

Attribution Gradients: Incrementally Unfolding Citations for Critical Examination of Attributed AI Answers

Hita Kambhampettu
Computer and Information
Science
University of Pennsylvania
Philadelphia, USA
hitakam@seas.upenn.edu

Alyssa Hwang
Computer and Information
Science
University of Pennsylvania
Philadelphia, USA
ahwang16@seas.upenn.edu

Philippe Laban
Microsoft Research
New York, New York, USA
plaban@microsoft.com

Andrew Head
Computer and Information
Science
University of Pennsylvania
Philadelphia, USA
head@seas.upenn.edu

Abstract

AI answer engines are a relatively new kind of information search tool: rather than returning a ranked list of documents, they generate an answer to a search question with inline citations to sources. But reading the cited sources is costly, and citation links themselves offer little guidance about what evidence they contain. We present attribution gradients, a technique to boost the informativeness of citations by consolidating scent and information prey in place. Its first feature is bringing evidence counts, supporting/contradictory excerpts, links to sources, and contextual explanations into one place. Its second feature is the ability to unravel second-degree citations in place. In a lab study, we demonstrate usage of the full gradient in a critical reading task and its support for deep engagement that increased the depth of what readers took away from the sources versus a standard citation and document QA design.

CCS Concepts

• Human-centered computing → Interactive systems and tools.

Keywords

intelligent reading tools, sensemaking, attributed AI-generated text

ACM Reference Format:

Hita Kambhampettu, Alyssa Hwang, Philippe Laban, and Andrew Head. 2026. Attribution Gradients: Incrementally Unfolding Citations for Critical Examination of Attributed AI Answers. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26)*, November 02–05, 2026, Detroit, MI, USA. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3830398.3830624>

1 Introduction

AI answer engines differ from their traditional search counterparts: rather than returning a ranked list of documents, they generate a written answer with inline citations attributed to source documents [20, 25]. Examples of such search engines include Perplexity [5], Microsoft Copilot [2], and You.com [7]. In these systems, the standard of presentation for citations remains a citation string: a number or label that links to a source document.

These citations are hard to verify. Answer engines frequently generate statements unsupported by their cited sources [25], surface

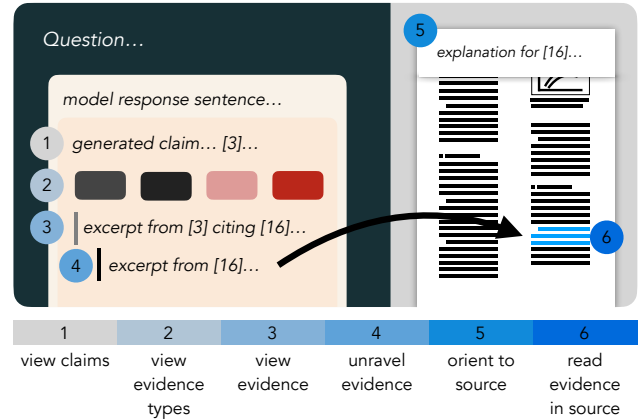


Figure 1: Attribution gradients. When using an attribution gradient, users begin with a claim in a generated answer (1) and see color-coded clusters of evidence (2): black indicates direct support; red, direct contradiction; gray, references to sources that support the claim; and pink, references to sources that contradict the claim. Users can explore evidence excerpts (3) that might include citations to secondary references (4). Users can view excerpts in the original PDF (6) with explanations of how the evidence relates to the claim (5). This graduated unfolding of context is designed to help users access a more nuanced understanding of support for a claim.

narrow perspectives [17], and reinforce biased views [23]. Yet users engage with cited sources less deeply than in traditional search [25]. From an information foraging perspective [22], the topology of a standard cited answer is not optimal: citation strings are low-scent, rarely conveying what evidence a source actually contains. The cost of looking up the source compounds when a cited source itself cites other work, what we call second-degree evidence, rather than providing direct evidence.

Prior work has developed parts of the solution. For instance, the ScholarQA literature-based question answerer [24] provides citation tooltips that show users source excerpts used in answer generation. Marco [13] lets users jump to excerpts of cited documents from tabulated answers about business document corpora. These tools reduce the cost of looking up the citations by surfacing excerpts and making it easy to jump to relevant passages.

In this work, we explore how to bring users yet more citation information with less effort. We articulate the idea of *attribution gradients*, where context about a citation is gradually unfolded to yield increasing information about citations under an incremental



This work is licensed under a Creative Commons Attribution 4.0 International License. <https://creativecommons.org/licenses/by/4.0/>

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2856-3/26/11

<https://doi.org/10.1145/3830398.3830624>

interaction step	scent	cost	added "partial prey"
view citation string	✓	low	(same as conventional answer)
identify subclaims	✓	low	is the fact I care about covered in the subclaims?
view evidence list	✓	low	excerpt snippets; counts of support/contradict, 1st/2nd-deg
jump to evidence in source		low	(same as conventional answer)
review source		moderate	parallel evidence highlighted; contextual explanations
view 2nd-deg evidence	✓	low	excerpt snippets
jump to 2nd-deg evidence in source		low	parallel evidence highlighted; contextual explanations

Figure 2: Articulation of support added by attribution gradients. Compared to conventional attributed AI answers, attribution gradients change the cost structure of looking up information about evidence by introducing new scent, lesser cost, or partial prey at the steps of interaction marked by a blue box.

cost structure. Our implementation of the gradient does this by letting a user to unpack a complex sentence into atomic claims, view supporting and contradictory excerpts from sources, jump to excerpts in their source documents, read contextual explanations of those excerpts, and unravel citations in excerpts to reach secondary literature, all with relatively little effort (Figure 1).

Most features of our gradient are known patterns from information interfaces across application areas. Evidence sorting, for instance, is a key feature of Selenite [18], which lets users review counts of common themes and valences in corpora of product reviews. Classification of citations into supporting and contrasting categories is a launched feature in the Scite scholarly search tool [21], which provides other classifications like the type of citing document. NEEDLE [11] supports parallel review of a user’s hypothesized claims and passages from a document corpus that support and contradict it. Paper Plain [9] and ScholarPhi [14] provide refined contextual explanation widgets in scholarly papers, targeting the explanation of dense and jargon-filled scholarly passages. Scim [12] facilitates discovery of salient document excerpts by highlighting them in place. These features have been individually validated in studies and deployments.

The first contribution of this work is the attribution gradients pattern. More than just a combination of past features, the idea of the gradient is to extend the citation with incremental, low-cost steps of increasing information scent and prey (Figure 2). Our second contribution is the feature of second-degree source unraveling. This part of the gradient lets users expand an excerpt’s citations in place to reach evidence in the sources it cites. Our third contribution is empirical evidence of the merit of the idea. In a lab study, we explored whether attribution gradients could drive deeper engagement with attributed sources and a more nuanced understanding of generated claims. We found that (1) participants used and found useful all parts of the gradient, (2) participants produced substantially higher-quality revisions of AI-generated claims, and (3) participants retained evidence they encountered, with effort comparable to a chat-with-papers baseline. Together, this work helps sketch out designs of cited AI answers that deliver more information about sources with less effort.

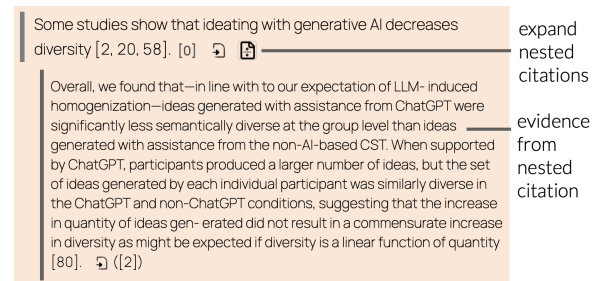


Figure 3: Unraveling citations. If a piece of second-degree evidence cites another source for its evidence, users can sometimes unravel the citation to that source. After clicking the button to expand nested citations, excerpts from the cited source are shown that pertain to the claim. In this case, evidence from citation “[2]” is shown that more precisely conveys how one of the cited studies shows that “ideating with generative AI decreases diversity.”

2 Design

The contribution of this paper is a prototype and evaluation of *attribution gradients*, which consolidate dense and rich evidence implied by a citation into an incrementally unfoldable evidence trail. The key idea is simple: rather than linking a citation directly to a source document, attribution gradients expand that single step into a series of low-cost interactions, each containing more information scent than the last. We instantiate attribution gradients as an extension to OpenSciLM [8], a scientific retrieval-augmented generation (RAG) QA system that answers questions by retrieving relevant open-access papers and generating an answer that cites them. Answers contain numerous sources, which in aggregate often tell a more complex narrative than the generated claims imply. Figure 4 shows the interface, annotated for each feature below.

Claim decomposition. A user begins by clicking on a cited sentence in the generated answer. The sentence is decomposed into atomic, testable claims, which are displayed beneath it. This gives users an entry point for looking up evidence: rather than evaluating a sentence as a whole, they can assess each claim independently.

Evidence overview. Selecting a claim displays a set of stacked bars summarizing the kinds of evidence that relate to it. Our gradient draws evidence from all of the answer’s citations and not just those in the current sentence, as we often found that the other citations had significant evidence to offer. Evidence can support or contradict the claim, and can be first-degree (the source provides the evidence itself) or second-degree (the source cites other sources for the evidence). This schematizes the evidence landscape, showing not just that a claim is cited, but whether its support is strong, weak, or mixed, and direct or indirect.

Evidence excerpts. Clicking into the overview shows the individual evidence excerpts. Excerpts are color-coded by type. This lets users scan for relevant evidence before committing to opening a source, lowering the cost of the citation link.

In-source viewing. Each excerpt has a “show source” button that opens the source PDF in a side panel, scrolled to and highlighting the

relevant passage. Alongside it, a generated contextual explanation relates the passage to the claim. Other extracted passages from the same source are shown in a less salient color, supporting further foraging within the document.

Second-degree source unraveling. When an excerpt is second-degree, citing other sources rather than providing evidence directly, users can expand it to see excerpts from those cited sources. To our knowledge, this feature of attribution gradients has not been explored in prior sensemaking systems. It connects readers to evidence buried in the citation chain without requiring them to manually track down additional sources (Figure 3).

2.1 Implementation

To develop and test the idea of attribution gradients, we developed a prototype implementation as an extension to the OpenSciLM interface [8] (see Appendix A.1 for details). Given a question, the system fetches an attributed answer from OpenSciLM’s web API and processes it through a separate LLM pipeline (we use GPT-5.1): claims are extracted from cited sentences, evidence excerpts are mined from source PDFs and classified by valence (supporting or contradicting) and degree (first- or second-degree), and each excerpt is paired with a generated contextual explanation. The valence classification is similar to prior work on citation-intent classification [10, 16]. For second-degree excerpts, the pipeline retrieves cited references via the Semantic Scholar API [15] and extracts excerpts from those sources. Excerpts are extracted from source PDFs using PaperMage [19] and highlighted in the UI via PaperCraft [4]. See Appendix A.3 for a preliminary analysis of the accuracy. In our analysis, valence classification was largely accurate, and degree classification was more mixed. We note however that it is often straightforward for a user to notice misclassified degree since second-degree excerpts visibly contain citations.

3 Evaluation

Attribution gradients are only useful if they help people interact with and learn from the sources. We sought evidence for (Q1) whether the gradients provide readers with a more nuanced understanding of the claims they read, grounded in the sources, and (Q2) whether readers used all features of the gradient, as our initial conjecture was that there was benefit to bringing them together.

We conducted a within-subjects lab study ($n = 15$) comparing attribution gradients to a baseline. Both conditions shared the same OpenSciLM response view; the baseline additionally provided ChatDOC [1], a chat-with-papers interface. This was intended as an alternative strong solution to problems attribution gradients sought to solve (QA for arbitrary questions with deep links into sources in the answers). For parity with attribution gradients, ChatDOC was preloaded with all sources cited across the answer and the second-degree sources that would appear in attribution gradients. When we observed participants, their use of the baseline resembled how a user might scan and query unordered lists of sources in a passage of a Deep Research answer. Participants overwhelmingly skipped the built-in citation link in the original answer and jumped right into scanning and querying papers in ChatDOC and clicking on citations in ChatDOC’s responses to read the sources.

Participants were mostly familiar with science papers, which were the kinds of documents they were required to consult in citations: they were doctoral (9), medical (5), and Master’s (1) students recruited from a private US university, all experienced with academic literature (93% had read over 20 papers) but intentionally varied in familiarity with the domain of the stimulus question (47% low, 33% moderate, 20% high). Sessions were held over Zoom. Before each task, participants received a tutorial for the interface used in that task. The experimenter demonstrated major features of the relevant interface, then asked the participant to practice those features on small practice tasks. Interface order was counterbalanced; participants were blinded to which interface we designed.

Each participant completed two 10-minute tasks, one per condition, reviewing evidence for a single claim as if preparing to explain it to a colleague. The same two study stimuli were used for all participants, with task order fixed and condition order counterbalanced. Reviewing evidence for the claim required reconciliation across multiple sources. The questions and answers that participants reviewed in the tasks were inspired by questions from ScholarQABench [8], OpenSciLM’s benchmark, lightly modified by the authors until the generated answers retrieved an interesting corpus of documents. The tasks were about recommendation algorithm effects on user preferences and ChatGPT’s effect on idea diversity. Each task pointed participants to one specific claim within a complete answer.

Engagement with citations was measured through: (1) a closed-book claim revision, in which participants rewrote the claim to reflect the evidence; (2) a summative assessment of the claim-evidence relationship. This was a multiple-choice question about how the evidence relates to the claim with options ranging from “almost certainly true” to “do not justify the claim” with the ground truth answer set by an external judge; (3) recall check questions, checking whether participants were able to recall specific pieces of evidence that they might have encountered during the task. Participants indicated whether 10 pieces of example evidence had appeared in their sources, including first-degree (8) and second-degree (2) evidence; among the options we provided three false negatives.

3.1 Analysis

An external PhD student with relevant domain expertise was recruited to review revised claims without knowledge of the condition in which they were written. Revised claims were evaluated on criteria of overall quality, specificity and quantity of evidence, faithfulness and correctness of the revised claim, and specificity with which the effect was described. Comparisons between conditions (for scores on revised claims, subjective Likert scores, behavioral events, and recall of evidence) were computed using the Wilcoxon signed-rank test ($\alpha = 0.05$). Qualitative accounts of behavior and reactions drew from a thematic analysis of post-task questionnaires and the experimenter’s structured notes.

4 Findings

Below, we describe our findings. Test results are reported with test statistics and p -values. Qualitative data is reported with attribution to participants using pseudonyms P1–15.

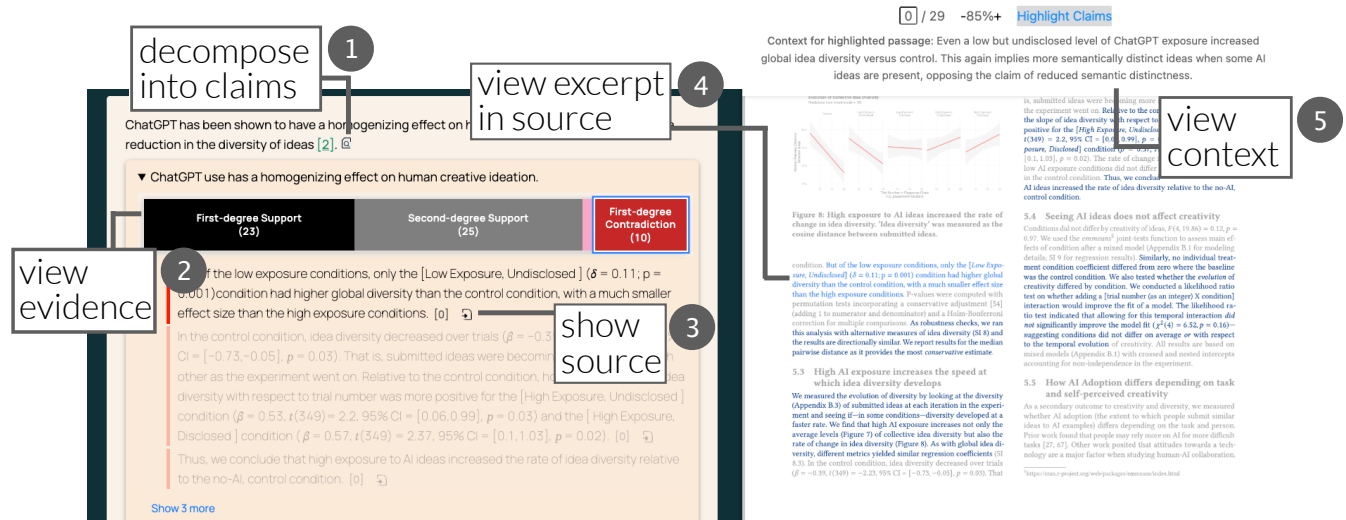


Figure 4: The full gradient. A reader can (1) decompose a sentence in the generated answer into claims, and (2) view evidence for a particular claim. A reader can see that attributed evidence is mixed for this claim. If a reader clicks on an evidence excerpt (3), they can (4) view it in the source and (5) read context for how the evidence excerpt relates to the claim. Counts shown in the evidence overview indicate the number of evidence excerpts of each type.

4.1 Effect on understanding of evidence (Q1)

Our measures suggested that participants better understood the nuance of the evidence for claims with attribution gradients and had mentally processed much of the evidence.

4.1.1 Effect on participants' revisions to the original claim. When asked to revise the claim to reflect the evidence encountered in a closed-book setting after interface use, participants produced revisions that were overall better in almost all dimensions judged. More specifically, significant differences favoring attribution gradients were found with regard to (Figure 5):

- Overall grade for the revised claim ($M = 4.00$, $SD = 0.93$ vs. $M = 2.87$, $SD = 0.92$; $W = 3.0$, $p = 0.004$)
- The specificity of evidence included ($M = 3.80$, $SD = 1.08$ vs. $M = 2.40$, $SD = 1.24$; $W = 2.5$, $p = 0.006$)
- The specificity with which the effect was described in the claim ($M = 3.93$, $SD = 0.80$ vs. $M = 2.60$, $SD = 1.06$; $W = 4.0$, $p = 0.003$)
- The faithfulness of the revision to the available evidence ($M = 4.00$, $SD = 0.85$ vs. $M = 2.87$, $SD = 0.99$; $W = 8.0$, $p = 0.008$)
- The factual correctness of the revised claim ($M = 4.13$, $SD = 0.74$ vs. $M = 3.07$, $SD = 1.16$; $W = 8.0$, $p = 0.007$)

To correct for multiple comparisons across the six revision-quality analyses, we applied Holm-Bonferroni correction; the five significant differences remained significant (adjusted $p \leq 0.024$).

The quantity of evidence in the revised claim had a positive but not statistically significant difference with attribution gradients ($M = 3.33$, $SD = 1.11$ vs. $M = 2.60$, $SD = 1.06$; $W = 18.0$, $p = 0.09$).

4.1.2 Indications of retention. We asked participants to recall evidence they could have come across during the task, reporting “yes,”

“no,” or “I don’t recall” for a sample of evidence (reworded so as not to be verbatim with excerpts shown in the interfaces).

Out of a possible score of 100%, participants in the attribution gradients condition answered 51.3% correctly ($SD = 21.3$), suggesting that they had encountered, and remembered, a large amount of evidence. In other words, it is unlikely they were merely skimming the evidence. 39.3% ($SD = 17.1$) of their answers were “I don’t recall” and these were marked wrong by default (participants only received points for a correct “yes” or “no”); therefore, a 51.3% score suggests when they committed to “yes” or “no,” they largely did so correctly. Across questions, recall was 56.0% for questions about first-degree evidence ($SD = 22.9$), 46.7% ($SD = 33.5$) for questions about second-degree evidence, and 46.7% ($SD = 30.3$) for questions where the correct answer was “no.”

We also asked participants in the baseline condition to recall the same pieces of evidence. This instrument was somewhat biased toward attribution gradients: evidence items were chosen through use of the attribution gradients interface (roughly uniformly sampling from top to bottom of the evidence list). This was because we mainly planned to use it as a recall check for the attribution gradients condition. Our recall items were biased towards primary sources, yet participants in the baseline might have spent less time with them given that secondary sources were more numerous and the baseline offered no visual distinction between primary and secondary sources. Still, participants could and should have come across much of this evidence in the baseline; the evidence represented key results from the sources. Baseline users were largely unable to correctly recall if these pieces of evidence were present ($M = 17.3\%$, $SD = 13.3$; $W = 0.0$, $p = 0.001$) for first- ($M = 22.7\%$, $SD = 18.3$; $W = 4.0$, $p = 0.002$) and second-degree evidence ($M = 12.0\%$, $SD = 14.7$; $W = 0.0$, $p = 0.005$). Participants in the baseline condition overall reported significantly less certainty

about whether they had previously seen evidence, reporting “I don’t recall” 60% of the time ($SD = 30.7$; $W = 22.0$, $p = 0.030$).

4.2 Use of the full gradient (Q2)

Participants’ behaviors and questionnaire responses helped clarify how participants used features of the gradient, both individually and in tandem, during the claim review task. In general, participants reported the features of attribution gradients as highly useful. To calibrate, all features below were rated on average as at least 6 out of 7, while for comparison, participants rated features in the baseline below 5 (ability to ask questions of papers, $M = 4.71$, $SD = 1.59$; ability to jump to passage from answer, $M = 4.79$, $SD = 1.93$).

Evidence overview. All participants filtered evidence types at least once when using attribution gradients. 15 toggled to view first-degree support, 13 to view second-degree support, and 7 to view first-degree contradictions (second-degree contradictions were not available in the provided stimuli). Participants reported the sorting of excerpts as highly useful, both along the dimensions of first- and second-degree evidence ($M = 6.79$, $SD = 0.58$) and supporting and contradictory valence ($M = 6.14$, $SD = 1.17$).

As P11 described, “in [the baseline], the onus was on me to generate questions from the contradictory perspective... [with attribution gradients] it was easier for me to see both sides of the claim.” Overall, we saw some indications that participants may have engaged more meaningfully with different valences of information with attribution gradients: of the eight participants whose revisions incorporated both supporting and contradictory evidence,¹ six had done so using attribution gradients. We note that all six of these participants had clicked to filter to contradictory evidence or open a contradictory excerpt during the task.

Second-degree source unraveling. When using attribution gradients, all 15 participants viewed at least one excerpt containing second-degree evidence, and 12 clicked to expand at least one second-degree excerpt to show excerpts from the sources it cited. P8 described that the second-degree excerpts helped them “go down more rabbit holes... if I wanted to know exactly what study was done.” Participants reported the ability to unravel citations to view original sources to be very useful ($M = 6.29$, $SD = 0.99$). This feature could support access to quite a lot of secondary sources; P4, for instance, opened seven sources providing second-degree evidence.

In-source viewing. The ability to open excerpts in the source was frequently exercised in the claim review task. Participants opened an excerpt in source a median of 20 times ($SD = 11.24$). Participants reported this feature as highly useful as well ($M = 6.79$, $SD = 0.58$). Participants’ descriptions aligned with the intent of the feature: P10 told us that with this feature they “found myself reading the source itself” and P4 told us that using the feature “made me more confident in citing the sources since I had more opportunities to read more papers and also more of each paper.” The source viewer also appeared to support another mode of foraging: six participants, for instance, clearly engaged with content in a paper aside from the excerpt they originally opened the paper to view. Sometimes

¹As assessed by an independent external grader during review of revised claims.

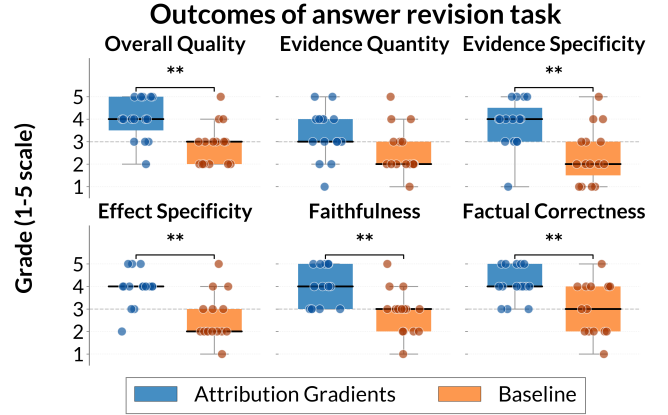


Figure 5: Comparisons of revised claims according to six quality dimensions. Participants using attribution gradients produced revisions that were rated significantly higher than the baseline on all dimensions except evidence quantity. Each box spans the interquartile range of grades, with the horizontal line marking the median. Each individual point is one participant’s grade. Brackets and asterisks indicate significance from paired Wilcoxon signed-rank tests comparing conditions within each dimension (** $p < .01$).

they scrolled to see additional content, and sometimes they clicked on different highlighted excerpts in the source.

Use of features together. Many participants reviewed lists in a way that coordinated features. For instance, we often saw a pattern resembling incremental, iterative processing of available evidence. Nine participants opened excerpts from the evidence overview in quick succession, viewing each briefly and working through on average about 50% of the list of excerpts that were displayed. P4 described the utility of this approach: “each amount of information [in the evidence excerpt] was small so I was able to quickly filter out which info was useful to me... I liked being able to be given all of the information so I can make that decision [of what to read].”

In addition to the features above, we validate use of the other features. By nature of interacting with the system at all, participants had descended into *claim decompositions*. By nature of opening sources, all participants had engaged with *evidence excerpts*.

5 Discussion

Our results suggest that structuring attributions as a gradient supports deeper engagement with the available evidence. Because we evaluated the gradient as a single integrated design, however, our results speak to the value of the assembled gradient rather than to any individual feature within it. We see four aspects of this design as potentially generative for attributed AI interfaces more broadly: *incrementality* (breaking review into smaller steps), *schematization* (pre-loading productive evidence schemas), *recursion* (treating sources as launching points into further literature), and *context* (providing orientation for decontextualized excerpts). These dimensions are not new to sensemaking research, but handling them well in attributed AI answers may change how users engage with

generated content that is linked back to sources. That said, our evidence comes only from scientific QA with readers experienced in academic literature; whether the pattern transfers to other settings is an open question we take up below.

5.1 Future studies

Our study might inspire future ones to be run across:

Different AI search engines. In principle, attribution gradients could apply wherever an answer provides citations, claims are in contention, and sources are available (e.g. Google AI Overviews and Deep Research [3]). Their efficacy may vary with source availability, source count, and evidence depth. Processing citations comes at a cost, and search engine builders will have to decide how many citations to process, how thoroughly.

Different participants. Are attribution gradients still useful when readers have more knowledge of claims and sources?

Different subsets of the gradient. We evaluated the full gradient. Which of its parts most benefited users? Follow-on work could isolate its individual parts, and imperfections in them (e.g., how misclassified evidence affects readers' conclusions, and whether readers detect and recover from such errors).

5.2 Future systems

We see several areas for fruitful future systems work.

Richer schemas. Attribution gradients could be expanded beyond the binary support-versus-contradict categorization to incorporate richer, multi-dimensional contextual relationships between claims and evidence. Users might glean more meaning that goes beyond whether evidence agrees or disagrees with a claim.

Richer evidence representations. If raw data, such as tables or numeric results, can be extracted from sources, then attribution gradients could provide greater context about the relationship of this data to claims. A gradient could incorporate information about whether statistical tests could be replicated, the uncertainty of a result, how a statistic was computed, or issues in a logical argument.

Richer evidence pools. Our gradients draw only on the sources cited in the original answer, so they help readers assess whether the answer represents its own citations. If the AI answer engine selects sources unevenly, the evidence shown in the overview inherits that bias. A fully critical future instantiation could do a post-hoc search for evidence on all sides of a claim before constructing the gradient.

6 Conclusion

We introduce attribution gradients to support the close reading of attributed AI answers. The gradient consolidates information about the claim, the valence and nature of the evidence, evidence excerpts, evidence chains, and contextual information, into one view. Our study showed participants revising claims in a way that suggested having acquired more information from attribution gradients than in the baseline. This work demonstrates how reframing citation inspection as a gradient of increasing information scent enables users to make better use of attributed AI answers.

Acknowledgments

We thank Princess Sampson, Mialy Rasetarinera, and Xiaorui Liu for grading participant responses, and Litao Yan for his feedback on figures and study design. This research was developed with funding from the Defense Advanced Research Projects Agency's SciFy program (Agreement No. HR00112520300). The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

- [1] [n. d.]. *chatdoc.com*. <https://chatdoc.com/> Last accessed 31 July 2026.
- [2] [n. d.]. *Microsoft Copilot*. <https://copilot.microsoft.com/> Last accessed 31 July 2026.
- [3] [n. d.]. *OpenAI Deep Research*. <https://openai.com/index/introducing-deep-research/> Last accessed 31 July 2026.
- [4] [n. d.]. *PaperCraft*. <https://openreader.semanticscholar.org/PaperCraft> Last accessed 31 July 2026.
- [5] [n. d.]. *Perplexity*. <https://perplexity.ai/> Last accessed 31 July 2026.
- [6] [n. d.]. *pypdf*. <https://pypi.org/project/pypdf/> Last accessed 31 July 2026.
- [7] [n. d.]. *you.com*. <https://you.com> Last accessed 31 July 2026.
- [8] Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'Arcy, David Wadden, Matt Latzke, Jenna Sparks, Jena D. Hwang, Varsha Kishore, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Daniel S. Weld, Doug Downey, Wen-tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. 2026. Synthesizing scientific literature with retrieval-augmented language models. *Nature* 650, 8103 (2026), 857–863. <https://doi.org/10.1038/s41586-025-10072-4>
- [9] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A Hearst, Andrew Head, and Kyle Lo. 2023. Paper Plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Transactions on Computer-Human Interaction* 30, 5, Article 74 (2023).
- [10] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural Scaffolds for Citation Intent Classification in Scientific Publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 3586–3596.
- [11] Moritz Dück, Steffen Holter, Robin Shing Moon Chan, Rita Sevastjanova, and Mennatallah El-Assady. 2025. Finding Needles in Document Haystacks: Augmenting Serendipitous Claim Retrieval Workflows. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Article 1003.
- [12] Raymond Fok, Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti Hearst, Andrew Head, and Daniel S Weld. 2023. Scim: Intelligent skimming support for scientific papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. ACM, 476–490.
- [13] Raymond Fok, Nedim Lipka, Tong Sun, and Alexa F Siu. 2024. Marco: Supporting Business Document Workflows via Collection-Centric Information Foraging with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM, Article 842.
- [14] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S Weld, and Marti A Hearst. 2021. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Article 413.
- [15] Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler C. Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The Semantic Scholar Open Data Platform. *ArXiv abs/2301.10140* (2023).
- [16] Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An Argument-Annotated Corpus of Scientific Publications. In *Proceedings of the 5th Workshop on Argument Mining*, Noam Slonim and Ranit Aharonov (Eds.). Association for Computational Linguistics, 40–46.
- [17] Nora Freya Lindemann. 2024. Chatbots, search engines, and the sealing of knowledges. *AI & SOCIETY* (2024), 5063–5076.

- [18] Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A Myers. 2024. Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM, Article 837.
- [19] Kyle Lo, Zejiang Shen, Benjamin Newman, Joseph Z Chang, Russell Authur, Erin Bransom, Stefan Candra, Yoganand Chandrasekhar, Regan Huff, Bailey Kuehl, et al. 2023. PaperMage: a unified toolkit for processing, representing, and manipulating visually-rich scientific documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 495–507.
- [20] Shahan Ali Memon and Jevin D West. 2024. Search engines post-ChatGPT: How generative artificial intelligence could make search less reliable. *arXiv preprint arXiv:2402.11707* (2024).
- [21] 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.
- [22] Peter Pirolli and Stuart Card. 1995. Information foraging in information access environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 51–58.
- [23] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative echo chamber? Effect of LLM-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM, Article 1033.
- [24] Amanpreet Singh, Joseph Chee Chang, Dany Haddad, Aakanksha Naik, Jena D. Hwang, Rodney Kinney, Daniel S Weld, Doug Downey, and Sergey Feldman. 2025. Ai2 Scholar QA: Organized Literature Synthesis with Attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Pushkar Mishra, Smaranda Muresan, and Tao Yu (Eds.). Association for Computational Linguistics, 513–523.
- [25] Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Yixin Mao, and Chien-Sheng Wu. 2025. Search engines in the AI era: A qualitative understanding to the false promise of factual and verifiable source-cited responses in LLM-based search. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 1325–1340.

A Appendix

A.1 Implementation details

Our implementation is a proof of concept combining off-the-shelf AI components with some human supervision. The pipeline is too slow for deployment in existing answer engines. This was an intentional tradeoff we made so that our study could better characterize the upper bound of the design’s utility. We used the strongest general-purpose model available to us at the time of development, GPT-5.1, in all stages of the pipeline. Prompts appear in Appendix A.2.

Fetching answers. We query OpenSciLM’s web API [8], which generates attributed answers over 45 million open-access papers and returns answers along with citations that have unique IDs for looking up cited papers.

Extracting claims. For each sentence containing a citation, we prompt the LLM to return minimal, testable scientific claims relating to the sentence’s main point.

Extracting excerpts. We fetch PDFs for cited papers via open-access links from the Semantic Scholar API [15] and extract their text with pyPDF [6]. We prompt the LLM with this text to extract evidence excerpts relating to each claim, classify each excerpt according to degree and valence, and generate an explanation to help make the passage understandable out of context.

Fetching second-degree sources. We resolve references to sources within second-degree excerpts using Asta text snippet search, or, failing that, by extracting relevant reference strings with an LLM and matching them to documents with the Semantic Scholar API.

Some sources could still not be resolved automatically; we looked up these references manually and downloaded their PDFs when not paywalled. This covered 74.5% of sources referenced in second-degree excerpts. Given the large number of second-degree sources to process, we opted for a more rapid excerpt extraction prompt for these sources, which limited the number of extracted excerpts to three per source.

UI. We processed PDFs to add interactions and annotations offline before study sessions. To highlight excerpts in source PDFs, we located them with PaperMage [19] by matching high-overlap token sequences. Issues arose when PDFs consisted of images of text rather than text itself, leading to the exclusion of one PDF, or when the LLM-extracted evidence was not verbatim. We found that passing the full text into GPT-5.1 and prompting it to extract verbatim excerpts mitigated this latter issue. We added UI components from the PaperCraft [4] library to the PDF to support scrolling to and highlighting the detected passages.

A.2 Main LLM prompts used in attribution gradients pipeline

All prompts below were submitted to GPT-5.1 with structured output format (Pydantic/JSON-schema). Source papers were chunked into 8,000–10,000 character segments to fit context limits.

A.2.1 Evidence extraction. Used to identify all passages in a source paper that support or contradict a given claim.

****TASK: COMPREHENSIVE EVIDENCE EXTRACTION****

Conduct a thorough analysis of the research paper below to extract ALL passages that provide credible evidence for or against the given statement. Apply rigorous scientific standards throughout your analysis.

****EXTRACTION REQUIREMENTS:****

1. ****Completeness**:** Review the entire text systematically. Do not skip any sections including abstract, methods, results, discussion, or conclusions.
2. ****Verbatim Accuracy**:** Extract passages exactly as written in the source, including:
 - All punctuation, parentheticals, and formatting
 - Complete citation strings (e.g., "[2]", "(Smith et al., 2023)")
 - Statistical notations, p-values, effect sizes, confidence intervals
 - Technical terminology and abbreviations as written
3. ****Evidence Standards**:** Only extract content that constitutes legitimate scientific evidence:
 - ****INCLUDE**:** Empirical results, statistical findings, experimental outcomes, systematic observations, meta-analysis results, validated measurements
 - ****INCLUDE**:** Citations to established research findings when relevant to the statement

- ****EXCLUDE****: Hypotheses, research objectives, methodology descriptions (unless methodology itself is the subject of evaluation), author opinions, speculation

4. ****Passage Length****: Typically one sentence, but may extend to multiple sentences if required for:
- Complete statistical reporting (e.g., including confidence intervals, effect sizes)
 - Coherent presentation of complex findings
 - Preservation of essential context

****EVIDENCE CLASSIFICATION****

For each extracted passage, provide these annotations:

* ****Valence****:

- "support": Evidence clearly supports or aligns with the statement
- "contradict": Evidence clearly opposes or contradicts the statement
- Apply strict standards, ambiguous or weak evidence should be excluded entirely

* ****Evidence Kind****:

- "primary": Original findings, data, or analyses presented for the first time in this paper
- "secondary": References to established findings from prior published research

* ****Explanation****: Provide essential context in clear, accessible language:

- Define technical terms or statistical concepts
- Explain the relevance to the statement
- Clarify what the evidence demonstrates
- Use simple, direct sentences appropriate for a general academic audience

Output the extracted passages and annotations as a JSON object with this format. This example assumes an input statement of "Eating eggs accelerates muscle growth.":

```
{
  "passages": [
    {
      "passage": "Prior studies have shown that consumption of eggs led to an increase in volume of the deltoids [1][2].",
      "evidence_kind": "secondary",
      "valence": "support",
      "explanation": "Deltoids are a muscle group. They are sometimes measured in studies of muscle growth."
    },
    {
      "passage": "Our ANOVA showed no significant effect (p = .31).",
      "evidence_kind": "primary",
      "valence": "contradict",
      "explanation": "The test looked for different muscle growth for egg eaters. The p-value is high. This suggests the difference was a fluke. So there is no evidence here for eggs affecting muscle growth."
    }
  ]
}
```

```
]
}
```

Statement: {STATEMENT}

Text: {TEXT}

A.2.2 Claim Decomposition. Used to break AI-generated answer summaries into atomic, independently verifiable claims.

****TASK: SYSTEMATIC CLAIM EXTRACTION AND DECOMPOSITION****

Analyze the provided AI-generated research summary to identify and decompose verifiable scientific claims that require evidence validation.

****STEP 1: SENTENCE IDENTIFICATION****

Parse the summary into individual sentences. Focus exclusively on sentences that:

- Contain explicit citation markers (e.g., "[1]", "[3]", "(Author et al., Year)")
- Present factual assertions that can be empirically verified
- Include substantive information relevant to the research question
- Make claims that would benefit from validation through cited source materials

****STEP 2: CLAIM DECOMPOSITION CRITERIA****

For each qualifying sentence, extract minimal, testable scientific claims according to these principles:

1. ****Atomicity****: Each claim should test exactly one scientific proposition
2. ****Testability****: Claims must be empirically verifiable through research evidence
3. ****Clarity****: Use precise, unambiguous language avoiding compound or complex structures
4. ****Relevance****: Claims should directly relate to the core assertions of the sentence
5. ****Uniqueness****: Avoid duplicate or highly similar claims

****CONSTRAINTS****

- Maximum 6 claims per sentence (typically 1-3)
- Claims must be complete, standalone statements
- Preserve scientific accuracy from the original text
- Maintain appropriate level of specificity

****QUALITY STANDARDS****

- ****INCLUDE****: Specific research findings, quantitative relationships, causal claims, comparative statements, methodological assertions
- ****EXCLUDE****: Vague generalizations, obvious statements, purely definitional content, speculative language

****EXAMPLE TRANSFORMATION****

Original: "Daily consumption of omega-3 fatty acids significantly improves cognitive performance and reduces inflammatory markers in elderly adults [2][5]."

Decomposed Claims:

- "Daily consumption of omega-3 fatty acids significantly improves cognitive performance in elderly adults."
- "Daily consumption of omega-3 fatty acids reduces inflammatory markers in elderly adults."

RESEARCH CONTEXT:

Question: {{QUESTION}}

TEXT TO ANALYZE:

{TEXT}

A.3 Accuracy of evidence extraction for stimuli

While the implementation is not a contribution of this work, we briefly describe the accuracy of the evidence extraction and classification shown to users in the study. This helps contextualize what users were seeing in the study—evidence that was largely extracted and classified correctly, but not perfectly. We recruited an HCI Ph.D. student with experience reading in the tasks' domains to examine all 113 evidence excerpts participants might see across the two tasks and tag their relationship to the claim.² Of excerpts that the judge tagged as supporting, 73.5% were correctly classified by our pipeline as such; of those classified otherwise, the vast majority (92.6%) were considered by the judge as broadly relevant to the claim but lacking specific evidence. All excerpts classified as contradicting were judged to be correctly classified. Classification into evidence kind was more mixed: for instance, what was classified as first-degree supporting evidence was judged to be evidence consisting of original data or results 36.4% of the time; theoretical arguments, synthesis, or interpretation of facts 40.0% of the time (which can sometimes be considered first-degree evidence), otherwise broadly relevant 20.0% of the time, and unjustified assertions 3.6% of the time. For second-degree supporting evidence, the judge reported that the evidence referenced other sources 46.8% of the time, unjustified assertions 12.8% of the time, and broadly relevant information 29.8% of the time. In other words, valence was largely correct and degree had more mixed results.

A.4 Effort and questions in the baseline

The baseline condition involved some kinds of work not present in the attribution gradients condition. Participants often engaged with documents through the conversational interface of ChatDOC, asking on average 2.4 questions. Some of the most common kinds of questions were:

- Information extraction and summarization ($n = 5$), such as "Provide the top 10 takeaways from these studies and what they say about how recommendation algorithms affect the formation of users' preferences"

- Clarification questions ($n = 5$), such as "Did the paper provide a control of the user's preference before and after their intervention?"
- Requests for contradiction ($n = 2$), such as "How would someone might point out the flaws in these experiments setup to make counterargument about the claim?"

Participants also opened the source PDFs, though less often than with attribution gradients, opening a median of 2 papers ($M = 1.7$). Participants' accounts suggested the chat interface was valued for orientation and also as a substitute for reading. P8 found the baseline "really good for the AI chatbot for an overview without going through details." P5 noted "I found myself using the AI chatbot a little too much," describing how scrolling through generated answers "funneled me into agreement." P12 reflected that with the baseline "I never really understood what the main point was."

Moving between answer, sources, and chat could take time. One participant spent 47 seconds on a single pass: reading through the generated answer, clicking into part of it, opening the source PDF, and then asking ChatDOC a question to make sense of what they had read. Reading the results section of one paper took the same participant 6.5 minutes of the 10-minute task. Some participants also asked ChatDOC for the kinds of information attribution gradients provided: 2 asked for contradictory evidence and 1 asked where the chatbot's answers came from.

²This omitted unraveled excerpts from second-degree sources, but did include second-degree excerpts (those containing citations to other papers).