These approaches enable 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 fine-tuning, 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?

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 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.

Evaluation on Rationality Benchmarks on rationality are scarce. Future research should prioritize the development of 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.

6. 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, cognitive architectures, and evaluations on rationality, 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

Aghajanyan, A., Huang, B., Ross, C., Karpukhin, V., Xu, H., Goyal, N., Okhonko, D., Joshi, M., Ghosh, G., Lewis, M., et al. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*, 2022.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

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

Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Bai, Y., Ying, J., Cao, Y., Lv, X., He, Y., Wang, X., Yu, J., Zeng, K., Xiao, Y., Lyu, H., et al. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36, 2024.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O. K., Aggarwal, K., Som, S., Piao, S., and Wei, F. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17682–17690, 2024.

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

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., and Ramesh, A. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., and Liu, Z. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

Chen, Y., Wang, R., Jiang, H., Shi, S., and Xu, R. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *arXiv preprint arXiv:2304.00723*, 2023.

Cheng, Y., Zhang, C., Zhang, Z., Meng, X., Hong, S., Li, W., Wang, Z., Wang, Z., Yin, F., Zhao, J., et al. Exploring large language model based intelligent agents: Definitions, methods, and prospects. *arXiv preprint arXiv:2401.03428*, 2024.

Cheng, Z., Xie, T., Shi, P., Li, C., Nadkarni, R., Hu, Y., Xiong, C., Radev, D., Ostendorf, M., Zettlemoyer, L., et al. Binding language models in symbolic languages. *arXiv preprint arXiv:2210.02875*, 2022.

Cheong, I., Xia, K., Feng, K., Chen, Q. Z., and Zhang, A. X. (a) i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. *arXiv preprint arXiv:2402.01864*, 2024.

Chern, S., Fan, Z., and Liu, A. Combating adversarial attacks with multi-agent debate. *arXiv preprint arXiv:2401.05998*, 2024.

Chiang, C.-H. and Lee, H.-y. A closer look into automatic evaluation using large language models. *arXiv preprint arXiv:2310.05657*, 2023.

Cohen, R., Hamri, M., Geva, M., and Globerson, A. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*, 2023.

Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.-D., et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 958–979, 2024.

Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H., and Su, Y. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36, 2024.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Durante, Z., Huang, Q., Wake, N., Gong, R., Park, J. S., Sarkar, B., Taori, R., Noda, Y., Terzopoulos, D., Choi, Y., et al. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*, 2024.

Echterhoff, J., Liu, Y., Alessa, A., McAuley, J., and He, Z. Cognitive bias in high-stakes decision-making with llms. *arXiv preprint arXiv:2403.00811*, 2024.

Eisenführ, F., Weber, M., and Langer, T. *Rational decision making*. Springer, 2010.

Fan, L., Wang, G., Jiang, Y., Mandlekar, A., Yang, Y., Zhu, H., Tang, A., Huang, D.-A., Zhu, Y., and Anandkumar, A. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35:18343–18362, 2022.

Fang, M., Deng, S., Zhang, Y., Shi, Z., Chen, L., Pechenizkiy, M., and Wang, J. Large language models are neurosymbolic reasoners. *arXiv preprint arXiv:2401.09334*, 2024.

Fu, J., Ng, S.-K., Jiang, Z., and Liu, P. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023.

Furuta, H., Nachum, O., Lee, K.-H., Matsuo, Y., Gu, S. S., and Gur, I. Multimodal web navigation with instruction-finetuned foundation models. *arXiv preprint arXiv:2305.11854*, 2023.

Gao, C., Lan, X., Lu, Z., Mao, J., Piao, J., Wang, H., Jin, D., and Li, Y. S^3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*, 2023a.

Gao, D., Ji, L., Zhou, L., Lin, K. Q., Chen, J., Fan, Z., and Shou, M. Z. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *arXiv preprint arXiv:2306.08640*, 2023b.

Gao, M., Ruan, J., Sun, R., Yin, X., Yang, S., and Wan, X. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*, 2023c.

Gardères, F., Ziaeefard, M., Abeloos, B., and Lecue, F. Conceptbert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 489–498, 2020.

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

Guerreiro, N. M., Alves, D. M., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., and Martins, A. F. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517, 2023.

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

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

Gur, I., Furuta, H., Huang, A., Safdari, M., Matsuo, Y., Eck, D., and Faust, A. A real-world webagent with planning, long context understanding, and program synthesis. *arXiv preprint arXiv:2307.12856*, 2023.

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

Han, S., Zhang, Q., Yao, Y., Jin, W., Xu, Z., and He, C. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*, 2024.

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

He, K., Mao, R., Lin, Q., Ruan, Y., Lan, X., Feng, M., and Cambria, E. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*, 2023.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., et al. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37, 2021.

Hong, S., Zheng, X., Chen, J., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023a.

Hong, W., Wang, W., Lv, Q., Xu, J., Yu, W., Ji, J., Wang, Y., Wang, Z., Dong, Y., Ding, M., et al. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*, 2023b.

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

Hu, Z., Iscen, A., Sun, C., Chang, K.-W., Sun, Y., Ross, D., Schmid, C., and Fathi, A. Avis: Autonomous visual information seeking with large language model agent. *Advances in Neural Information Processing Systems*, 36, 2024.

Huang, J. and Chang, K. C.-C. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Jiang, B., Zhuang, Z., Shivakumar, S. S., Roth, D., and Taylor, C. J. Multi-agent vqa: Exploring multi-agent foundation models in zero-shot visual question answering. *arXiv preprint arXiv:2403.14783*, 2024.

Kang, H. and Liu, X.-Y. Deficiency of large language models in finance: An empirical examination of hallucination. *arXiv preprint arXiv:2311.15548*, 2023.

Ke, Y. H., Yang, R., Lie, S. A., Lim, T. X. Y., Abdullah, H. R., Ting, D. S. W., and Liu, N. Enhancing diagnostic accuracy through multi-agent conversations: Using large language models to mitigate cognitive bias. *arXiv preprint arXiv:2401.14589*, 2024.

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., and Perez, E. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*, 2024.

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

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything. *arXiv:2304.02643*, 2023.

Koo, R., Lee, M., Raheja, V., Park, J. I., Kim, Z. M., and Kang, D. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*, 2023.

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

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

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

LeCun, Y. 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, January 2024. URL https://www.ece.uw.edu/wp-content/uploads/2024/01/lecun-20240124-uw-lyttle.pdf. Slide presentation retrieved from University of Washington.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., Gao, J., et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214, 2024a.

Li, H., Chong, Y. Q., Stepputtis, S., Campbell, J., Hughes, D., Lewis, M., and Sycara, K. Theory of mind for multi-agent collaboration via large language models. *arXiv preprint arXiv:2310.10701*, 2023a.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Li, J., Wang, S., Zhang, M., Li, W., Lai, Y., Kang, X., Ma, W., and Liu, Y. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024b.

Li, X., Zhao, R., Chia, Y. K., Ding, B., Bing, L., Joty, S., and Poria, S. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *arXiv preprint arXiv:2305.13269*, 2023c.

Li, Y., Wang, S., Ding, H., and Chen, H. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 374–382, 2023d.

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

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Tu, Z., and Shi, S. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

Liu, E. Z., Guu, K., Pasupat, P., Shi, T., and Liang, P. Reinforcement learning on web interfaces using workflow-guided exploration. *arXiv preprint arXiv:1802.08802*, 2018.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Liu, H., Yan, W., Zaharia, M., and Abbeel, P. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024b.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024c.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023b.

Liu, Z., Zhang, Y., Li, P., Liu, Y., and Yang, D. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*, 2023c.

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

Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K.-W., Wu, Y. N., Zhu, S.-C., and Gao, J. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Luo, Z., Xie, Q., and Ananiadou, S. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*, 2023.

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

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Marino, K., Chen, X., Parikh, D., Gupta, A., and Rohrbach, M. 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*, pp. 14111–14121, 2021.

Mohtashami, A., Hartmann, F., Gooding, S., Zilka, L., Sharifi, M., et al. Social learning: Towards collaborative learning with large language models. *arXiv preprint arXiv:2312.11441*, 2023.

Mukherjee, A. and Chang, H. H. Heuristic reasoning in ai: Instrumental use and mimetic absorption. *arXiv preprint arXiv:2403.09404*, 2024.

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

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

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

OpenAI. Gpt-4v(ision) system card. 2023. URL https://api.semanticscholar.org/CorpusID:263218031.

OpenAI. Gpt-4o. Software available from OpenAI, 2024. URL https://openai.com/index/hello-gpt-4o/. Accessed: 2024-05-20.

Pan, L., Albalak, A., Wang, X., and Wang, W. Y. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*, 2023.

Pan, Z., Luo, H., Li, M., and Liu, H. Chain-of-action: Faithful and multimodal question answering through large language models. *arXiv preprint arXiv:2403.17359*, 2024.

Prasad, A., Koller, A., Hartmann, M., Clark, P., Sabharwal, A., Bansal, M., and Khot, T. Adapt: As-needed decomposition and planning with language models. *arXiv preprint arXiv:2311.05772*, 2023.

Qian, C., Cong, X., Yang, C., Chen, W., Su, Y., Xu, J., Liu, Z., and Sun, M. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023.

Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F., and Chen, H. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*, 2022.

Qiao, S., Zhang, N., Fang, R., Luo, Y., Zhou, W., Jiang, Y. E., Lv, C., and Chen, H. Autoact: Automatic agent learning from scratch via self-planning. *arXiv preprint arXiv:2401.05268*, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.-b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.

Sclar, M., Kumar, S., West, P., Suhr, A., Choi, Y., and Tsvetkov, Y. Minding language models'(lack of) theory of mind: A plug-and-play multi-character belief tracker. *arXiv preprint arXiv:2306.00924*, 2023.

Shaw, P., Joshi, M., Cohan, J., Berant, J., Pasupat, P., Hu, H., Khandelwal, U., Lee, K., and Toutanova, K. N. From pixels to ui actions: Learning to follow instructions via graphical user interfaces. *Advances in Neural Information Processing Systems*, 36, 2024.

Shen, C., Cheng, L., Nguyen, X.-P., You, Y., and Bing, L. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4215–4233, 2023.

Shen, W., Li, C., Chen, H., Yan, M., Quan, X., Chen, H., Zhang, J., and Huang, F. Small llms are weak tool learners: A multi-llm agent. *arXiv preprint arXiv:2401.07324*, 2024a.

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., and Zhuang, Y. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36, 2024b.

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., Schärli, N., and Zhou, D. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pp. 31210–31227. PMLR, 2023.

Shi, T., Karpathy, A., Fan, L., Hernandez, J., and Liang, P. World of bits: An open-domain platform for web-based agents. In *International Conference on Machine Learning*, pp. 3135–3144. PMLR, 2017.

Shi, Z., Gao, S., Chen, X., Yan, L., Shi, H., Yin, D., Chen, Z., Ren, P., Verberne, S., and Ren, Z. Learning to use tools via cooperative and interactive agents. *arXiv preprint arXiv:2403.03031*, 2024.

Shinn, N., Labash, B., and Gopinath, A. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*, 2023.

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

Stureborg, R., Alikaniotis, D., and Suhara, Y. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*, 2024.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

Sun, Q., Yin, Z., Li, X., Wu, Z., Qiu, X., and Kong, L. Corex: Pushing the boundaries of complex reasoning through multi-model collaboration. *arXiv preprint arXiv:2310.00280*, 2023.

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

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

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

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

Tang, Q., Deng, Z., Lin, H., Han, X., Liang, Q., and Sun, L. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*, 2023.

Tversky, A. and Kahneman, D. Rational choice and the framing of decisions. *Decision making: Descriptive, normative, and prescriptive interactions*, pp. 167–192, 1988.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.

Wang, J., Liang, Y., Meng, F., Sun, Z., Shi, H., Li, Z., Xu, J., Qu, J., and Zhou, J. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*, 2023b.

Wang, P., Xiao, Z., Chen, H., and Oswald, F. L. Will the real linda please stand up... to large language models? examining the representativeness heuristic in llms. *arXiv preprint arXiv:2404.01461*, 2024.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022a.

Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023c.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022b.

Wang, Z., Wan, W., Chen, R., Lao, Q., Lang, M., and Wang, K. Towards top-down reasoning: An explainable multi-agent approach for visual question answering. *arXiv preprint arXiv:2311.17331*, 2023d.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wikipedia contributors. Plagiarism — Wikipedia, the free encyclopedia, 2004. URL https://en.wikipedia.org/w/index.php?title=Plagiarism&oldid=5139350. [Online; accessed 22-July-2004].

Wong, L., Mao, J., Sharma, P., Siegel, Z. S., Feng, J., Korneev, N., Tenenbaum, J. B., and Andreas, J. Learning adaptive planning representations with natural language guidance. *arXiv preprint arXiv:2312.08566*, 2023.

Wu, J., Lu, J., Sabharwal, A., and Mottaghi, R. Multi-modal answer validation for knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2712–2721, 2022.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., and Wang, C. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

Wu, S., Xie, J., Chen, J., Zhu, T., Zhang, K., and Xiao, Y. How easily do irrelevant inputs skew the responses of large language models? *arXiv preprint arXiv:2404.03302*, 2024.

Xie, J., Chen, Z., Zhang, R., Wan, X., and Li, G. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*, 2024a.

Xie, Y., Mallick, T., Bergerson, J. D., Hutchison, J. K., Verner, D. R., Branham, J., Alexander, M. R., Ross, R. B., Feng, Y., Levy, L.-A., et al. Wildfiregpt: Tailored large language model for wildfire analysis. *arXiv preprint arXiv:2402.07877*, 2024b.

Xiong, K., Ding, X., Cao, Y., Liu, T., and Qin, B. Diving into the inter-consistency of large language models: An insightful analysis through debate. *arXiv preprint arXiv:2305.11595*, 2023.

Xu, S., Pang, L., Shen, H., Cheng, X., and Chua, T.-s. Search-in-the-chain: Towards the accurate, credible and traceable content generation for complex knowledge-intensive tasks. *arXiv preprint arXiv:2304.14732*, 2023.

Xu, X., Wang, Y., Xu, C., Ding, Z., Jiang, J., Ding, Z., and Karlsson, B. F. A survey on game playing agents and large models: Methods, applications, and challenges. *arXiv preprint arXiv:2403.10249*, 2024.

Yang, Z. and Zhu, Z. Curiousllm: Elevating multi-document qa with reasoning-infused knowledge graph prompting. *arXiv preprint arXiv:2404.09077*, 2024.

Yang, Z., Chen, G., Li, X., Wang, W., and Yang, Y. Doraemongpt: Toward understanding dynamic scenes with large language models. *arXiv preprint arXiv:2401.08392*, 2024.

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

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022b.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Yao, W., Heinecke, S., Niebles, J. C., Liu, Z., Feng, Y., Xue, L., Murthy, R., Chen, Z., Zhang, J., Arpit, D., et al. Retroformer: Retrospective large language agents with policy gradient optimization. *arXiv preprint arXiv:2308.02151*, 2023.

Yasunaga, M., Aghajanyan, A., Shi, W., James, R., Leskovec, J., Liang, P., Lewis, M., Zettlemoyer, L., and Yih, W.-t. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*, 2022.

Ye, H., Gui, H., Zhang, A., Liu, T., Hua, W., and Jia, W. Beyond isolation: Multi-agent synergy for improving knowledge graph construction. *arXiv preprint arXiv:2312.03022*, 2023.

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., and Tenenbaum, J. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018.

Yin, D., Brahman, F., Ravichander, A., Chandu, K., Chang, K.-W., Choi, Y., and Lin, B. Y. Lumos: Learning agents with unified data, modular design, and open-source llms. *arXiv preprint arXiv:2311.05657*, 2023.

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

Zelikman, E., Huang, Q., Poesia, G., Goodman, N., and Haber, N. Parsel: Algorithmic reasoning with language models by composing decompositions. *Advances in Neural Information Processing Systems*, 36:31466–31523, 2023.

Zhang, J., Hou, Y., Xie, R., Sun, W., McAuley, J., Zhao, W. X., Lin, L., and Wen, J.-R. Agentcf: Collaborative learning with autonomous language agents for recommender systems. *arXiv preprint arXiv:2310.09233*, 2023.

Zhang, Y., Mao, S., Ge, T., Wang, X., de Wynter, A., Xia, Y., Wu, W., Song, T., Lan, M., and Wei, F. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*, 2024a.

Zhang, Z., Bo, X., Ma, C., Li, R., Chen, X., Dai, Q., Zhu, J., Dong, Z., and Wen, J.-R. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*, 2024b.

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

Zhao, Y., Lin, Z., Zhou, D., Huang, Z., Feng, J., and Kang, B. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023.

Zhao, Z., Ma, K., Chai, W., Wang, X., Chen, K., Guo, D., Zhang, Y., Wang, H., and Wang, G. Do we really need a complex agent system? distill embodied agent into a single model. *arXiv preprint arXiv:2404.04619*, 2024.

Zheng, B., Gou, B., Kil, J., Sun, H., and Su, Y. Gpt-4v (ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614*, 2024a.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024b.

Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., Zhu, C., Ji, H., and Han, J. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*, 2022.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023a.

Zhu, X., Chen, Y., Tian, H., Tao, C., Su, W., Yang, C., Huang, G., Li, B., Lu, L., Wang, X., et al. 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*, 2023b.

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.

B. LLM-based Evaluations

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 & Lee (2023); Zheng et al. (2024b); Fu et al. (2023); Liu et al. (2023b) 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, i.e., the orderability of preferences.

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. Multi-Agent VQA (Jiang et al., 2024) relies on a group of LLM-based graders for evaluating zero-shot, open-world visual question answering, where exact answer matches are no longer feasible.

C. More Explanations on Scope

Rationality, by definition, is not equivalent to *reasoning* (Khardon & Roth, 1997; Huang & 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. Furthermore, reasoning is crucial when faced with conflicting data or competing objectives. It helps systems to weigh the evidence, consider alternative perspectives, and make trade-offs between different courses of action. Through reasoning, individuals can weigh the evidence, consider alternative perspectives, and make

trade-offs between different courses of action. This process allows for more nuanced and context-dependent decision-making while navigating the intricacies in the real world. , all of which are fundamental steps in making rational decisions.

Rationality is also different from Theory of Mind (ToM) (Apperly & Butterfill, 2009; Nye et al., 2021; Oguntola et al., 2021; Hagendorff, 2023; Li et al., 2023a; Sclar et al., 2023; 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.

D. More Related Work on Multi-Modal Models

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., 2023b), (Open) Flamingo (Alayrac et al., 2022; Awadalla et al., 2023), LLaVA (Liu et al., 2024a; 2023a), CogVLM (Wang et al., 2023c), 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.

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. We leave more discussions to Appendix D. 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 interfaces (GUIs) 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.

The concept of invariance is the cornerstone of Visual Question Answering (VQA) agents (Chen et al., 2022; Jiang et al., 2024; Wang et al., 2023d; Yi et al., 2018; Wang et al., 2022a; Bao et al., 2022; Zhao & 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., 2023d; Jiang et al., 2024; Zhao & 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.

E. More Related Work on Knowledge Retrieval

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 & 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., 2023d) 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., 2023c), WildfireGPT (Xie et al., 2024b), and Agent Hospital (Li et al., 2024b). Chain-of-Knowledge (Li et al., 2023c) even discovers that integrating multiple knowledge sources enhances performance by 2.1% compared to using a single source in its experiments.

F. More Related Work on Using Tools

VisProg (Gupta & Kembhavi, 2023), ViperGPT (Surís et al., 2023), and Parsel (Zelikman et al., 2023) generate Python programs to reliably execute subroutines. Gupta & 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, and 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.

To boil down the task into its logical essence, Multi-Agent VQA (Jiang et al., 2024), as an example, 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 & 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 of agents from irrelevant contexts, ensuring that their operations are streamlined and focused solely on necessary information.

G. More Related Work on Neuro-Symbolic Reasoning

Coherent Orderability of Preference SymbolicToM (Sclar et al., 2023) 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 substantial performance drops by 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 Beyond detailed symbolic reasoning steps, these modules typically expect a standardized input formats, similar to API calls of tool usage. 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%.

H. More Related Work on Reflection, Debate, and Memory

Corex (Sun et al., 2023) 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., 2023) equips the single-agent Reflexion (Shinn et al., 2023) 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., 2023e) 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., 2023c), 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 & 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., 2023c) 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 & 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.