Deep Research in the Era of Agentic AI: Requirements and Limitations for Scholarly Research

Mohamad Yaser Jaradeh[1,2,*], Sören Auer[2,1]

[1]L3S Research Center, Leibniz University Hannover, Hanover, Germany
[2]TIB - Leibniz Information Centre for Science and Technology, Hanover, Germany

Abstract

In the fast-evolving era of agentic AI, Large Language Models (LLMs) from major providers and open-source alternatives offer unprecedented capabilities for "deep search", enabling complex, iterative information retrieval and synthesis crucial for academic endeavors. However, their application in scientific research and paper writing necessitates strict requirements and a critical awareness of inherent limitations, including the risks of unreviewed content, temporal biases, and access barriers such as paywalls. This vision paper discusses a list of requirements that a scientific deep research system should have to become a viable candidate (i.e., to become a valuable system for researchers). As well as a list of limitations that are observed from current systems (industry-grade and community-developed). We also outline a path forward for harnessing agentic AI in scientific discovery and scholarly communication.

Keywords

Agentic AI, Deep (Re)Search, Information Asymmetry, Unreviewed Content Risks

1. Introduction

Recent advances in Large Language Models (LLMs) have shifted the focus of Artificial Intelligence (AI) research from static prediction to **agentic autonomy**. In agentic systems, an LLM iterates between natural-language reasoning with external actions—taking API calls, executing code, and retrieving new documents, while iteratively revising its action plan. The ReAct [1] framework first formalized this plan-act-reflect loop, demonstrating that explicit reasoning traces boost performance on multi-step tasks and provide hooks for tool integration.

Commercial providers have rapidly productized these ideas. OpenAI's Assistants and function-calling APIs, built on GPT-4 and its multimodal successor GPT-4o, expose browsing, code-execution, and memory primitives that let developers compose end-to-end research agents [2]. Google's Gemini family extends the paradigm to images, audio, and video while supporting one million token contexts.

We term these capabilities **agentic deep search**: long-horizon research workflows in which an autonomous agent formulates queries, harvests evidence from heterogeneous APIs, evaluates source quality, and synthesizes structured outputs. By shrinking days of literature search or patent mining into hours, agentic deep search promises to accelerate discovery across disciplines.

Yet raw capability does not guarantee full reliability. LLMs embed a dated snapshot of the world; factual accuracy diminishes (or completely lacks) for events or findings that post-date their last training cut-off. Even frontier models' system cards (i.e., a document that provides transparent, standardized information about the model's characteristics, capabilities, limitations) cautions that alignment safeguards remain a "work in progress", with residual risks of hallucination and persuasive misinformation [3]. Autonomy introduces further hazards: an unchecked agent may cite unreviewed blogs, omit paywalled studies, or leak sensitive data while invoking external tools. Without rigorous provenance logs, researchers cannot audit how conclusions were reached – an unacceptable opacity in scholarly contexts.

5th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment (Sci-K), November 2-6, 2025, Nara, Japan
*Corresponding author.
jaradeh@l3s.de (M. Y. Jaradeh); auer@tib.eu (S. Auer)
0000-0001-8777-2780 (M. Y. Jaradeh); 0000-0002-0698-2864 (S. Auer)

This vision paper argues that agentic AI can become a trustworthy partner in scientific inquiry only if its design satisfies research-specific requirements. We explore our vision with two core components:

1. **Requirements:** We formulate eight criteria—verifiability, bias mitigation, temporal awareness, selective memory, and more—that distinguish research-grade agents from consumer chatbots.
2. **Limitations:** We catalog current failure modes, including unvetted inputs, concept drift, paywall bias, hallucination, context bottlenecks, privacy constraints, and operational costs.

By categorizing these principles and clarifying them, we aim to guide developers, funders, and research communities toward agentic AI that is trustworthy and to be used by scientists and researchers to help them perform complex tasks.

2. Background: Agentic AI and Deep Search Capabilities

Agentic AI refers to autonomous systems designed to pursue complex goals with minimal human intervention, demonstrating adaptability, advanced decision-making, and self-sufficiency in evolving environments [4]. They couple a LLM's internal chain-of-thought reasoning with external actions in a feedback loop of plan $\rightarrow$ act $\rightarrow$ reflect. Paradigms such as ReAct interleave natural-language reasoning traces with tool calls, letting the model update its plan as new evidence arrives [1]. More generally, chain-of-thought prompting shows that exposing intermediate reasoning steps markedly boosts accuracy on complex tasks, establishing an essential primitive for autonomy [5].

Major providers now ship APIs that turn foundation models into multi-modal research agents. OpenAI's GPT-4/4o [2] family supports function-calling, web browsing, and code execution, while Anthropic's Claude 3/4 series [6] advertises sustained "computer use" workflows and long context reasoning suitable for document ranking and assessments. Academic and open-source efforts, e.g., Toolformer, where the model teaches itself when and how to call external APIs [7], further illustrate this trend toward tightly integrated tool use.

With browsing and API hooks, LLM agents can conduct multi-step literature and data dives that previously required human curators:

- **Web-scale retrieval & citation**. For instance, WebGPT [8] fine-tunes GPT-3 to navigate the web, gather evidence, and output answers with inline references, outperforming human answers (collected from Reddit) in blind tests.
- **Systematic literature reviews**. Products like Elicit [9] use LLMs to rank and summarize hundreds of abstracts for evidence synthesis, while open-source LLAssist [10] automates key parts of PRISMA-style[1] reviews.
- **Domain-specific mining**. LLM agents already appear in patent invalidity or prior-art searches, delivering faster recall of obscure filings [11], and in bibliometric trend analysis pipelines that classify thousands of papers to surface emerging topics [12].

Collectively, these examples show that agentic AI search approaches can ingest and process 10^2–10^3 raw sources, draft structured summaries, and iterate as new information arrives.

However, scientific inquiry sets a higher bar than consumer question answering. Researchers need transparent provenance, reproducibility, domain-aware ranking, and the ability to handle uncertainty - requirements [13]. The next sections highlight where the current systems fall short and outline a list of what is needed for trustworthy Deep (Re-)search in academia.

3. Requirements for Applying Agentic AI Search in Research

Casual question and answering agents can afford occasional gaps or fuzzy information provenance; scholarly workflows cannot. To be admissible in a paper, every claim, dataset, and result produced

[1] https://www.prisma-statement.org/

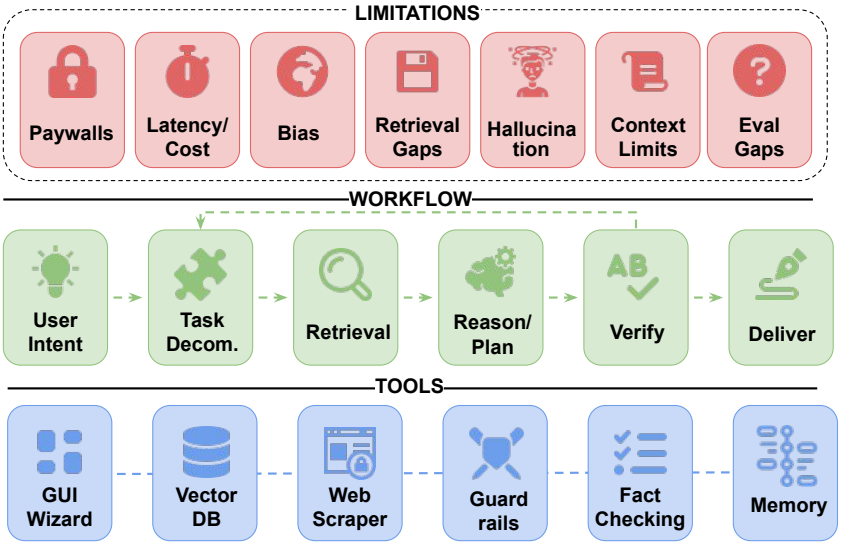

Figure 1: Vision Overview: Listing limitations, tools, and agentic workflow for Deep Research

by an LLM agent must withstand peer review and re-analysis for years to come. This section distills eight core requirements that are required to turn an autonomous search agent into a reliable Deep (Re-)search companion tool.

Verifiability and Traceability. An agent must emit a machine-readable "audit trail" that links each generated statement to (i) the exact retrieval query, (ii) stable identifiers (e.g., DOI, PubMed ID, patent number), and (iii) the intermediate reasoning step that justified inclusion. Frameworks such as TRiSM [14] (Trust, Risk & Security Management) propose logging every tool call and chain-of-thought token for post-hoc inspection, while LLMAuditor [15] demonstrates how a second LLM plus human spot-checks can scalably detect citation drift or hallucination.

Ethical and Bias Mitigation. Agentic search must actively counter selection biases that skew evidence toward English-language, high-impact venues or majority demographics. Gallegos et al. [16] survey on bias in LLMs' catalogs mitigation approaches, from counterfactual prompts to calibrated relevance scoring, that should be embedded as first-class modules when using agentic AI search. Recent biomedical screening studies confirm that bias-aware prompts recover ∼12% more non-English RCTs compared with naïve keyword filters [17].

Scalable Breadth with Human Depth. Hybrid pipelines such as LatteReview [18] or GPT-assisted screening prototypes [19] show that LLM agents can filter through tens of thousands of abstracts, yet final selection, annotation, and interpretation still require expert researcher review. Designing checkpoints where humans judge borderline cases keeps false positives below systematic-review thresholds without sacrificing the speed and comfort of using agentic AI search techniques.

Seamless Workflow Integration. To fit existing toolchains, agents should use bibliographic as well as natural language. Feeding candidates straight into Zotero, exporting RIS/BibTeX, querying the arXiv and Crossref APIs, and respecting journal impact-factor filters is essential. Tutorials on GPT-Zotero bridges and OpenAI's plugin/GPT-store model illustrate how modest wrappers can embed deep-search capabilities inside everyday reference managers used by researchers [20, 21].

Temporal Awareness & Concept-Drift Control. Scientific consensus evolves; an agent that combines 2012 and 2025 findings can produce misleading results. Temporal-drift frameworks such as

Table 1
Requirements Overview: Key-Mechanisms and Example Implementations

Requirement	Key Mechanism	Example Implementation
Verifiability	Immutable audit logs	TRiSM + LLMAuditor
Bias control	Fairness metrics & counterfactual prompts	Bias-aware abstract screening
Human oversight	Checkpoints & adjudication queues	LatteReview hybrid loop
Workflow fit	Bibliographic APIs & plugin hooks	Zotero-GPT bridge
Temporal rigor	Drift-aware ranking & dependency checks	Zilean + Byam
Narrative scope	Declarative in-/out-scope schemas	Prompt Engineering for RE
Source quality	Journal/conference allow-lists	Impact-factor filter
Memory model	Detachable episodic stores	Safe episodic memory design

Zilean [22] and CORAL [23] adapt retrieval and ranking to shifting evidence bases, while code-centric studies like Byam [24] show LLMs repairing projects broken by outdated dependencies. Conversely, using LLMs for coding highlights the effects of temporal drift, where models use old and conflicting packages to write code, unaware of newer releases or security vulnerabilities.

Researcher-Controlled Narrative Scope. Deep (re-)searches must be bounded. Prompt-engineering studies in requirements engineering propose declarative "in-scope/out-of-scope" schemas that agents use to accept or reject documents, ensuring that a cancer-biology review does not drift into plant genomics unless explicitly instructed by the researcher [25]. Specification work on multi-agent LLM systems further argues that formal task definitions are a prerequisite for reproducibility and safety.

Peer-Reviewed-Only Source Constraints. Agents need configurable whitelists (e.g., journals ranked $\geq$ Q1, A^* conferences, or datasets with open licenses) and should fail fast if no qualifying evidence exists. Title and abstract screening benchmarks show that adding an impact-factor filter and peer-review flag cuts noise by 35% without hurting recall when compared to baseline semantic search [17].

Selective (Long-Term vs Short-Term) Memory. Researchers accumulate domain knowledge (long-term) but isolate project-specific facts (short-term). Emerging work on episodic memory [26] for AI agents recommends detachable, human-editable memory slots. Persistent memories store lab standards or preferred datasets. Temporary slots hold the current paper set and disappear after publication. This separation supports both cumulative expertise and clean-slate reproducibility.

Meeting these eight requirements turns an LLM-based search agent from a helpful assistant into a research-grade collaborator. Table 1 shows a condensed overview of the requirements discussed above.

4. Key Limitations of Agentic AI in Deep Research

Scientific rigor requires a high level of traceability and trustworthiness for agentic AI search agents to be used on a large scale. We now present a taxonomy of the most limiting factors that still separate today's autonomous "deep-search" stacks from reliable academic workflows.

Unvetted or Non-Peer-Reviewed Inputs. Browser-enabled agents easily scrape blogs, press releases, or social-media threads that have not undergone editorial or peer scrutiny. A recent survey of LLM hallucinations notes that one-third of observed confabulations could be traced to unreviewed web pages being pulled into the context window without provenance checks [27]. In systematic-review works, up to 18% of citations surfaced by naive agents pointed to pre-print servers or personal websites rather than journals indexed in Web of Science or MEDLINE – forcing experts to step in to filter the

noise [28]. As a mitigation, agents must default to curated APIs (PubMed, Crossref, Dimensions) or enforce whitelists such as Q1 journals and A* conferences before loosening to grey literature.

Temporal Staleness and Concept Drift. LLMs freeze a snapshot of the world at their last training cut-off; GPT-4's base weights, for example, contain little post-2023 knowledge. Studies probing "effective cut-off" dates reveal uneven freshness across domains, e.g., biomedicine lags by over a year compared with computer science within the same model [29]. Follow-up work shows performance drops of 15-20% on factoid questions whose answers shifted after the model's cut-off [30].

Paywalls, Licensing, and Institutional Access. Because autonomous agents have no legal right to bypass subscription barriers, deep searches systematically over-sample open-access content. Bibliometric analyses of Wikipedia citations confirm a 44% OA bias even before agentic filtering. For instance some environmental-sciences reviews note that paywalled articles are routinely omitted in global syntheses, especially for low- and middle-income country authors and institutes [31, 32]. These numbers are skewed whenever key findings reside behind JSTOR, IEEE-Xplore, or Elsevier access walls.

Mitigation: Workflow engineers must pair agents with institutional proxy resolvers, maintain error-out policies when access fails, or flag "coverage gaps" for manual moderation.

Hallucination and Mis-synthesis. Even when sources are valid, language models can fabricate links or misattribute findings. Controlled evaluations detect high hallucination rates in retrieval-augmented settings and higher in zero-retrieval modes when using specific models. Mitigation methods like entropy-based uncertainty estimators catch only a subset of these errors [33]. In climate-change case studies, agents have promoted retracted pre-prints as definitive evidence, illustrating how unreviewed claims can be converted into citable facts in research papers.

Context-Length and Multimodal Bottlenecks. Scientific artifacts are long: a systematic review's appendix can exceed 100k tokens, and figures embed essential data. Yet performance degrades as context length stretches; recent work shows sigmoid-like decay curves once input exceeds 32k tokens [34]. Attempts to push to 128k tokens (e.g., LongRoPE2 [35]) restore accuracy only with specialized fine-tuning that few commercial APIs expose. Meanwhile, extracting tables, charts, and images still requires separate vision pipelines; although LLM-based "ChatExtract" tools achieve promising figure-level accuracy, they remain brittle on complex multi-panel plots [36].

Scalability, Cost, and API Fragmentation. Token fees, rate limits, and heterogeneous API formats limit most end-to-end automation efforts. PubMed and arXiv are freely searchable, but Web of Science and Scopus impose subscription APIs. Many domain-specific repositories (e.g., ICSD for materials science) have no programmatic interfaces for automated access. Even when APIs exist, processing millions of abstracts incurs high costs on popular cloud endpoints, constraining reproducibility for the general public.

Privacy and Data Protection. When research involves sensitive human data (e.g., genomic sequences, patient notes, or even private datasets from internal resources), routing data through third-party LLM endpoints can breach GDPR or institutional data privacy regulations. Surveys on LLM-agent security enumerate data-exfiltration vectors, from prompt-leaks to malicious tool-wrapper code, underscoring the need for on-premise or privacy-enhanced deployments [37].

Given these risks, credible Deep (Re-)search pipelines must curtail full autonomy. Recommended safeguards include:

- **Whitelist enforcement** for peer-reviewed venues before fallback to gray literature.

- **Recency checks** that mark claims older than a configurable horizon or outside the model's knowledge window.
- **Access-gap alerts** when paywalled or API-inaccessible content prevents coverage parity.
- **Auto-generated audit trails** plus mandatory human veto on any citation lacking stable identifiers.
- **Resource budgets** that track cumulative emissions and token spend, prompting the researcher to justify large-scale runs.

By incorporating these considerations into agent orchestration layers, we can take advantage of LLM efficiency while respecting the integrity and sustainability required of modern academia (cf. Figure 1).

5. Conclusion

Agentic AI has vaulted from laboratory curiosity to production-ready real-world application, giving researchers unprecedented power to orchestrate deep searches that span thousands of papers, code bases, and datasets in minutes. Yet this paper has shown that transforming raw capability into research-grade reliability demands an explicit contract: verifiable provenance, bias controls, temporal awareness, paywall navigation, privacy safeguards, and human-in-the-loop checkpoints. Without these guardrails, the very speed and scale that make LLM agents attractive in the first place can increase misinformation and negatively impact reproducibility.

Our forward-looking vision is a "research-grade agent" framework that layers domain-specific enhancements on top of existing approaches and implementations:

- **Built-in quality filters:** peer-review flags, journal-rank allow-lists, and recency cut-offs—applied before content enters the context window.
- **Hybrid retrieval architectures:** that sequence open-access mirrors, institutional proxy resolvers, and licensed APIs, ensuring comprehensive coverage while respecting copyright and privacy constraints.
- **Episodic memory partitions:** that separate durable disciplinary knowledge from project-specific scratchpads, enabling cumulative expertise without affecting reproducibility.
- **Standardized reliability benchmarks:** that score agents on citation accuracy, coverage under paywalls, and resistance to concept drift.
- **Policy hooks:** requiring disclosure of AI assistance in grant proposals, systematic reviews, and journal submissions, creating incentives for transparent and auditable workflows.

With the current advancements of LLMs we stand at a fork in the road. One path turns LLMs' autonomy into an academic enhancer, reducing months of literature search into hours while still elevating the rigor of the scientific method. The other path, if unchecked, would broadcast misinformation at machine speed.

Acknowledgments

We thank our colleague Allard Oelen for his valuable comments in reviewing this paper.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini in order to: Paraphrase and reword, Improve writing style, and Grammar and spelling check. As well as Deep Research to find more related work. After using these service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, ReAct: Synergizing reasoning and acting in language models, in: International Conference on Learning Representations (ICLR), 2023.

[2] GPT-4o System Card — openai.com, https://openai.com/index/gpt-4o-system-card/, 2024. [Accessed 22-07-2025].

[3] K. Robison, OpenAI says its latest GPT-4o model is 'medium' risk — theverge.com, https://www.theverge.com/2024/8/8/24216193/openai-safety-assessment-gpt-4o, 2024. [Accessed 22-07-2025].

[4] D. B. Acharya, K. Kuppan, B. Divya, Agentic ai: Autonomous intelligence for complex goals—a comprehensive survey, IEEE Access 13 (2025) 18912–18936. doi:10.1109/ACCESS.2025.3532853.

[5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.

[6] Introducing Claude 4 — anthropic.com, https://www.anthropic.com/news/claude-4, 2025. [Accessed 22-07-2025].

[7] T. Schick, J. Dwivedi-Yu, R. Dessí, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: language models can teach themselves to use tools, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[8] R. Nakano, J. Hilton, S. Balaji, J. Wu, O. Long, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, X. Jiang, K. Cobbe, T. Eloundou, G. Krueger, K. Button, M. Knight, B. Chess, J. Schulman, Webgpt: Browser-assisted question-answering with human feedback, ArXiv abs/2112.09332 (2021). URL: https://api.semanticscholar.org/CorpusID:245329531.

[9] J. Kung, Elicit (product review), J. Can. Health Libr. Assoc. 44 (2023).

[10] C. Y. Haryanto, Llassist: Simple tools for automating literature review using large language models, 2024. URL: https://arxiv.org/abs/2407.13993. arXiv:2407.13993.

[11] H. Dadri, A New Era for Efficient Patent Invalidity Searches - XLSCOUT — xlscout.ai, https://xlscout.ai/a-new-era-for-efficient-patent-invalidity-searches-with-llms, 2025. [Accessed 22-07-2025].

[12] H. Raja, A. Munawar, N. Mylonas, M. Delsoz, Y. Madadi, M. Elahi, A. Hassan, H. Abu Serhan, O. Inam, L. Hernandez, H. Chen, S. Tran, W. Munir, A. Abd-Alrazaq, S. Yousefi, Automated category and trend analysis of scientific articles on ophthalmology using large language models: Development and usability study, JMIR Form Res 8 (2024) e52462.

[13] J. de la Torre-López, A. Ramírez, J. R. Romero, Artificial intelligence to automate the systematic review of scientific literature, Computing 105 (2023) 2171–2194. URL: https://doi.org/10.1007/s00607-023-01181-x. doi:10.1007/s00607-023-01181-x.

[14] S. Raza, R. Sapkota, M. Karkee, C. Emmanouilidis, Trism for agentic ai: A review of trust, risk, and security management in llm-based agentic multi-agent systems, 2025. URL: https://arxiv.org/abs/2506.04133. arXiv:2506.04133.

[15] M. Amirizaniani, J. Yao, A. Lavergne, E. S. Okada, A. Chadha, T. Roosta, C. Shah, Llmauditor: A framework for auditing large language models using human-in-the-loop, 2024. URL: https://arxiv.org/abs/2402.09346. arXiv:2402.09346.

[16] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N. K. Ahmed, Bias and fairness in large language models: A survey, Computational Linguistics 50 (2024) 1097–1179. URL: https://aclanthology.org/2024.cl-3.8/. doi:10.1162/coli_a_00524.

[17] F. Dennstädt, J. Zink, P. M. Putora, J. Hastings, N. Cihoric, Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain, Systematic Reviews 13 (2024) 158. URL: https://doi.org/10.1186/s13643-024-02575-4. doi:10.1186/s13643-024-02575-4.

[18] P. Rouzrokh, M. Shariatnia, Lattereview: A multi-agent framework for systematic re-

view automation using large language models, 2025. URL: https://arxiv.org/abs/2501.05468. arXiv:2501.05468.

[19] C. Galli, A. V. Gavrilova, E. Calciolari, Large language models in systematic review screening: Opportunities, challenges, and methodological considerations, Information 16 (2025). URL: https://www.mdpi.com/2078-2489/16/5/378. doi:10.3390/info16050378.

[20] Zotero GPT Integration: No Coding — anara.com, https://anara.com/blog/zotero-gpt-integration, 2024. [Accessed 22-07-2025].

[21] ChatGPT plugins — openai.com, https://openai.com/index/chatgpt-plugins, 2024. [Accessed 22-07-2025].

[22] Z. Deng, Q. Feng, B. Lin, G. G. Yen, Zilean: A modularized framework for large-scale temporal concept drift type classification, Information Sciences 712 (2025) 122134. URL: https://www.sciencedirect.com/science/article/pii/S002002552500266X. doi:https://doi.org/10.1016/j.ins.2025.122134.

[23] K. Xu, L. Chen, S. Wang, Coral: Concept drift representation learning for co-evolving time-series, 2025. URL: https://arxiv.org/abs/2501.01480. arXiv:2501.01480.

[24] F. Reyes, M. Mahmoud, F. Bono, S. Nadi, B. Baudry, M. Monperrus, Byam: Fixing breaking dependency updates with large language models, 2025. URL: https://arxiv.org/abs/2505.07522. arXiv:2505.07522.

[25] K. Huang, F. Wang, Y. Huang, C. Arora, Prompt engineering for requirements engineering: A literature review and roadmap, 2025. URL: https://arxiv.org/abs/2507.07682. arXiv:2507.07682.

[26] C. DeChant, Episodic memory in ai agents poses risks that should be studied and mitigated, 2025. URL: https://arxiv.org/abs/2501.11739. arXiv:2501.11739.

[27] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Trans. Inf. Syst. 43 (2025). URL: https://doi.org/10.1145/3703155. doi:10.1145/3703155.

[28] L. Schmidt, I. Cree, F. Campbell, WCT EVI MAP group, Digital tools to support the systematic review process: An introduction, J Eval Clin Pract 31 (2025) e70100.

[29] J. Cheng, M. Marone, O. Weller, D. Lawrie, D. Khashabi, B. V. Durme, Dated data: Tracing knowledge cutoffs in large language models, 2024. URL: https://arxiv.org/abs/2403.12958. arXiv:2403.12958.

[30] C. Zhu, N. Chen, Y. Gao, Y. Zhang, P. Tiwari, B. Wang, Is your LLM outdated? a deep look at temporal generalization, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 7433–7457. URL: https://aclanthology.org/2025.naacl-long.381/. doi:10.18653/v1/2025.naacl-long.381.

[31] K. A. Wood, L. L. Jupe, F. C. Aguiar, A. M. Collins, S. J. Davidson, W. Freeman, L. Kirkpatrick, T. Lobato-de Magalhães, E. McKinley, A. Nuno, J. F. Pagès, A. Petruzzella, D. Pritchard, J. P. Reeves, S. M. Thomaz, S. A. Thornton, H. Yamashita, J. L. Newth, A global systematic review of the cultural ecosystem services provided by wetlands, Ecosystem Services 70 (2024) 101673. URL: https://www.sciencedirect.com/science/article/pii/S2212041624000809. doi:https://doi.org/10.1016/j.ecoser.2024.101673.

[32] P. Yang, A. Shoaib, R. West, G. Colavizza, Open access improves the dissemination of science: insights from wikipedia, Scientometrics 129 (2024) 7083–7106. URL: https://doi.org/10.1007/s11192-024-05163-4. doi:10.1007/s11192-024-05163-4.

[33] S. Farquhar, J. Kossen, L. Kuhn, Y. Gal, Detecting hallucinations in large language models using semantic entropy, Nature 630 (2024) 625–630. URL: https://doi.org/10.1038/s41586-024-07421-0. doi:10.1038/s41586-024-07421-0.

[34] Z. Dong, J. Li, J. Jiang, M. Xu, W. X. Zhao, B. Wang, W. Chen, Longred: Mitigating short-text degradation of long-context large language models via restoration distillation, 2025. URL: https://arxiv.org/abs/2502.07365. arXiv:2502.07365.

[35] N. Shang, L. L. Zhang, S. Wang, G. Zhang, G. Lopez, F. Yang, W. Chen, M. Yang, LongroPE2: Near-lossless LLM context window scaling, in: Forty-second International Conference on Machine Learning, 2025. URL: https://openreview.net/forum?id=jwMjzGpzi4.

[36] M. P. Polak, D. Morgan, Extracting accurate materials data from research papers with conversational language models and prompt engineering, Nature Communications 15 (2024) 1569. URL: https://doi.org/10.1038/s41467-024-45914-8. doi:10.1038/s41467-024-45914-8.

[37] B. Yan, K. Li, M. Xu, Y. Dong, Y. Zhang, Z. Ren, X. Cheng, On protecting the data privacy of large language models (llms) and llm agents: A literature review, High-Confidence Computing 5 (2025) 100300. URL: https://www.sciencedirect.com/science/article/pii/S2667295225000042.