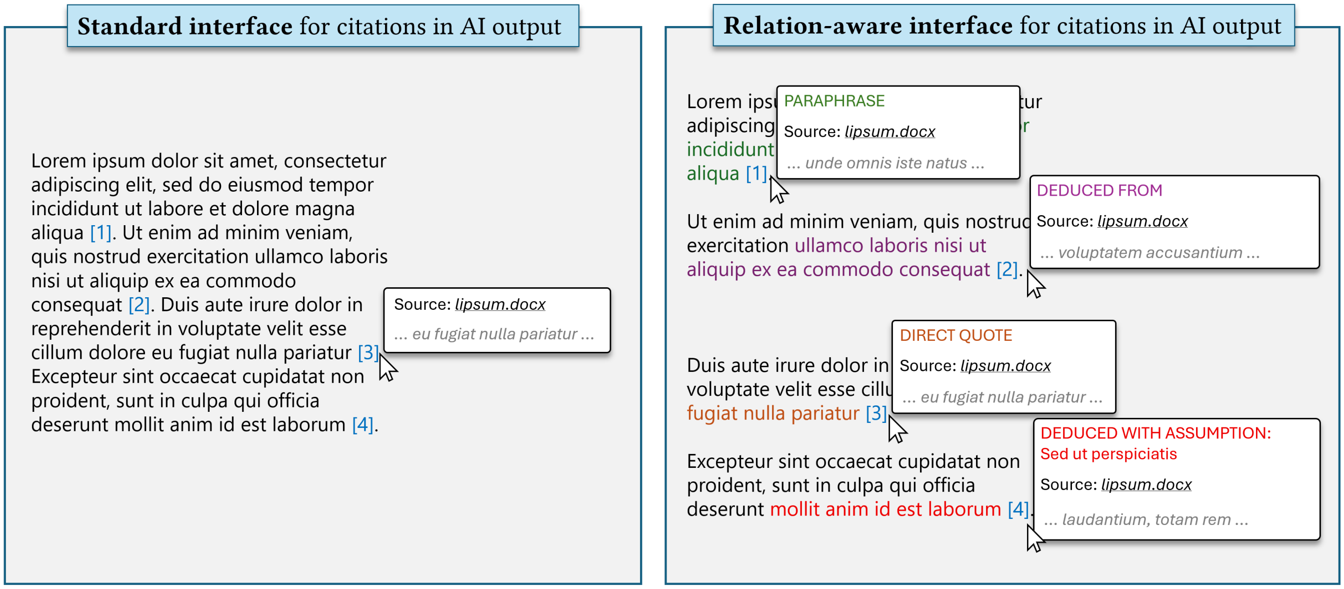
From Binary Groundedness to Support Relations: Towards a Reader-Centred Taxonomy for Comprehension of AI Output

Advait Sarkar
Microsoft Research
Cambridge, United Kingdom
University of Cambridge
Cambridge, United Kingdom
University College London
London, United Kingdom
advait@microsoft.com

Christian Poelitz
Microsoft Research
Cambridge, United Kingdom
cpoelitz@microsoft.com

Viktor Kewenig
Microsoft Research
Cambridge, United Kingdom
a-vikewenig@microsoft.com



Standard interface for citations in AI output

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua [1]. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat [2]. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur [3]. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum [4].

Source: *lipsum.docx*
... eu fugiat nulla pariatur ...

Relation-aware interface for citations in AI output

Lorem ipsu ... tur adipiscing ... r incididunt ... aliqua [1].

PARAPHRASE
Source: *lipsum.docx*
... unde omnis iste natus ...

Ut enim ad minim veniam, quis nostru exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat [2].

DEDUCED FROM
Source: *lipsum.docx*
... voluptatem accusantium ...

Duis aute irure dolor in voluptate velit esse cill fugiat nulla pariatur [3].

DIRECT QUOTE
Source: *lipsum.docx*
... eu fugiat nulla pariatur ...

Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum [4].

DEDUCED WITH ASSUMPTION: Sed ut perspiciatis
Source: *lipsum.docx*
... laudantium, totam rem ...

Figure 1: Left: standard citation-enabled responses from a language model. Right: a hypothetical interface that distinguishes between different types of syntactic manipulation (e.g. direct quote vs. paraphrase) and interpretation (e.g. induction, deduction, deduction subject to assumptions) involved in the production of language model output. We propose the development of a taxonomy of reader-centric *support relations* that would enable such interfaces, thereby leading to better critical engagement of readers with language model output and understanding of how it relates to the sources.

Abstract

Generative AI tools often answer questions using source documents, e.g., through retrieval augmented generation. Current groundedness and hallucination evaluations largely frame the relationship between an answer and its sources as binary (the answer is either supported or unsupported). However, this obscures both the syntactic moves (e.g., direct quotation vs. paraphrase) and the interpretive moves (e.g., induction vs. deduction) performed when models reformulate evidence into an answer. This limits both benchmarking and user-facing provenance interfaces.

We propose the development of a reader-centred taxonomy of grounding as a set of support relations between generated statements and source documents. We explain how this might be synthesised from prior research in linguistics and philosophy of language, and evaluated through a benchmark and human annotation protocol. Such a framework would enable interfaces that communicate not just whether a claim is grounded, but how.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI); Natural language interfaces**.

Keywords

provenance, explainability, fact verification, citation, faithfulness

ACM Reference Format:
Advait Sarkar, Christian Poelitz, and Viktor Kewenig. 2026. From Binary Groundedness to Support Relations: Towards a Reader-Centred Taxonomy for Comprehension of AI Output. In *Proceedings of Workshop on Science and Technology for Augmenting Reading at CHI 2026 (CHI 2026 STAR Workshop).*

1 Beyond Binary Hallucination and Groundedness

When an Generative AI tool responds to a query by using information from a document or from the Web, "groundedness" is often treated as a binary property: either the answer is supported by a source, or it is not. Yet even the most straightforward fact retrieval scenario shows that this framing is too blunt.

Consider a short quarterly report paragraph about a fictional company (Acme Corporation), including the sentence "Total revenue reached $847.2 million…" alongside surrounding narrative about performance and growth. Now imagine the user asks a concrete question tied to that excerpt: "What was Acme Corp's revenue in Q2 2025?". When we tested this query on four commercial systems, they produced responses that were all plausibly "grounded," but that differed materially in how they relate to the text: one repeated the exact figure verbatim ("$847.2 million"); another quoted the figure but adds immediate surrounding context "Total revenue reached $847.2 million"; another rounds the figure ("Total revenue was over $847 million"); and another paraphrases ("Revenue in Q2 2025 was US$ 847.2 million") in a way that more directly fits the query as asked, but quietly introduces an interpretive assumption (treating "revenue" as equivalent to "total revenue," which may not match the user's interpretation). All of these responses are "supported" in the simple sense, but they represent different support relations; different ways of transforming source material into an answer.

The issue becomes unavoidable for interpretive questions. Using the same report-style excerpt, a user might ask: "Is Acme doing well?" Here, an answer cannot simply point to a single span; it must decide what counts as evidence (e.g., revenue growth percentage, or qualitative phrases that might appear in the report such as "strong performance," "resilience"), compress multiple sentences into a judgment, and reframe descriptive language as an evaluative conclusion. These are interpretive moves that embed implicit criteria for "doing well" and that must necessarily go beyond what the document literally supports.

Consider another toy example. If the source text is "The cat sat on the mat" and the user asks "Did an animal sit on the mat?", a "yes" answer relies on a background assumption (a cat is an animal). If the user asks "Is the cat able to sit?" the answer is a deduction from the described event. If the user asks "What did the cat sit on?" the answer can be a minimal direct quote ("a mat") or a more complete quote ("The cat sat on the mat") or a partial paraphrase ("It sat on the mat"). These responses differ in syntactic and epistemic status: direct quotation, paraphrases, deduction, and deduction contingent on ancillary assumptions, despite all being "supported" claims. Put differently, two answers can both be "grounded" while warranting very different levels of scrutiny: one may be traceable to a verbatim span; another may be a defensible but assumption-laden inference.

Prevailing groundedness frameworks that focus on binary "supported/unsupported" checks fail to model the syntactic and interpretive reformulation that occurs when text is turned into an answer to a user query. This matters because, as readers increasingly rely on Generative AI tools to process documents and synthesise answers for them, it becomes crucial for them to understand not merely whether support is present, but what kind of support.

Readers rarely have the time, expertise, or motivation to audit every claim in an AI-generated answer; by distinguishing these support relations, we can help readers assess where scrutiny is most warranted. Prior work has argued that the greatest risk of Generative AI to knowledge work is not hallucination, but the erosion of critical thinking through passive reliance [37], with survey evidence showing that higher confidence in AI is associated with less critical thinking and a shift from information gathering to verification [21]. Binary groundedness labels may exacerbate this: if a claim is labelled "supported," readers may not interrogate *how* it is supported. More broadly, AI has been argued to shift knowledge work from material production to the critical integration of AI output [34], and the metacognitive demands this places on users (monitoring what the AI did, evaluating its output, and deciding how much to rely on it) have been identified as a key usability challenge [39]. A taxonomy of support relations can be understood as a resource for developing metacognitive tools that helps readers meet these demands. Our design intent follows the argument that AI systems should challenge and provoke critical engagement rather than merely accelerate workflows [35].

2 Towards a Taxonomy of Support Relations

Given a generated statement S and a source document (or set of documents) D, we wish to determine a support relation (or relations) $R(S, D)$ drawn from an explicitly defined inventory (e.g., direct quotation, paraphrase, deductive support, inductive support, support contingent on ancillary assumptions). How might we develop such an operational taxonomy of support relations, and how might we evaluate it?

Prior work could provide strong theoretical starting points. For instance, Toulmin's model of argumentation [18, 40] decomposes argumentative statements into claim, grounds/data, and warrant, with optional backing, qualifier, and rebuttal. Toulmin's scheme is attractive here because it renders explicit the inferential link between evidence and an assertion. A related starting point is work on support in argumentation systems [8], which surveys distinct interpretations of support (e.g., deductive support, evidential support, necessary support, backing).

Another starting point is pragmatics. Grice's account of "conversational implicature" [12] describes how interpreters attribute speaker meaning beyond what is literally said, under cooperative principles and contextual assumptions. This suggests categories for content that is not entailed by D but is nevertheless inferable in context.

Finally, scholarship on representing scholarly discourse relations provides relevant inspiration for knowledge work settings: the "ScholOnto" project [4, 5, 25, 41] , for example, aims to support scholarly interpretation by enabling researchers to represent claims and their relationships to the literature as an explicit semantic network. The "scite" metric [30] aims to contextualise scientific citations as supportive or unsupportive.

Yet, it is not sufficient for a taxonomy to be theoretically grounded. It must also be practical on two counts: it must be possible for human and machine annotators to robustly classify support relations, and these support relations must be useful and understandable by readers of LLM-generated text. This could be achieved by treating taxonomy construction as an iterative design process, with explicit attention to operationalisation.

The initial step might be to conduct a structured literature review across the above traditions to produce a longlist of candidate support relations and boundary cases, followed by collapsing this longlist into a minimal working taxonomy guided by two pragmatic criteria: discriminability (can trained annotators reliably distinguish the categories) and actionability (does the distinction matter for downstream uses such as provenance interfaces). This could involve iterative reviews of the taxonomy by expert annotators (e.g. those familiar with linguistic theories of support) as well as non-expert readers.

To quantitatively validate the taxonomy, the next step is to produce a full annotation specification, definitions, canonical examples, counterexamples, and decision rules, explicitly targeted at usability by both human annotators and potentially LLMs-as-judge [45] with the goal of scaling annotation to larger datasets and enabling future online annotation. A practical path is to construct a benchmark of statement-source pairs by enriching existing groundedness and hallucination corpora (some examples are given in Section 3.2) with support-relation labels. Human reliability and construct validation could first be established with a human annotation study on a stratified sample of statement-source pairs. This would quantify inter-annotator agreement and reveal systematic confusions between relation types, enabling iterative refinement, and ideally result in a "gold standard" ground truth dataset of classified statement-source pairs. It would then be straightforward to benchmark the performance of frontier models on this annotation task by measuring their ability to match human annotations.

3 Related Work

3.1 Science and Technology for Augmented Reading

This project could build on prior HCI and design work in technologically-assisted reading. The Semantic Reader project [24], which encompasses much prior work in this area, outlines how citations in scientific papers can be enriched to support readers. For instance, inline citations can be coloured to indicate whether the reader has already recently encountered or read them, and can display personalised cards that explain how the work relates to their interests. The "CiteRead" [33] system integrates commentary from subsequently-published work to support evaluation of citations.

The "InkSync" system [20], while not aimed at helping readers verify LLM text generated from sources, adopts a "warn-verify-audit" approach where LLM-generated text that appears to contain new information is highlighted as such, and the user is prompted to manually verify it (e.g., through a web search). Similarly, the "GenAudit" system [19] identifies errors and suggests correct edit suggestions. Systems like these are helpful, but are focused on helping the user avoid factual errors, rather than evaluate the particular relationship of *supported* claims to the source text, as we propose.

A relevant prior is the "Traceable Texts" interface [14], which annotates phrases in AI-generated summaries with links to corresponding phrases in the source, helping with fact checking and indexing into sources for deeper reading. The DATATALES interface [38] employs a similar brushing/linking interaction for textual narratives generated from data charts. Going further, the "attribution gradients" interface [15] decomposes LLM-generated statements into claims, which are in turn linked to evidence sources, each of which is classified according to a 2 × 2 framework (first-degree vs. second-degree, support vs. contradiction). Another close precedent is the FACTS&EVIDENCE system [3], which helps users verify the factuality of a passage by decomposing it into claims and verifying the support for each claim against web sources. A related line of work proposes "co-audit" as a general framework for helping humans double-check AI-generated content, including design principles such as grounding outputs with sources and not allowing the LLM to audit itself [11]. Similarly, interactive task decomposition interfaces that surface editable assumptions and execution plans have been shown to improve users' ability to steer and verify AI-generated analyses [16].

In the programming domain, the "Trailblazer" interface [42] helps developers understand LLM-generated answers about a codebase by visualising the trace of the agent's exploration of the codebase. This is an interesting approach because, as we will observe in the next section regarding much knowledge work, answers often do not draw on simple, single sources, but are distributed across a code (or more generally, knowledge) base. A somewhat similar interactive hierarchical exploration for academic papers is exemplified in the "Qlarify" [9] and "TreeReader" [44] interfaces.

3.2 Hallucination Detection and Groundedness Benchmarks

There is a vast literature on hallucination detection and various benchmarks that we cannot exhaustively review here (a review is given by Kazlaris et al. [17]). Prominent examples include HaluEval [22], HaDes [23], FactCHD [7], RAGTruth [31], DiaHalu [6], and Hallulens [2]. These are noted as potential starting points, since these datasets are often composed of source-statement pairs annotated with binary labels, that we could expand with a richer set of support relation labels.

Prior research has gone beyond binary categorisations, albeit not in the direction of classifying different types of positive support as we have proposed here. Rather, here the focus is in understanding different types of errors. For instance, the FRANK benchmark [32] identifies a typology of seven errors, such as relation error, entity error, out-of-article error, and grammatical errors. Similarly, the LibreEval dataset [1] identifies six hallucination subtypes. The FActScore benchmark [28] uses a ternary top level categorisation: supported, not supported and irrelevant; however in a qualitative analysis they note finer-grained types pertaining to the scope of the error, such as single-sentence contradiction and page-level contradiction, as well as pertaining to the semantic manipulation, such as "subjective" judgements, and cases where the source is itself wrong or inconsistent. Similarly, AttrScore [43] uses a ternary classification: the statement is either attributable (i.e., supported), extrapolatory (i.e., unsupported), or contradictory.

The Claimify method [26] notes that identifying how LLM-generated statements map onto knowledge claims is itself a complex and nonstandard procedure, and the method proposed can extract atomic claims from complex sentences, and identify amiguous claims that cannot be resolved. The claims so extracted can then be used as part of a verification pipeline such as VeriTrail [27]. It may also be possible to frontload the identification of claim spans at generation time, using a technique such as symbolically grounded generation [13] or citation-enabled LLMs (a review is given by Gao et al. [10]).

4 Discussion and Open Questions

There is a real tension between a taxonomy that is theoretically justified and one that is operationally usable at scale. Many candidate relation types depend on background knowledge and context, and the family of relations that involve support contingent on ancillary assumptions can proliferate rapidly if not constrained, because almost any inference can be described as requiring a tacit premise. Any attempt to import the full richness of scholarly discourse on argumentation, linguistics, and philosophy of language risks producing an ontology that is too fine-grained for annotators to apply consistently, or too abstract to be useful to readers as decision support. This implies an open design question: what is the smallest set of support relations that remains meaningful for annotators and readers, remaining theoretically advised by but without "overfitting" to philosophically motivated distinctions?

There are basic questions of the "units" of analysis: we have proposed defining the task as relating a generated statement S to evidence in a document D, but in practice it is rarely obvious what should count as a single statement, how to segment complex sentences, or when to decompose an answer into smaller atomic propositions. Existing pipelines highlight both opportunity and hazard: atomic decomposition can make verification tractable, but it introduces a consequential modelling choice about how fine-grained the decomposition should be and what contextual information must remain attached so that the unit remains faithful and interpretable [26]. For a support-relations taxonomy, this raises an open question about the division of labour between claim extraction and support labelling.

Similarly, many answers are supported not by a single contiguous span, but by multiple excerpts distributed across a document, and sometimes by excerpts that appear to pull in different directions. Real knowledge work is rarely partitioned into isolated documents: what matters is often a cloud of concepts, commitments, themes, and discussions distributed across conversations and artefacts, with users wanting to trace where else something was discussed and how it connects to other work. This raises questions about how to scale the representation when the graph of relevant materials becomes larger. Our design intent is not for the reader's evaluation process to fall back on "the AI knows and tells me," but to preserve users' ability to know where statements come from and to dive into the right parts of the underlying material. For the benchmark and annotation scheme, this implies open questions about whether the support relation should be defined between S and a single excerpt, between S and a set of excerpts, or between S and a structured evidential object that can represent corroboration and tension within D and across $D_1, \ldots, D_n$. It also forces an explicit choice about how to treat internal inconsistency: whether contradictory support should be surfaced to readers as a distinct relation, a meta-property of the evidence set, or discursively within the generated statement itself.

Another open issue is domain transfer. Support relations may manifest differently across summarisation, document-grounded question answering, and more analytical or evaluative queries, and even knowledge work domains (e.g., legal work, medical work, scientific research, etc.) because the permissible interpretive moves, the expected level of compression, and the availability of an agreed evidential standard vary by task and domain. We must also be sensitive to the fragilities of the LLM-as-judge paradigm: model behaviour and evaluation can depend strongly on task framing and prompts. Recent work has shown that even highly detailed rubric instructions yield only marginal improvements in LLM-as-judge alignment with human judgements, and that simpler measures such as perplexity can sometimes perform comparably [29]. Moreover, LLM-generated "explanations" of their own reasoning have been shown to be unreliable, since they do not reflect the model's actual mechanism [36].

We close by inviting collaboration precisely because we believe these to be productive research questions. The project calls for interdisciplinary input on which theoretical frameworks offer the best starting points for an operational taxonomy. We particularly welcome discussion of alternative choices about claim granularity and decomposition, and discussion on principled ways to represent support when evidence is distributed across multiple spans or multiple documents, including cases where evidence is cumulative or internally in tension. Finally, we hope to discuss benchmarking best practices and prompt design, including protocols that make prompt sensitivity and task transfer explicit. The objective is to build support for understanding provenance and critical reading in a form that remains legible and actionable for end users.

References

[1] Arize AI. [n. d.]. LibreEval: The Open-Source Benchmark for RAG Hallucination Detection. https://arize.com/llm-hallucination-dataset/. Accessed 4 February 2026.

[2] Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: Llm hallucination benchmark. *arXiv preprint arXiv:2504.17550* (2025).

[3] Varich Boonsanong, Vidhisha Balachandran, Xiaochuang Han, Shangbin Feng, Lucy Lu Wang, and Yulia Tsvetkov. 2025. FACTS&EVIDENCE: An Interactive Tool for Transparent Fine-Grained Factual Verification of Machine-Generated Text. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, Nouha Dziri, Sean (Xiang) Ren, and Shizhe Diao (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 437–448. doi:10.18653/v1/2025.naacl-demo.35

[4] Simon Buckingham Shum, Enrico Motta, and John Domingue. 2000. ScholOnto: an ontology-based digital library server for research documents and discourse. *International Journal on Digital Libraries* 3, 3 (2000), 237–248. doi:10.1007/s007990000034

[5] Simon J. Buckingham Shum, Victoria Uren, Gangmin Li, Bertrand Sereno, and Clara Mancini. 2007. Modelling naturalistic argumentation in research literatures: representation and interaction design issues. *International Journal of Intelligent Systems* 22, 1 (2007), 17–47. doi:10.1002/int.20188

[6] Kedi Chen, Qin Chen, Jie Zhou, He Yishen, and Liang He. 2024. Diahalu: A dialogue-level hallucination evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 9057–9079.

[7] Xiang Chen, Duanzheng Song, Honghao Gui, Chenxi Wang, Ningyu Zhang, Yong Jiang, Fei Huang, Chengfei Lyu, Dan Zhang, and Huajun Chen. 2024. FactCHD: benchmarking fact-conflicting hallucination detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (Jeju, Korea) *(IJCAI '24)*. Article 687, 9 pages. doi:10.24963/ijcai.2024/687

[8] Andrea Cohen, Sebastian Gottifredi, Alejandro J. García, and Guillermo R. Simari. 2014. A survey of different approaches to support in argumentation systems. *The Knowledge Engineering Review* 29, 5 (2014), 513–550. doi:10.1017/S0269888913000325

[9] Raymond Fok, Joseph Chee Chang, Tal August, Amy X. Zhang, and Daniel S. Weld. 2024. Qlarify: Recursively Expandable Abstracts for Dynamic Information Retrieval over Scientific Papers. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 145, 21 pages. doi:10.1145/3654777.3676397

[10] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6465–6488. doi:10.18653/v1/2023.emnlp-main.398

[11] Andrew D. Gordon, Carina Negreanu, José Cambronero, Rasika Chakravarthy, Ian Drosos, Hao Fang, Bhaskar Mitra, Hannah Richardson, Advait Sarkar, Stephanie Simmons, Jack Williams, and Ben Zorn. 2024. Co-audit: tools to help humans double-check AI-generated content. *Proceedings of the 14th annual workshop on the intersection of HCI and PL (PLATEAU 2024)* (5 2024). doi:10.1184/R1/25587552.v1

[12] Paul Grice. 1991. *Studies in the Way of Words.* Harvard University Press.

[13] Lucas Torroba Hennigen, Shannon Shen, Aniruddha Nrusimha, Bernhard Gapp, David Sontag, and Yoon Kim. 2024. Towards Verifiable Text Generation with Symbolic References. arXiv:2311.09188 [cs.CL] https://arxiv.org/abs/2311.09188

[14] Hita Kambhamettu, Jamie Flores, and Andrew Head. 2025. Traceable Texts and Their Effects: A Study of Summary-Source Links in AI-Generated Summaries. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 538, 7 pages. doi:10.1145/3706599.3719830

[15] Hita Kambhamettu, Alyssa Hwang, Philippe Laban, and Andrew Head. 2025. Attribution Gradients: Incrementally Unfolding Citations for Critical Examination of Attributed AI Answers. arXiv:2510.00361 [cs.HC] https://arxiv.org/abs/2510.00361

[16] Majeed Kazemitabaar, Jack Williams, Ian Drosos, Tovi Grossman, Austin Zachary Henley, Carina Negreanu, and Advait Sarkar. 2024. Improving Steering and Verification in AI-Assisted Data Analysis with Interactive Task Decomposition. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 92, 19 pages. doi:10.1145/3654777.3676345

[17] Ioannis Kazlaris, Efstathios Antoniou, Konstantinos Diamantaras, and Charalampos Bratsas. 2025. From Illusion to Insight: A Taxonomic Survey of Hallucination Mitigation Techniques in LLMs. *AI* 6, 10 (2025). doi:10.3390/ai6100260

[18] Charles W Kneupper. 1978. Teaching argument: An introduction to the Toulmin model. *College Composition & Communication* 29, 3 (1978), 237–241.

[19] Kundan Krishna, Sanjana Ramprasad, Prakhar Gupta, Byron C. Wallace, Zachary C. Lipton, and Jeffrey P. Bigham. 2025. GenAudit: Fixing Factual Errors in Language Model Outputs with Evidence. arXiv:2402.12566 [cs.CL] https://arxiv.org/abs/2402.12566

[20] Philippe Laban, Jesse Vig, Marti Hearst, Caiming Xiong, and Chien-Sheng Wu. 2024. Beyond the Chat: Executable and Verifiable Text-Editing with LLMs. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 20, 23 pages. doi:10.1145/3654777.3676419

[21] Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1121, 22 pages. doi:10.1145/3706598.3713778

[22] Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6449–6464. doi:10.18653/v1/2023.emnlp-main.397

[23] Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 6723–6737. doi:10.18653/v1/2022.acl-long.464

[24] Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, Fangzhou Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Kinney, Aniket Kittur, Hyeonsu B. Kang, Egor Klevak, Bailey Kuehl, Michael J. Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita Rao, Paul Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. 2024. The Semantic Reader Project. *Commun. ACM* 67, 10 (Sept. 2024), 50–61. doi:10.1145/3659096

[25] Clara Mancini and Simon J. Buckingham Shum. 2006. Modelling discourse in contested domains: A semiotic and cognitive framework. *International Journal of Human-Computer Studies* 64, 11 (2006), 1154–1171. doi:10.1016/j.ijhcs.2006.07.002

[26] Dasha Metropolitansky and Jonathan Larson. 2025. Towards Effective Extraction and Evaluation of Factual Claims. *arXiv preprint arXiv:2502.10855* (2025).

[27] Dasha Metropolitansky and Jonathan Larson. 2025. VeriTrail: Closed-Domain Hallucination Detection with Traceability. *arXiv preprint arXiv:2505.21786* (2025).

[28] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 12076–12100.

[29] Bhuvanashree Murugadoss, Christian Poelitz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. 2025. Evaluating the Evaluator: Measuring LLMs' Adherence to Task Evaluation Instructions. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 18 (Apr. 2025), 19589–19597. doi:10.1609/aaai.v39i18.34157

[30] Josh M Nicholson, Milo Mordaunt, Patrice Lopez, Ashish Uppala, Domenic Rosati, Neves P Rodrigues, Peter Grabitz, and Sean C Rife. 2021. scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative science studies* 2, 3 (2021), 882–898.

[31] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 10862–10878.

[32] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4812–4829.

[33] Napol Rachatasumrit, Jonathan Bragg, Amy X. Zhang, and Daniel S Weld. 2022. CiteRead: Integrating Localized Citation Contexts into Scientific Paper Reading. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 707–719. doi:10.1145/3490099.3511162

[34] Advait Sarkar. 2023. Exploring Perspectives on the Impact of Artificial Intelligence on the Creativity of Knowledge Work: Beyond Mechanised Plagiarism and Stochastic Parrots. In *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work* (Oldenburg, Germany) *(CHIWORK '23)*. Association for Computing Machinery, New York, NY, USA, Article 13, 17 pages. doi:10.1145/3596671.3597650

[35] Advait Sarkar. 2024. AI Should Challenge, Not Obey. *Commun. ACM* (Sept. 2024), 5 pages. doi:10.1145/3649404 Online First.

[36] Advait Sarkar. 2024. Large Language Models Cannot Explain Themselves. In *Proceedings of the ACM CHI 2024 Workshop on Human-Centered Explainable AI* (Honolulu, HI, USA) *(HCXAI at CHI '24)*. doi:10.48550/arXiv.2405.04382

[37] Advait Sarkar, Xiaotong (Tone) Xu, Neil Toronto, Ian Drosos, and Christian Poelitz. 2024. When Copilot Becomes Autopilot: Generative AI's Critical Risk to Knowledge Work and a Critical Solution. In *Proceedings of the Annual Conference of the European Spreadsheet Risks Interest Group (EuSpRIG 2024)*.

[38] Nicole Sultanum and Arjun Srinivasan. 2023. DATATALES: Investigating the use of Large Language Models for Authoring Data-Driven Articles. In *2023 IEEE Visualization and Visual Analytics (VIS)*. 231–235. doi:10.1109/VIS54172.2023.00055

[39] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The Metacognitive Demands and Opportunities of Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 680, 24 pages. doi:10.1145/3613904.3642902

[40] Stephen E Toulmin. 2003. *The Uses of Argument.* Cambridge University Press, Cambridge, England.

[41] Victoria Uren, Simon Buckingham Shum, Michelle Bachler, and Gangmin Li. 2006. Sensemaking tools for understanding research literatures: design, implementation and user evaluation. *International Journal of Human-Computer Studies* 64, 5 (2006), 420–445. doi:10.1016/j.ijhcs.2005.09.004

[42] Litao Yan, Jeffrey Tao, Lydia B Chilton, and Andrew Head. 2025. Answering Developer Questions with Annotated Agent-Discovered Program Traces. In

Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25). Association for Computing Machinery, New York, NY, USA, Article 29, 14 pages. doi:10.1145/3746059.3747652

[43] Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic Evaluation of Attribution by Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 4615–4635. doi:10.18653/v1/2023.findings-emnlp.307

[44] Zijian Zhang, Pan Chen, Fangshi Du, Runlong Ye, Oliver Huang, Michael Liut, and Alán Aspuru-Guzik. 2025. TreeReader: A Hierarchical Academic Paper Reader Powered by Language Models. In *2025 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 286–292. doi:10.1109/VL-HCC65237.2025.00039

[45] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.