Traceable Texts and Their Effects: A Study of Source-Summary Links in AI-Generated Summaries

Hita Kambhamettu
hitakam@seas.upenn.edu
University of Pennsylvania
Philadelphia, PA, USA

Jamie Flores
Jamie.Flores@pennmedicine.upenn.edu
Perelman School of Medicine
Philadelphia, PA, USA

Andrew Head
head@seas.upenn.edu
University of Pennsylvania
Philadelphia, PA, USA

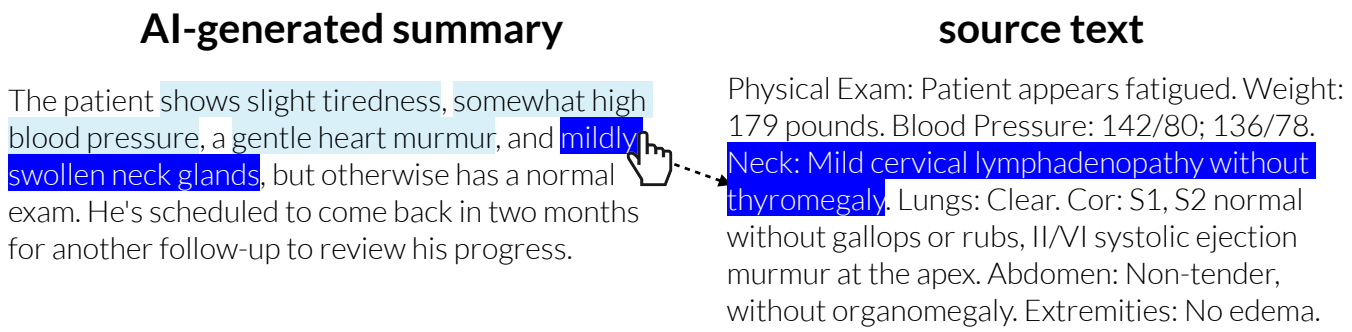

Figure 1: Interacting with a traceable text. We call an AI-generated summary with backlinks to a source text a "traceable text." Pictured is the kind of traceable text we studied in this paper. The generated summary is augmented with phrase-level links (e.g., "mildly swollen neck glands") to corresponding passages in a source text (e.g., "Neck: Mild cervical lymphadenopathy without thyromegaly"). The links help a reader to check the veracity of a passage in a summary and also offer an index into the source text if the reader wishes to learn more.

ABSTRACT

As AI-generated texts proliferate, how can people better understand the veracity of these texts? One increasingly common solution is linking AI-generated text to the sources from which they were derived. We call this design pattern *traceable text*. In this paper, we present a usability study of the effects of a traceable text on reading an AI-generated summary alongside a source document. We focus on a variant of traceable text that we expect to be increasingly common and general: phrase-level links from summary to source texts. With the traceable text, readers answered questions about the source text more quickly and more accurately when generated summaries contained hallucinations. In an open-ended task, participants used traceable text to aid understanding and to index into the texts. We define a broader design space of traceable text informed by examples in the literature.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; **Empirical studies in HCI**.

KEYWORDS

augmented text, inter-text links, summaries

CHI EA '25, April 26-May 1, 2025, Yokohama, Japan

ACM Reference Format:
Hita Kambhamettu, Jamie Flores, and Andrew Head. 2025. Traceable Texts and Their Effects: A Study of Source-Summary Links in AI-Generated Summaries. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3706599.3719830

1 INTRODUCTION

Advances in natural language processing (NLP) have significantly transformed the landscape of text simplification and summarization tasks, enabling AI models to distill complex information into more accessible and concise forms. From medical text [10] to legal documents [24], scientific literature [8] to online documents [40], interest is mounting in using AI to generate summaries to help users quickly understand the essence of lengthy, complicated texts. However, the convenience of summaries does not always eliminate the need for deeper engagement with the source material. Particularly in contexts where accuracy and detail are paramount, such as medical or legal documents, readers may need to verify the information in an AI-generated summary or explore a source text in greater depth. This necessity becomes even more critical when summaries contain errors, jargon, or misleading information.

To address this challenge, this paper explores how summary-to-source linkages can help readers engage with generated summaries and their sources. We term texts with such linkages *traceable texts*. We see such texts as potentially helping readers build trust in AI-generated content and engage more deeply with source texts. To assess the impact of traceable text on the reading experience, we conducted a usability study (Section 4). The study uses an instantiation of traceable text we expect to become common: a generated

summary with phrasal links back to related phrases in the source text. The study focused on a domain where source texts are complex, summaries are useful, and hallucinations matter: patients understanding their medical records [13, 27]. Participants read medical records augmented with traceable text and answered questions that required understanding of summaries and sources.

We found that the traceable text helped readers answer questions about the content of the source text significantly more quickly. When answering questions about hallucinatory content, we observed a marked improvement in correctness (from 12.5% to 70%); when answering questions about verified summaries, there was a statistically insignificant but noticeable improvement (75% to 90%). Participants brought their own medical notes for a final open-ended reading task. They found traceable text useful for understanding the details of the source note, for quickly locating relevant content, and for motivating deeper engagement with the source text.

2 BACKGROUND & RELATED WORK

Recent research has explored how to enrich reading experiences with links from texts to supporting information. Text has been linked to visualizations [34, 48], data tables [5, 29], code [51], diagrams [26], and math formulas [20]. Some of this research, like ours, has observed that modern LLMs can be used to produce these linkages (e.g., [48, 51]). This prior work highlights a community interest in supporting holistic reading of complex documents via lightweight augmentations to those documents. Our work shares this value and explores the effects of a related interaction primitive specifically for linking phrases in an AI-generated summary to corresponding phrases in its source text.

Significant research effort has gone into advancing automatic text summarization technologies (e.g., [9, 12, 30]). Some of this work has strived to produce concise textual summaries of specialized, lengthy documents [2, 8, 49]. Interactive systems have incorporated such summaries in new ways. For instance, some summaries have served as an entry point into documents [4, 14]. Other systems have brought the affordance of dynamic text expansion [6] to generated document summaries [14]. Relatedly, some systems allow parallel exploration of summary and source content, as in the case of linked video and crowdsourced summaries [44] and linked videos, captions, scripts, and plot summaries [43]. Others have allowed users to participate in the summarization process so that summaries better reflect individuals' needs [18, 53]. Our research complements this work with a tightly-focused exploration of a specific linking interaction—links from phrases in a summary to relevant passages in a source text—and how they impact reading experience.

Our work aligns with recent research for helping users verify the text generated by AIs. Prior work has offered new mechanisms for users to validate AI-generated content, for instance by visualizing AI assistance involved in writing a document [23], by presenting suggested edits and evidence to users [31, 36], or by warning authors when an LLM introduces new information within a document being edited [33]. Additionally, prior work has explored how to present many LLM responses at once by helping end users gauge similarities and differences across different outputs [17]. Hennigen et al. [21] describe a kind of traceable text interaction, where generated text contains explicit symbolic references to fields present within a data source. There is established precedent for incorporating reference links in existing AI systems. Many AI chat platforms, such as Perplexity AI [45], Microsoft Copilot [41], OpenScholar [3], and You.com [52], employ academic paper-style citations that link directly to external sources, enabling users to manually verify the model's responses. Some tools also offer additional features. For instance, the online playground for Olmo [19] uses phrase-level highlights to cross-check the documents from the training data that contain exact text matches with the model [37]. Additionally, Deep Research [42] provides a status bar that outlines intermediate steps in the output formulation process, allowing users to trace the model's logical reasoning and understand why specific references were selected. The above work sets the stage for our investigation into traceable text as a means to connect generated summaries and their source materials for the purposes of reader comprehension and trust. We explore the effect of a more specific instantiation of traceable text—phrase-level, summary-source linkages—on readers' experience with a generated text and its source.

3 TRACEABLE TEXT

3.1 Interaction design

This paper is broadly interested in generated texts that link back to source materials. We call these *traceable texts*. Our study focuses on one instantiation of traceable text: a generated summary that links its phrases to corresponding passages in a source text. Our instantiation triggers such links on hover: when the reader hovers over a phrase in the summary text, corresponding passages in the source text are highlighted (Figure 1). The goal of this design was to help readers verify content accuracy, identify potential hallucinations, and dig into the source text. Our prototype included bidirectional linking: when a modal key is held, the reader can hover over source passages to see corresponding passages in the summary. With our design, the method for checking for hallucinations involves looking at links from generated summary phrases. A hallucinated phrase will either not be highlighted in the first place (suggesting it may have no basis in the source text), or it will connect to a related source passage that could contain refuting evidence.

3.2 Design space

While our study focuses on one kind of traceable text, there are many ways to provide traceability. Here, we describe how interactive systems already or might provide traceability by articulating a design space of traceable text. We describe dimensions of traceable text, alternatives within those dimensions, and highlight where our prototype fits into the space (Figure 2).

Our work is not the first to describe a design space with this focus. For instance, Worledge et al. [50] plotted out a spectrum of text generations from extractive to abstractive, emphasizing the varying degrees to which each kind of generation supports attribution. We also draw inspiration (and our "forms of attribution" dimension) from a design space discussed in a talk by Shen et al. [46]. Below, we introduce a design space relevant to our framing of traceable text, along with examples of alternatives from the literature, and description of where our design fits in the space.

Dimension	Alternatives
Granularity	Phrase-level Sentence-level Passage-level Document-level
Modality of source	Text Data table Image
Invoking interaction	User-suggested AI-suggested Both
Form of attribution	Exact quote Paraphrase Interpretation
Comprehension scaffolds	Term definition Phrase simplifications Claim veracity
Arity of linkages	1-to-1 1-to-many many-to-1 many-to-many

Figure 2: A design space for *traceable text*. Alternatives selected for inclusion in our design are highlighted in gray.

Granularity. The first dimension of the design space is granularity. What size of chunks should the generated text be broken into before linking back to a source? Presumably, the finer grained the chunk, the more focused readers can be in gathering context in the linked source relevant to the chunk. Prior work has implemented chunking of links at the granularity of the phrase [21], sentence [16, 31], passage [14], and document [28, 35].

Modality of source. What mode of source information should the traceable text link to? In many cases, summaries are generated from source texts. As such, a traceable text might link into that source text (as in the case of [31]). When generated texts are based on other modalities of information (e.g., figures or tables from research papers, image descriptions), a user might benefit from links into figures or tables [21, 48] or, more generally, into images.

Invoking interaction. How do the traceability links come to be? Links could be created on-demand in response to a user request (e.g., [14]). They could also be generated by an AI [21, 31, 39, 48]—current production LLM systems such as SearchGPT, You.com, and Microsoft Copilot also employ AI-generated reference links. Additionally, systems might support both methods of linking [14, 34].

Form of attribution. How are the source and destination information related? Sometimes AI-generated content directly quotes the source material, so the generated text and source passage are the same (e.g. [15]). In other cases the AI-generated text is a paraphrase of the source which aims to retain the original meaning [4, 21, 31]. Finally, the AI-generated content could be an interpretation of the source text that potentially introduces new information [38].

Comprehension scaffolds. If the source content is difficult to understand in its own right, what scaffolds does the interface provide to help readers understand it? If the generated summary is simpler (as in [39]), the summary itself might serve as a scaffold. Additionally, sources could be annotated with term definitions [4] or in-situ support for claim verification [33].

Arity of linkages. Lastly, how much fan-out is in the links? Generated content could be based on more than one passage in the source. In that case, there could be 1-to-many links from part of the generated content to multiple source passages (e.g., [4, 23, 28, 31]). Many commercial LLM systems (e.g. Deep Research, You.com, Microsoft Copilot) link multiple papers to an AI response [41, 42, 52]. A simpler alternative is for each passage in a generation to link to one passage in the source (1-to-1, as in [16, 21, 23]). It could also be possible to provide many-to-1 linkages, where several generated passages draw upon a single source passage, as well as many-to-many linkages, where multiple generated passages correspond to several source passages.

Within this space, we focused on an instantiation of traceable text that we saw as likely to be generally useful: textual summaries generated from a source text. We provided phrase-level links because phrases contained the kinds of claims that we thought users would want to verify in the texts we focused on. Sites of linking were chosen by the AI because our tests showed promising accuracy in automated phrase selection (see Section 3.3).

3.3 Implementation

We implemented our traceable text with LLM prompts that identified summary-source links and with a React-based UI.

Prompting. To identify summary-source links, we used a chain of prompts. We first summarized the source text using a few-shot prompt [7, 47] to the OpenAI GPT-4 API [1]. Then, we prompted the LLM to segment the summary into phrase-level claims. We requested that these claims be individual verifiable statements, and that those claims be made as granular as possible. We provided few-shot examples to steer the LLM to identify as many claims in the summary as it could. Then, we prompted it once more to relate the claims to passages in the source text.[1]

Study interface. To support our study, we developed an interface that rendered our traceable text. In the interface, the source text and AI-generated summary were shown in side-by-side panels. Each summary claim and its corresponding source passage were wrapped in HTML spans. When the user hovered over a linked span, the corresponding passage in the other text would be highlighted. Our implementation was in React. It primarily involved programmatic text styling and hover-event handling. Since these features are available in most web frameworks, our implementation would likely be straightforward to port to other frameworks.

[1] All prompts appear in the supplemental material.

Preliminary analysis of accuracy. After we used the above approach to generate summaries and links, we asked clinical informatics fellows to review the accuracy of the phrase links (i.e., whether they reflected the source text). Accuracy was good, albeit not perfect. Of 159 summary-to-source links for 10 source texts, 129 were completely correct, 18 had "semantic issues," and 12 were incorrect. "Semantic issues" occurred when the summary claim lightly diverged from the semantics of the source passage.[2] More concretely, clinical informatics fellows were instructed to mark sections where the source text was accurately linked to the summary, but where they would phrase the linkage differently, and to explain their reasoning. For example, one annotator marked a "semantic issue" for a claim reading 'sore throat and cough' linking to 'rhinopharyngitis' (a condition that often—but not always—exhibits sore throats and coughs) in the source note. Incorrect links largely arose due to hallucinations in the generated summary. They also included instances where a claim linked to only part of the relevant content in the source (e.g., only 1 of 2 relevant source sentences).[3]

A pilot correction workflow. We piloted a correction workflow wherein we asked domain experts (clinical informatics fellows) to validate the generated summaries and linkages we used in our user study (presented to them as commented Microsoft Word documents), and then propose fixes. This was used to produce the validated summaries used in Phase 2 of the study. Validation of source/summary pairs typically took experts 10–25 minutes per pair.

4 STUDY

We conducted a usability study to address the following questions:

RQ1: How does a traceable text influence how readers notice hallucinations? We posited that participants would more consistently identify hallucinations in generated summaries when there were links to related claims in the source text.

RQ2: How do phrase links affect readers' ability to answer questions about the source text? We anticipated that the traceable text would help readers better answer questions about information in the summary and source note.

RQ3: How do readers use the traceable text? When is it used during reading? What do readers think it is useful for?

The documents at the center of our study were medical progress notes—that is, notes that clinicians take during visits with patients. We choose this document as it was relatable to study participants—it is a requirement in some countries that patients can access and read their notes to review details of their care [11]. Furthermore, progress notes are a document where hallucinations matter, as they risk misleading patients about their own health information. Finally, notes are a document where participants could bring documents of personal interest to them (i.e., their own notes).

4.1 Methods

We recruited 21 participants from several channels. Given the focus on medical documents, 16 patients were recruited from a university-affiliated patient and family advisory council. The remaining participants were recruited from academic mailing lists and social media posts. 19 of 21 participants provided demographic information. Of these, 70% identified as female and the rest (30%) as male. Ages ranged between 23 and 90 years old, with a median age of 60. When asked to report their occupation, 7 participants described themselves as retired, 4 as graduate students, 2 as educators, 1 as self-employed, 1 as a finance manager, 1 as an audio engineer. 2 reported that they were disabled and did not otherwise report their occupation. 1 participant did not report their occupation. Participants were compensated $35 USD. One participant did not complete the study and was excluded from quantitative analyses. Our procedure received ethics review from our institution's IRB and interviews were conducted over Zoom. The study followed a within-subjects design and consisted of three phases.

Phase 1: Summaries with hallucinations. Participants were asked questions about the source texts and accompanying summaries where they were made aware the AI-generated summary may contain hallucinations. For half of the questions, participants had access to the traceable text; participants were counterbalanced so that each task was performed with each interface by an equal number of participants. Questions were designed to be tricky, requiring the reader to notice and resolve disagreement between the summary and source text. Participants answered four questions total, each about a different source/summary pair.

Phase 2: Validated summaries and links. This phase was the same as Phase 1 (involving four new but similar questions). However, questions focused only on fully verified summaries and (when in the traceable text condition) links. The purpose of this phase was to understand the effect on reading when the generated text had no hallucination-related issues. Participants were told that all summaries had been expert-validated.

Phase 3: Unstructured reading time. Participants were given the remainder of the time (typically 5–10 minutes) to read one of their own medical notes alongside an AI-generated summary of their note using the traceable text.[4] This was intended to better reflect realistic contexts of use than Phases 1 and 2.

Questions. Questions were written to reflect those readers may have of their medical notes, including questions about care recommendations [27]. Answers were multi-select with 5 options each (i.e., all correct options needed to be selected). This was to incentivize readers to more closely review the summary and source text, rather than stopping after they found one answer in the text.[5] With

[2]We note that "semantic issues" may have had slightly varying meaning to different fellows. We asked fellows to look for them without defining "semantic issue" explicitly.

[3]The license for the n2c2 dataset prohibits us from distributing example notes. However, we are happy to share indexes of selected notes, and our summaries and questions, over personal correspondence with readers who have dataset use approval.

[4]We collected and processed participants' notes prior to the study session. All participants were made aware that their notes (with identifying and sensitive information manually removed) would be passed into a commercial large language model. All participants explicitly gave their consent for us to do this.

[5]An example question is: "Which tests has the doctor recommended to monitor the patient's blood pressure levels? Select all that apply." The answer choices are 'ECG requested', 'ETT requested', 'New values requested for LDL levels', 'HbA1c requested', and 'No tests requested'.

this design, participants had a $(1/2)^5$ probability of randomly guessing the right answer, or about 3%. Questions and answers were validated by medical experts (clinical informatics fellows).

Measures. For Phases 1 and 2, we measured the following:

- *Correctness* — A binary variable indicating whether the participant's response to the question was correct.
- *Speed* — The number of seconds taken by the participant to answer the question.
- *Difficulty* — A five-point Likert scale variable indicating the participant's self-assessment of the following prompt: "I found this task difficult."
- *Open Form Responses* — Responses obtained both from oral comments during the study and from open-ended questions in the final questionnaire.

Stimuli. Source texts were medical notes sampled from the n2c2 NLP Research dataset [32], specifically the 2014 corpus of patients managing coronary artery disease diagnoses. For Phase 1, summaries were selected that contained real hallucinations from GPT-4. Two kinds of hallucinations were represented, because they were the hallucinations that GPT-4 produced.[6] The first kind of hallucination consisted of contradictions between the summary and the source text, and the second kind were extrapolations (i.e., claims that were not necessarily incorrect, though not explicitly entailed by the source text). These represent both intrinsic and extrinsic hallucinations as they are termed in a review of hallucinations [25]. Phase 2 required expert-validated summaries and links. We recruited two medical experts—clinical informatics fellows, who are also board-certified primary care physicians—to perform this validation.

Analysis. We compared correctness with and without traceable text with the χ^2 test, and compared difficulty with the Wilcoxon unpaired two-sample test. Differences in time were analyzed by fitting mixed effects models, with reading interface, task number, and the interaction between them as fixed effects, and participant ID as a random effect. Statistical significance was assessed using F-tests with Holm-Bonferroni correction [22] and an α level of .05. RQ3 was answered following a single-author thematic analysis of notes taken during study sessions. In reporting results, we refer to participants with the pseudonyms P1—P21.

5 RESULTS

5.1 RQ1. Impact on noticing hallucinations

Participants more reliably noticed hallucinations in summaries when answering questions with traceable text. In tasks involving hallucinations, participants answered correctly 70% of the time with traceable text, and 12.5% without ($\chi^2 = 27.3$, $p = 1.8 \times 10^{-7}$). They answered questions significantly more quickly, with an average of 1.8 minutes with traceable text ($\sigma = 0.8$) compared to 2.9 minutes with the baseline ($\sigma = 1.0$; $F = 45.0$, $p = 3.64 \times 10^{-8}$). Despite answering questions more correctly, there was no statistically significant difference in self-reported difficulty ($W = 919$, $p = .24$).

[6]To characterize the kinds of hallucinations GPT-4 produces in summaries in this domain, we generated 34 summaries of medical notes, and asked experts (medical residents) to review them. They reported that 3 of the summaries contained fabrications—completely made up information—and 3 included extrapolations—where the model makes a (potentially correct) interpretation of what was in the source note, but that interpretation is not in the source note.

5.2 RQ2. Impact on answer correctness

Across all tasks (both with hallucinations and verified texts), participants answered questions significantly more correctly when using traceable text ($\chi^2 = 22.2$, $p = 2.4 \times 10^{-6}$). The difference was particularly pronounced for questions involving hallucinations (see Section 5.1). For the questions involving verified summaries, participants answered correctly 90% of the time with traceable text, and 75% without, though the difference was not statistically significant ($\chi^2 = 3.1$, $p = .08$). Participants answered questions significantly more quickly with traceable text, taking an average of 1.4 minutes with traceable text ($\sigma = 0.8$) versus 2.5 minutes with the baseline ($\sigma = 1.0$). This difference was statistically significant ($F = 95.4$, 1.2×10^{-16}). There was no statistically significant difference in the self-reported difficulty of completing tasks with traceable text versus the baseline ($W = 3553$, $p = .22$).

5.3 RQ3. The experience of using traceable text

Participants' impressions of traceable text were positive. In their Likert scale feedback after the unstructured reading task, 10 participants indicated strong agreement and 10 participants indicated agreement to the statement "I would use this interface to read my progress notes in the future." Based on their use in these tasks, participants described phrase links as "really cool" (P15), "actually really smart" (P5), and something they "absolutely love" (P10).

Functions of phrase links. Participants found several aspects of using phrase links valuable. First, the links served as an effective index into the note. Participants described links as providing "an entry point to the note" (P6) and something that made them "want to read the note a little more" (P8).

The links allowed them to "zoom in" on specific phrases within the source document and subsequently "zoom out" to the summary, facilitating an understanding of the note's broader context (P2). P2 described a situation in which they systematically clicked through all available links in the summary to "verify each statement" in the note. They emphasized how this helped them feel more confident in the summary's accuracy: "Especially when I don't recognize a term, I can see exactly where it came from. I know the summary isn't making stuff up." Participants also appreciated having links from technical terms in the source text to their simplified forms in the summary ($N = 4$). P6 highlighted how they could "zoom in" to the summary starting with the note: "I want to read the doctor's note first… but if I see a phrase I don't understand, I'll click the link to check how it's worded in simpler terms in the summary." One participant was even able to find an error in their source note using the phrase links (P15).

Navigation patterns. Some participants ($N = 6$) read notes with a *summary-first* approach, reading through the summary and using links to refer back to details in the source text. Some of these participants ($N = 3$) read through the linked passages exhaustively, exercising every link. For instance, P3 systematically used summary-to-source links to cross-check the summary's statements: "I needed to be sure the summary was correct. So I'd read a line, click the link, see if it lines up." In P2's words, this allowed them to "highlight which aspect of the progress note [each claim in the summary] came from." Conversely, participants who indicated that they were

more comfortable reading their medical notes ($N = 5$) followed a *source-first* approach, using links to clarify information in the note. This helped them understand otherwise confusing sections of the note (P10, P12). P10, for example, started from the source, and "whenever I hit a phrase I'm unsure about, I jump to the summary link to see a clearer explanation. It saves me from feeling stuck or having to Google on my own." Some participants ($N = 2$) used both summary-to-note and note-to-summary links, largely to ensure completeness and context alignment. For instance, P7 explained "going back and forth [between summary and source] helps me feel I'm not missing anything." Altogether, the study revealed that traceable text is an effective interaction for deepening engagement with the summary and source document.

6 LIMITATIONS AND FUTURE WORK

Limitations and risks. Our study represents a limited subset of tasks, involving just one kind of source text, and a narrow kind of question. Assessing the utility of traceable texts more broadly requires evaluation on a broader set of tasks of various domains and levels of complexity and other kinds of traceable texts. Additional evaluation is needed to explore more nuanced effects of the traceable text, such as learning and cognitive load. Our work does not examine adverse effects resulting from traceability features, like potentially disincentivizing complete reading of a source text. Additionally, relying on AI models to generate linkages could have negative effects—if linkages are incorrect, it could lead a reader to believe a summary claim represents an unrelated source passage, and thus overestimate the veracity of the claim.

Future Work. Our instantiation of traceable text is only one point in a broader design space for linking AI-generated summaries to their sources (see Section 3.2). Prior work has introduced other points in the space, and future work should continue to flesh out the space. As AI becomes more capable in understanding and decomposing other modalities (e.g., figures, tables, imagery), it may be possible to provide finer-grained linkages into sources of these other modalities. Additionally, future studies could explore whether the granularity of linked content (phrase, sentence, etc.) has any bearing on outcomes like retrieval speed, comprehension, and trust in the generated text. Finally, as retrieval-augmented generation approaches continue to develop, traceable texts might be developed to provide fine-grained linkages into external sources.

In domains where texts inform high-stakes decisions (like health), it may be critical that the correctness of a traceable text is verified. We piloted clinician-in-the-loop correction workflow in Section 3.3 that helped us to verify and fix traceable texts. However, at 10–25 minutes per text, it is likely not scalable. In the future, alternative semi-automatic or crowdsourced approaches may be necessary to generate verified traceable texts at scale.

7 CONCLUSION

In this paper, we discuss traceable texts. Traceable texts link from AI-generated texts back to the sources they are based on. We explore and evaluate an instantiation of traceable text, where an AI-generated summary of source text is augmented with phrase-level links to passages in the source text. A usability study showed this traceable text reduced the amount of time it took for readers to answer questions involving inspection of the summary and source text. This effect held particularly when the summary included hallucinations. When readers used traceable text on a personal source text, they found the links useful for clarifying sections of the source text. This study provides evidence of the usefulness of traceability links in helping to verify generated texts.

ACKNOWLEDGMENTS

This research was developed with funding from the Defense Advanced Research Projects Agency's (DARPA) SciFy program (Agreement No. HR00112520300). The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Nouf Ibrahim Altmami and Mohamed El Bachir Menai. 2022. Automatic Summarization of Scientific Articles: A Survey. *Journal of King Saud University-Computer and Information Sciences* 34, 4 (2022), 1011–1028.

[3] Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, et al. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199* (2024).

[4] 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 (2023).

[5] Sriram Karthik Badam, Zhicheng Liu, and Niklas Elmqvist. 2018. Elastic Documents: Coupling Text and Tables through Contextual Visualizations for Enhanced Document Reading. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2018), 661–671.

[6] Mark Bernstein. 2009. On Hypertext Narrative. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*. 5–14.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[8] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme Summarization of Scientific Documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 4766–4777.

[9] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 615–621.

[10] Tirth Dave, Sai Anirudh Athaluri, and Satyam Singh. 2023. ChatGPT in Medicine: An Overview of its Applications, Advantages, Limitations, Future Prospects, and Ethical Considerations. *Frontiers in Artificial Intelligence* 6 (2023).

[11] Tom Delbanco, Jan Walker, Jonathan D. Darer, Joann G. Elmore, Henry J. Feldman, Suzanne G. Leveille, James D. Ralston, Stephen E. Ross, Elisabeth Vodicka, and Valerie D. Weber. 2010. Open Notes: Doctors and Patients Signing On. *Annals of Internal Medicine* 153, 2 (2010), 121–125.

[12] Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. Discourse-Aware Unsupervised Summarization for Long Scientific Documents. In *Conference of the European Chapter of the Association for Computational Linguistics*.

[13] Tobias Esch, Roanne Mejilla, Melissa Anselmo, Beatrice Podtschaske, Tom Delbanco, and Jan Walker. 2016. Engaging Patients Through Open Notes: An Evaluation Using Mixed Methods. *British Medical Journal Open* 6, 1 (2016).

[14] 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*. Association for Computing Machinery, Article 145.

[15] 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*. 476–490.

[16] 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*. Article 842.

[17] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Article 838.

[18] Samira Ghodratnama, Mehrdad Zakershahrak, and Fariborz Sobhanmanesh. 2020. Adaptive Summaries: A Personalized Concept-based Summarization Approach by Learning from Users' Feedback. In *International Conference on Service-Oriented Computing*. Springer, 281–293.

[19] Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the Science of Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 15789–15809.

[20] Andrew Head, Amber Xie, and Marti A. Hearst. 2022. Math Augmentation: How Authors Enhance the Readability of Formulas using Novel Visual Design Practices. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Article 491.

[21] Lucas Torroba Hennigen, Shannon Shen, Aniruddha Nrusimha, Bernhard Gapp, David Sontag, and Yoon Kim. 2024. Towards Verifiable Text Generation with Symbolic References. In *Proceedings of the Conference on Language Modeling (CoLM) 2024*.

[22] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6 (1979), 65–70.

[23] Md Naimul Hoque, Tasfia Mashiat, Bhavya Ghai, Cecilia D. Shelton, Fanny Chevalier, Kari Kraus, and Niklas Elmqvist. 2024. The HaLLMark Effect: Supporting Provenance and Transparent Use of Large Language Models in Writing with Interactive Visualization. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Article 1045.

[24] Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2021. Summarization of Legal Documents: Where Are We Now and the Way Forward. *Computer Science Review* 40 (2021).

[25] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, Article 248 (2023). Issue 12.

[26] Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. 2023. Graphologue: Exploring large language model responses with interactive diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. Article 3.

[27] Hita Kambhamettu, Danaé Metaxa, Kevin Johnson, and Andrew Head. 2024. Explainable Notes: Examining How to Unlock Meaning in Medical Notes with Interactivity and Artificial Intelligence. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Article 449.

[28] Hyeonsu B. Kang, Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. 2023. Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, Article 43.

[29] Dae Hyun Kim, Enamul Hoque, Juho Kim, and Maneesh Agrawala. 2018. Facilitating Document Reading by Linking Text and Tables. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 423–434.

[30] Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An Empirical Survey on Long Document Summarization: Datasets, Models, and Metrics. *Comput. Surveys* 55, 8, Article 154 (2022).

[31] Kundan Krishna, Sanjana Ramprasad, Prakhar Gupta, Byron C. Wallace, Zachary C. Lipton, and Jeffrey P. Bigham. 2024. GenAudit: Fixing Factual Errors in Language Model Outputs with Evidence. *arXiv preprint arXiv:2402.12566* (2024).

[32] Vishesh Kumar, Amber Stubbs, Stanley Shaw, and Özlem Uzuner. 2015. Creation of a New Longitudinal Corpus of Clinical Narratives. *Journal of Biomedical Informatics* 58 (2015), S6–S10.

[33] 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*. Article 20.

[34] Shahid Latif, Zheng Zhou, Yoon Kim, Fabian Beck, and Nam Wook Kim. 2021. Kori: Interactive synthesis of text and charts in data documents. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 184–194.

[35] Yoonjoo Lee, Hyeonsu B. Kang, Matt Latzke, Juho Kim, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. PaperWeaver: Enriching Topical Paper Alerts by Contextualizing Recommended Papers with User-collected Papers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Article 19.

[36] Susan Lin, Jeremy Warner, J.D. Zamfirescu-Pereira, Matthew G. Lee, Sauhard Jain, Shanqing Cai, Piyawat Lertvittayakumjorn, Michael Xuelin Huang, Shumin Zhai, Björn Hartmann, and Can Liu. 2024. Rambler: Supporting Writing With Speech via LLM-Assisted Gist Manipulation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Article 1043.

[37] Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens. *arXiv preprint arXiv:2401.17377* (2024).

[38] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D. Gordon. 2023. "What It Wants Me To Say"': Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Article 598.

[39] 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*. Article 837.

[40] Raiza Martin and Steven Johnson. 2023. Introducing NotebookLM.

[41] Mircosoft. [n. d.]. *Copilot*. https://copilot.microsoft.com/

[42] OpenAI. [n. d.]. *Deep Research*. https://openai.com/index/introducing-deep-research/

[43] Amy Pavel, Dan B. Goldman, Björn Hartmann, and Maneesh Agrawala. 2015. SceneSkim: Searching and Browsing Movies Using Synchronized Captions, Scripts and Plot Summaries. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 181–190.

[44] Amy Pavel, Colorado Reed, Björn Hartmann, and Maneesh Agrawala. 2014. Video Digests: A Browsable, Skimmable Format for Informational Lecture Videos.. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. 573–582.

[45] Perplexity. [n. d.]. *Perplexity*. https://perplexity.ai/

[46] Shannon Zejiang Shen, Lucas Torroba Hennigen, Yung Sung Chuang, Ben Cohen-Wang, Linlu Qiu, Yoon Kim, and David Sontag. [n. d.]. *Generating Easily Verifiable Attributions with Large Language Models*. https://www.szj.io/assets/files/talks/2024-Nov-Verifiable-Attribution Accessed 11 March, 2025.

[47] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. 2023. LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2998–3009.

[48] 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)*. IEEE, 231–235.

[49] Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F. Rousseau, Chunhua Weng, and Yifan Peng. 2023. Evaluating Large Language Models on Medical Evidence Summarization. *NPJ Digital Medicine* 6, 1 (2023).

[50] Tatsunori Worledge, Theodora Hashimoto and Carlos Guestrin. 2024. The Extractive-Abstractive Spectrum: Uncovering Verifiability Trade-offs in LLM Generations. *arXiv preprint arXiv:2411.17375* (2024).

[51] Ryan Yen, Jiawen Stefanie Zhu, Sangho Suh, Haijun Xia, and Jian Zhao. 2024. CoLadder: Manipulating Code Generation via Multi-Level Blocks. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. Article 11.

[52] you.com. [n. d.]. *you.com*. https://you.com

[53] Xiaoyu Zhang, Jianping Li, Po-Wei Chi, Senthil Chandrasegaran, and Kwan-Liu Ma. 2023. ConceptEVA: Concept-Based Interactive Exploration and Customization of Document Summaries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Article 204.