What is the Agent Doing? Visualizing Agentic AI Querying Workflows

Harry Li, Gabriel Appleby, Kenneth Alperin, Steven R. Gomez, and Ashley Suh

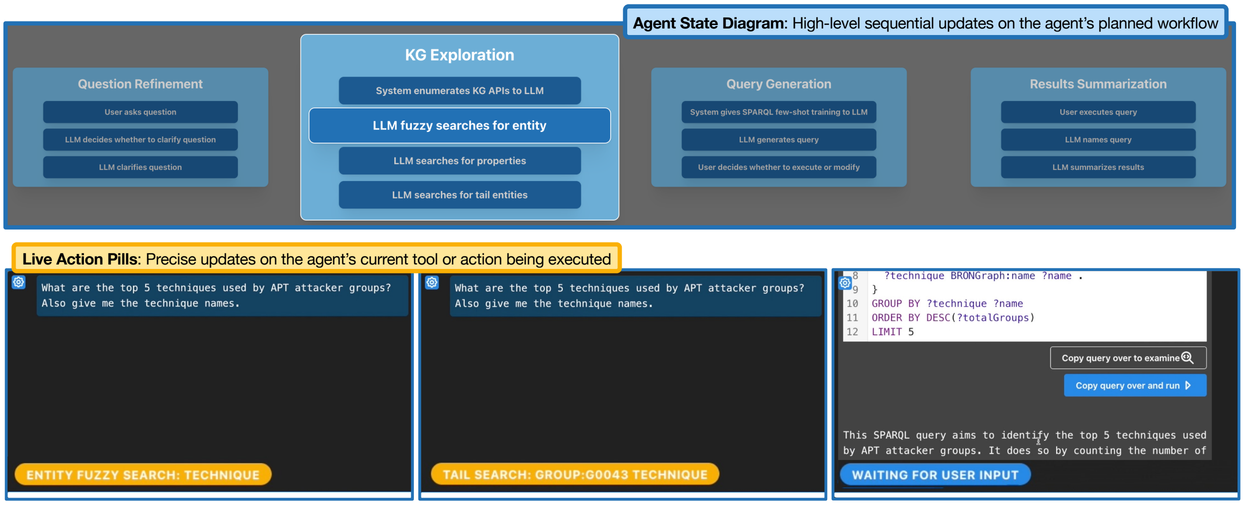

Fig. 1: The Agentic Trace Display helps users follow what an AI agent is doing as it works through querying a knowledge graph. At the top, the State Diagram shows what stage the agent is in within its larger workflow. Meanwhile, the Live Action Pill at the bottom offers real-time, detailed updates about the tool or action the agent is executing. With these two views, users can track both the overall process and the agent's specific actions as they happen.

Abstract—We explore how visualizations can help users understand what an AI agent is doing as it builds and runs queries over data. As part of the LinkQ system, a natural language interface for querying knowledge graphs with a large language model (LLM), we designed two complementary views: a State Diagram that shows where the agent is within a larger workflow, and a Live Action Pill that gives real-time updates about the agent's current task. In a study with 14 practitioners, we found that these visuals helped participants build stronger mental models of the agent's behavior while also increasing their confidence in the system. However, we also observed that users sometimes trusted incorrect outputs simply because the agent appeared to be doing the "right" thing. Our findings point to both the value and risk of visualizing agent behavior in interactive AI systems.

Index Terms—Natural language interface, knowledge graphs, large language models, query builder, trustworthy design, explainable AI

1 Introduction

Agentic AI systems are increasingly being used to support data analysis tasks. However, these systems are often opaque, leading to miscalibrated user trust. Users may blindly trust the agentic model and overlook critical failure modes, or, without a proper understanding of what the model is doing, be too skeptical to use the tool altogether.

In this paper, we present a dual *Agentic Trace Display* designed to enhance the transparency in LinkQ [18], an agentic natural language interface for querying knowledge graphs [15]. The display consists of two complementary visual components. The first is a *State Diagram* that outlines the high-level stages of the agentic workflow while highlighting the agent's current state within the pipeline. The second is a *Live Action Pill* situated within the chat panel, which provides a real-time description of the agent's current task (e.g., executing the action *"Fuzzy searching for the 'The Godfather' entity..."*).

- *Harry Li, Kenneth Alperin, Steven R Gomez, and Ashley Suh are with MIT Lincoln Laboratory. E-mail: {Harry.Li, Kenneth.Alperin, Steven.Gomez, Ashley.Suh}@mit.ll.edu*
- *Gabriel Appleby is with the National Renewable Energy Laboratory. E-mail: Gabriel.Appleby@nrel.gov.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

We intentionally build on familiar visualizations, such as workflow diagrams and real-time status indicators, to explore how visual traces of agentic workflows can influence user understanding and trust. We conducted a think-aloud study with 14 practitioners on both targeted and open-ended exploratory knowledge graph querying tasks. Participants reported that the Agentic Trace Display improved their mental model of what the agent was doing (its behavior) and what it would do next (its intentions), resulting in higher reported trust and confidence in the LinkQ system as a whole. At times we observed that the Agentic Trace Display instilled confidence in LinkQ's *incorrect* answers, suggesting a nuanced role for trace visualizations in agentic AI tools.

For the remainder of this paper, we briefly discuss related work in this area, describe the design of the Agentic Trace Display, and report the findings of our think-aloud study. Finally, we discuss the implications for calibrating trust in agentic AI systems with visualization design.

2 Background & Related Work

The Agentic Trace Display visualizes the agent's workflow in LinkQ [18], a natural language interface that supports querying knowledge graphs (KGs). KGs are a common data structure for modeling complex relationships in data, often used for question-answer and recommendation systems in domains such as web search, life sciences, and cyber-security [9, 36]. For a thorough background on KGs, including their use cases and historical challenges, see [15, 17, 20].

For this paper, we follow Anthropic's definition of *agentic workflows*: structured processes where large language models (LLMs) interact with tools through predefined paths to ensure reliability in completing tasks[1]. LinkQ's workflow utilizes LLMs for dynamic tool-calling, where tools are different functions that represent components of the KG querying pipeline (e.g., API calls, entity resolution, query construction, and so on). Sapkota et al. [29] outline the conceptual differences between AI agents and agentic AI. Dhanoa et al. [8] provide design patterns for agentic visualization with respect to human sensemaking [27]. For the remainder of the work, we use *LLM* and *agent* interchangeably, as well as a *tool* and *function*.

The usage of LLMs for facilitating query building (particularly for KGs) has been proposed dating back to 2019 [2, 26], and their unification has been studied in two distinct roadmaps [19, 24]. The past two years has shown increasing promise for utilizing LLMs for various types of data exploration and visual analysis tasks [1, 4, 11, 33, 34, 34, 35]. The resurgence of agentic AI, as well as its cataloged challenges (e.g., hallucinations, accountability) demonstrates that visualization plays an essential role in illuminating the behavior and explainability of these technologies [8, 21].

This said, the integration of LLM agents into visualization tools introduces new risks—particularly in how users interpret, trust, and rely on LLM outputs, regardless of their correctness [28]. Research in explainable AI has shown that LLMs can inadvertently lead to overreliance on AI [5], resulting in overtrust for model recommendations (even when errors are present) [12, 13], or can confuse users when LLM outputs clash with their mental models of the data [6]. Our qualitative study (Section 5) shows that the Agentic Trace Display provides high transparency into the agent's workflow, empowering users to 'trust' the LLM. This trust still needs to be fine tuned, as prior research (as well as some feedback from our study) indicates that well-designed agentic visualizations can foster overtrust.

3 Agentic Trace Display

Briefly, LinkQ is a natural language interface designed to let users ask questions about data in a KG and have an agent generate a query to answer those questions. Tools that the agent may use to answer the user's question include: (1) iterative question refinement, (2) entity resolution, (3) property or relation resolution, (4) query generation, (5) query execution, (6) results summarization. For more detailed information about LinkQ, see the open source project at https://github.com/mit-ll/linkq.

4 Workflow Visualizations

As LinkQ progresses through the agentic workflow, the system tracks both general and specific workflow states and displays them to the user.

State Diagram: The state diagram is located in the top panel of LinkQ (see Figures 1 and 2) and provides users with a transparent mental model of how the system works. The state diagram displays four primary protocol states: (1) Question Refinement, (2) KG Exploration, (3) Query Generation, and (4) Results Summarization. Each state also shows a detailed breakdown of sub-states. As the user interacts with LinkQ, the diagram highlights the precise stage the system is currently in, e.g., *KG Exploration → LLM fuzzy searches for entity* – and also reduces the size and brightness of the inactive states. This helps the user identify what the LLM is currently doing and which steps come next. As users become more familiar with the system, they can choose to display only the high level states, or hide the state diagram altogether (Figure 3).

Live Action Pill: For simplicity, LinkQ's chat panel hides agentic responses from the LLM that are not directed at the user (although power users can choose to show all messages). While the user waits for LinkQ to finish answering their question, the chat panel shows the Live Action Pill, a fine-grained loading message that explains what the LLM is doing at every step, shown in Figure 4. The Live Action Pill provides more specific detail than the State Diagram, showing messages such as:

- ENTITY FUZZY SEARCH: *[search string]*
- PROPERTIES SEARCH: *[entity ID]*
- TAIL SEARCH: *[head entity ID, property ID]*

Although not explicitly discussed in their release notes[2], OpenAI has recently implemented a similar 'live action pill' to represent the thinking process of GPT-5. From our understanding, GPT-5 labels the live action pill itself (because the chat is free-form) while LinkQ's live action pill is labeled in a structured way based on which stage of the workflow it is in.

5 Qualitative Study

We conducted a think-aloud study [23] focusing on practitioners' analysis workflows, exploration strategies, and overall confidence in an agent when using LinkQ. The study design was largely based on recent evaluations for LLM data analysis systems [11, 33, 35].

We recruited participants from several departments across two organizations. All 14 participants conduct data analysis on a daily basis and had at least some experience with either LLMs or KGs (see Table 1). The study consisted of exploring and answering both targeted and exploratory questions about two different knowledge graphs, Wikidata [37] and BRON (a cybersecurity KG [14]).

Experience	with KGs	with LLMs
1 (None)	0 / 14	0 / 14
2 (Slight)	4 / 14	0 / 14
3 (Some)	2 / 14	3 / 14
4 (Moderate)	3 / 14	0 / 14
5 (Extreme)	5 / 14	11 / 14

Table 1: A breakdown of our participants' experience with KGs and LLMs from our qualitative study. Experience was self-reported on a Likert scale of 1-5, where each level was described in terms of how often or thoroughly they developed, trained, interacted with, or queried KGs and LLMs in their work.

For this workshop paper, we discuss the overall feedback we received about the Agentic Trace Display and remark on how participants' workflows and confidence was impacted by the agent itself. More details about the study protocol is available in our Appendix A.

5.1 Visualization Usage

Feedback: The state diagram (Figure 2) was by far the participants' preferred visual component in LinkQ, with participants remarking that it greatly contributed to their understanding and trust in the agent's workflow. Participants frequently mentioned that the state diagram "shed light" on the agent's process, making its operations more transparent and comprehensible: *"I really like the state diagram while it's running because it's giving so much insight into what's going on. It makes the LLM workflow more explainable."* This visual aid, paired with the live action pill, not only made the process more engaging but also provided users with a clearer sense of the workflow, allowing them to anticipate potential issues and debug more effectively. Another participant noted, *"One, it's a lot more engaging than waiting for the query, and two, it's helping me understand what's actually happening in the process."* Overall, nearly every participant told us they would prefer if all LLM-assisted tools incorporated an 'agentic trace display' to better illuminate an agent's behavior.

Areas for Improvement: Some participants suggested that the level of detail in the state diagram could be reduced over time as users become more familiar with the system. One participant commented, *"Once you get familiar with the system, then an abbreviated version of the state diagram might be enough. You won't need to see every single state."* Agentic AI systems can seek to balance detail and simplicity in order to maintain the effectiveness of their agent-based state diagrams. For example, for expert users, the live action pill could remain visible for

[1] https://www.anthropic.com/engineering/building-effective-agents

[2] https://help.openai.com/en/articles/6825453-chatgpt-release-notes

Fig. 2: The *State Diagram* (described in Section 4) highlights the high-level workflow of the agent as well as the specific sub-stage (*Question Refinemnt* → *User asks question*). The other stages that are not active have reduced size and brightness.

Fig. 3: The *State Diagram* collapsed (using the top right dropdown), with only primary steps in the agent's workflow on display and sub-stages hidden. Each state is highlighted as the agent works through its pipeline.

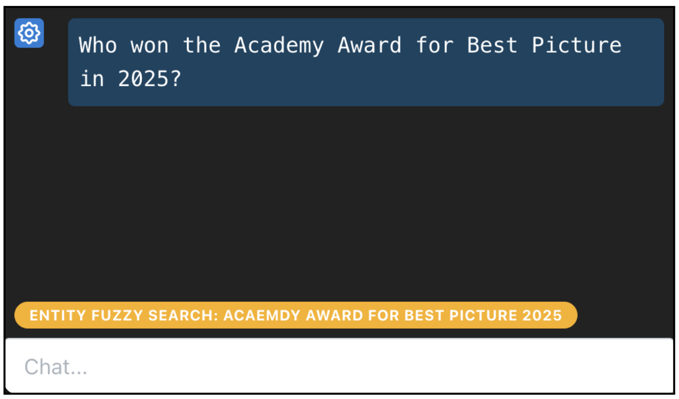

Fig. 4: While the user waits for the LLM to generate a query, the chat panel shows a *Live Action Pill* that explains that the LLM is fuzzy searching for the entity *Academy Award for Best Picture 2025*.

small-scale updates on the agent's workflow, while a state diagram can be hidden or displayed by choice. An example is shown in Figure 3.

Future Opportunities: Another interesting direction discussed by participants was the ability to 'edit' or 'restart' the agent's workflow by interacting with the state diagram itself. For example, if users found that the agent did not resolve an entity correctly, they wanted to click on the *"LLM fuzzy searches for entity"* state, and begin a dialogue either visually or through the chat panel to help correct the agent's misstep. Participants were critical about LLM-based systems often having an agent restart its workflows from scratch when given user feedback, rather than picking up from a specific point in time. Interactive trace visualizations could incorporate this type of timestamped agent-editing in the future.

5.2 Potential for Overtrust

At times during our study, asking a question in LinkQ would result in an incorrect answer (e.g., because of an incorrectly written query or wrong entity resolution). We found participants sometimes *overtrusted* the agent's incorrect answer and would use the supporting visualizations to rationalize reasons why the agent might be right. For example, when one participant was searching for a list of movies that won Best Picture, the agent generated a query that incorrectly flagged *Gladiator* as a winner in 2023. The participant remarked: *"It says Gladiator won best picture in 2023, but isn't that movie old? I know a second one came out recently, so maybe that's when it won?"*

Even when performing question-answering over domain-specific data that participants were familiar with (i.e. with the BRON cybersecurity KG), participants still justified incorrect outputs from LinkQ that they were uncertain about. For example, when retrieving vulnerabilities associated with Google Chrome, LinkQ (incorrectly) returned no results. One participant told us, *"I think that makes sense, since Google Chrome probably has a lot of support, it probably doesn't have any vulnerabilities."* We believe that the added visual transparency, largely in part through the agentic trace display, may have resulted in overconfidence in the agent's outputs. The insights into what the LLM is doing at every step tended to sway participants into believing the agent was always accurately performing its tasks.

Implications: Our findings align with recent work that shows how an LLM's tendency to "smooth over" inconsistencies or outright lie can result in overreliance on AI [5,13,16,25,31]. *Transparently-designed* visualizations, while useful for revealing LLM reasoning and agentic behavior, may inadvertently exacerbate overtrust by reducing a user's scrutiny of AI outputs – as suggested (unintentionally) by one participant: *"I really liked that the interface was flexible enough that someone could accept the information at face value."*

Similar to current research in explainable AI [3,32], it is likely that visualization designs for agentic AI workflows will need to shift towards *challenging* users' assumptions around AI outputs—rather than purely clarifying system behavior—especially in this new era of LLM-based data analysis tools. Potential strategies could include visualizations that explicitly highlight where uncertainty exists in the system workflow (which does not exist in our State Diagram), as well as proactively suggesting contradictory or alternative interpretations via in-line visualizations. Work in trustworthy visualization design [12,22] can help developers understand when trust might be falsely gained through paired visualization and LLM design.

6 Limitations & Future Work

While our Agentic Trace Display focused on simple, widely recognizable visualizations to isolate their effects on user perception, future work could build on the findings in this paper by exploring more novel or expressive ways of visualizing agentic behavior. For example, future designs might incorporate richer temporal traces, uncertainty cues, history of state transitions, or interactive debugging elements.

Our think-aloud study involved 14 participants belonging to different departments within two organizations. Our participants are highly educated and experienced in data science, and displayed inquisitive behaviors when interacting with an agentic AI system. All participants had at least *some* to very high LLM experience, suggesting that our findings may not generalize to "every day users" who are unfamiliar with LLMs or trustworthy design. Consequently, future studies can extend the research presented in this paper to examine how trust is impacted in users who are *not* inherently skeptical of AI, and consequently may not make the effort to question LLM-based systems.

7 Conclusion

This workshop paper presented an Agentic Trace Display, composed of a paired State Diagram and Live Action Pill to improve transparency in agentic AI workflows. Integrated into LinkQ, a natural language interface for knowledge graph querying, these visualizations gave users real-time insight into the agent's internal reasoning and execution pipeline.

Our think-aloud study revealed that participants found the trace visualizations highly useful for forming accurate mental models of the agent's behavior, which in turn fostered confidence in system outputs. However, we also observed instances where the same visual cues contributed to overtrust, particularly when the agent produced incorrect results. These findings point to a potential design tension in agentic systems: visualizations that clarify system behavior may also unintentionally reinforce user belief in incorrect outputs.

Going forward, we encourage designers of LLM-based systems to incorporate trace-based visualizations, but also promote critical reflection through trust-calibrating mechanisms. As agentic AI continues to advance, visualization will play a critical role in balancing usability with user skepticism.

Acknowledgments

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001 or FA8702-25-D-B002. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

This work was authored in part by the National Renewable Energy Laboratory (NREL), operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

A Appendix

A.1 Study Design

Protocol: Participants were recruited from a professional network of practitioners who currently or have previously worked with KGs and LLMs. Each evaluation lasted between 1–1.5 hours and were conducted either in-person or virtually using a video conferencing tool. All participants consented to the study prior to its start. Two authors conducted a qualitative, think-aloud study that consisted of questions that had pre-determined correct or incorrect answers, as well as open-ended questions that had no correct or incorrect answer. Finally, participants were asked a series of follow-up questions to gather their feedback on the tool. The resultant qualitative data was analyzed with a focus on emergent themes [7] related to our original research goals.

Participants: We followed a saturation-based approach [10] in which participants were recruited until no new themes emerged during data collection. Our evaluation consisted of 14 participants who have a BS (3/14), an MS (7/14) or a PhD (4/14). Participants ranked their experience for KGs and LLMs on a Likert scale of 1 (none) to 5 (extreme), shown in Table 1. All of the participants in our study work in AI, ML, and data science with a focus in the cybersecurity domain.

Knowledge Graphs: Participants first used LinkQ with the Wikidata KG [37] and were tasked with finding answers for randomly sampled questions from the the Mintaka question bank [30]. Next they used the BRON cybersecurity KG [14] to answer questions about cybersecurity, for example *"Find critical vulnerabilities for Google Chrome in 2024"*. We selected Wikidata KG because of its domain-agnostic, general-purpose use that all participants were familiar with. In contrast, we chose the BRON KG because of its domain-specificity in cybersecurity. While all of our participants are knowledgeable in cybersecurity, only 3/14 had used BRON previously.

References

[1] J. Alexander, P. Nanda, K.-C. Yang, and A. Sarvghad. Can gpt-4 models detect misleading visualizations? In *2024 IEEE Visualization and Visual Analytics (VIS)*, pp. 106–110. IEEE, 2024. 2

[2] B. AlKhamissi, M. Li, A. Celikyilmaz, M. Diab, and M. Ghazvininejad. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*, 2022. doi: 10.48550/arXiv.2204.06031 2

[3] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, article no. 81, 16 pages. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3411764.3445717 3

[4] R. C. Basole and T. Major. Generative ai for visualization: Opportunities and challenges. *IEEE Computer Graphics and Applications*, 44(2):55–64, 2024. 2

[5] Z. Buçinca, M. B. Malaya, and K. Z. Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21, 2021. 2, 3

[6] A. Crisan, N. Butters, and Zoe. Exploring subjective notions of explainability through counterfactual visualization of sentiment analysis. In *2024 IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV)*, pp. 15–24. IEEE, 2024. 2

[7] L. A. Curry, I. M. Nembhard, and E. H. Bradley. Qualitative and mixed methods provide unique contributions to outcomes research. *Circulation*, 119(10):1442–1452, 2009. 4

[8] V. Dhanoa, A. Wolter, G. M. León, H.-J. Schulz, and N. Elmqvist. Agentic visualization: Extracting agent-based design patterns from visualization systems. *arXiv preprint arXiv:2505.19101*, 2025. 2

[9] L. Ehrlinger and W. Wöß. Towards a definition of knowledge graphs. *Proc. ESWC Posters and Demos Track*, 48(1-4):2, 2016. 1

[10] G. Guest, A. Bunce, and L. Johnson. How many interviews are enough? an experiment with data saturation and variability. *Field methods*, 18(1):59–82, 2006. 4

[11] J. Ha, H. Jeon, D. Han, J. Seo, and C. Oh. Clochat: Understanding how people customize, interact, and experience personas in large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–24, 2024. 2

[12] S. Ha, S. Monadjemi, and A. Ottley. Guided by ai: Navigating trust, bias, and data exploration in ai-guided visual analytics. In *Computer Graphics Forum*, vol. 43, p. e15108. Wiley Online Library, 2024. 2, 3

[13] G. He, N. Aishwarya, and U. Gadiraju. Is conversational xai all you need? human-ai decision making with a conversational xai assistant. *arXiv preprint arXiv:2501.17546*, 2025. 2, 3

[14] E. Hemberg, J. Kelly, M. Shlapentokh-Rothman, B. Reinstadler, K. Xu, N. Rutar, and U.-M. O'Reilly. Linking threat tactics, techniques, and patterns with defensive weaknesses, vulnerabilities and affected platform configurations for cyber hunting. *arXiv preprint arXiv:2010.00533*, 2020. doi: 10.48550/arXiv.2010.00533 2, 4

[15] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann. Knowledge graphs. *ACM Comput. Surv.*, 54(4), article no. 71, 37 pages, 2021. doi: 10.1145/3447772 1

[16] S. S. Y. Kim, J. W. Vaughan, Q. V. Liao, T. Lombrozo, and O. Russakovsky. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, article no. 420, 19 pages. Association for Computing Machinery, New York, NY, USA, 2025. doi: 10.1145/3706598.3714020 3

[17] H. Li, G. Appleby, C. D. Brumar, R. Chang, and A. Suh. Knowledge graphs in practice: Characterizing their users, challenges, and visualization opportunities. *IEEE Trans. Vis. Comput. Graph.*, 30(1):584–594, 2024. doi: 10.1109/TVCG.2023.3326904 1

[18] H. Li, G. Appleby, and A. Suh. Linkq: An llm-assisted visual interface for knowledge graph question-answering. In *2024 IEEE Visualization and Visual Analytics (VIS)*, pp. 116–120, 2024. doi: 10.1109/VIS55277.2024.00031 1

[19] H. Li, G. Appleby, and A. Suh. A preliminary roadmap for llms as assistants in exploring, analyzing, and visualizing knowledge graphs. In *IEEE VIS NLVIZ Workshop: Exploring Research Opportunities for Natural*

Language, Text, and Data Visualization, 2024. 2

[20] M. Lissandrini, D. Mottin, K. Hose, and T. B. Pedersen. Knowledge graph exploration systems: are we lost? In *CIDR*, vol. 22, pp. 10–13, 2022. 1

[21] Y. Liu, S. K. Lo, Q. Lu, L. Zhu, D. Zhao, X. Xu, S. Harrer, and J. Whittle. Agent design pattern catalogue: A collection of architectural patterns for foundation model based agents. *Journal of Systems and Software*, 220:112278, 2025. 2

[22] O. McKinley, S. Pandey, and A. Ottley. Trustworthy by design: The viewer's perspective on trust in data visualization. *arXiv preprint arXiv:2503.10892*, 2025. 3

[23] C. North. Toward measuring visualization insight. *IEEE computer graphics and applications*, 26(3):6–9, 2006. 2

[24] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Trans. Knowl. Data Eng.*, pp. 1–20, 2024. doi: 10.1109/TKDE.2024.3352100 2

[25] P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 2024. 3

[26] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller. Language models as knowledge bases? In *Proc. EMNLP/IJCNLP*, pp. 2463–2473. ACL, Hong Kong, 2019. doi: 10.18653/v1/D19-1250 2

[27] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *In Proc. International Conf on Intelligence Analysis*, vol. 5, pp. 2–4. McLean, VA, USA, 2005. 2

[28] V. Rawte, A. Sheth, and A. Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023. 2

[29] R. Sapkota, K. I. Roumeliotis, and M. Karkee. Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges. *arXiv preprint arXiv:2505.10468*, 2025. 2

[30] P. Sen, A. F. Aji, and A. Saffari. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. *arXiv preprint arXiv:2210.01613*, 2022. 4

[31] R. Serafini, A. Yardim, and A. Naiakshina. Exploring the impact of intervention methods on developers' security behavior in a manipulated chatgpt study. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, article no. 520, 26 pages. Association for Computing Machinery, New York, NY, USA, 2025. doi: 10.1145/3706598.3713989 3

[32] A. Suh, K. Alperin, H. Li, and S. R. Gomez. Don't just translate, agitate: Using large language models as devil's advocates for ai explanations. In *ACM CHI 2025 Workshop on Human-Centered Explainable AI (HCXAI)*, 2025. 3

[33] N. Sultanum and A. Srinivasan. Datatales: Investigating the use of large language models for authoring data-driven articles. In *Proc. VIS*, pp. 231–235, 2023. 2

[34] Y. Tian, W. Cui, D. Deng, X. Yi, Y. Yang, H. Zhang, and Y. Wu. Chartgpt: Leveraging llms to generate charts from abstract natural language. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[35] Z. Wang, Y. Huang, D. Song, L. Ma, and T. Zhang. Promptcharm: Text-to-image generation through multi-modal prompting and refinement. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–21, 2024. 2

[36] H. Wei. Combining knowledge graphs, quickly and accurately. *Amazon Science*, 2020. 1

[37] Wikipedia. Wikidata Statistics. `https://www.wikidata.org/wiki/Wikidata:Statistics`. Accessed: 2025-03-18. 2, 4