PaperTrail: A Claim-Evidence Interface for Grounding Provenance in LLM-based Scholarly Q&A

Anna Martin-Boyle
mart5877@umn.edu
University of Minnesota
Minneapolis, Minnesota, USA

Cara A.C. Leckey
cara.ac.leckey@nasa.gov
NASA Langley Research Center
Hampton, Virginia, USA

Martha C. Brown
martha.c.brown@nasa.gov
NASA Langley Research Center
Hampton, Virginia, USA

Harmanpreet Kaur
harmank@umn.edu
University of Minnesota
Minneapolis, Minnesota, USA

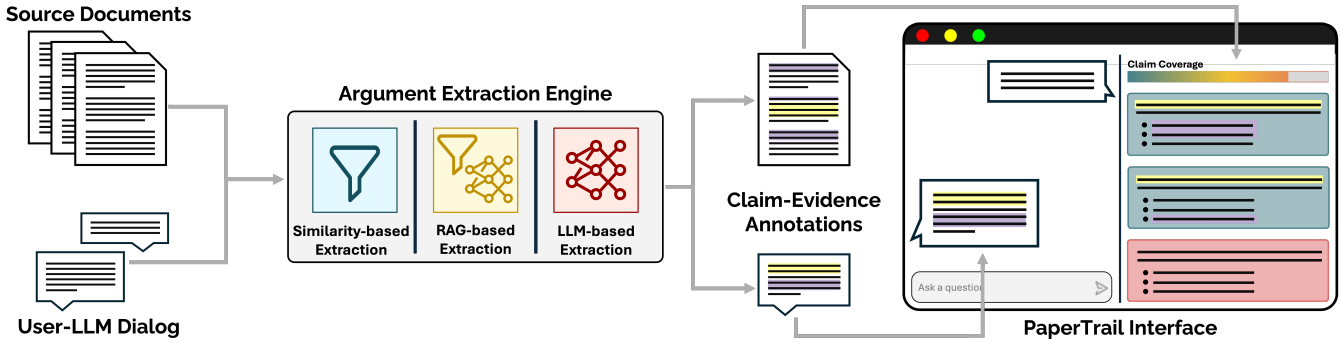

Figure 1: PaperTrail system architecture showing the three-stage pipeline for argument-grounded provenance. Paper PDFs and user-LLM dialog feed into the Argument Extraction Engine, which combines three different extraction methods deployed strategically across pipeline stages based on design-time tradeoffs between computational cost and semantic capability. The system produces claim-evidence annotations that power the PaperTrail interface for scholarly question-answering.

Abstract

Large language models (LLMs) are increasingly used in scholarly question-answering (QA) systems to help researchers synthesize vast amounts of literature. However, these systems often produce subtle errors (e.g., unsupported claims, errors of omission), and current provenance mechanisms like source citations are not granular enough for the rigorous verification that scholarly domain requires. To address this, we introduce PaperTrail, a novel interface that decomposes both LLM answers and source documents into discrete claims and evidence, mapping them to reveal supported assertions, unsupported claims, and information omitted from the source texts. We evaluated PaperTrail in a within-subjects study with 26 researchers who performed two scholarly editing tasks using PaperTrail and a baseline interface. Our results show that PaperTrail significantly lowered participants' trust compared to the baseline. However, this increased caution did not translate to behavioral changes, as people continued to rely on LLM-generated scholarly edits to avoid a cognitively burdensome task. We discuss the value of claim-evidence matching for understanding LLM trustworthiness in scholarly settings, and present design implications for cognition-friendly communication of provenance information.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; **Empirical studies in HCI**; • **Computing methodologies** → *Natural language processing*.

Keywords

Large Language Models, Provenance, Scientific Literature

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791101

ACM Reference Format:
Anna Martin-Boyle, Cara A.C. Leckey, Martha C. Brown, and Harmanpreet Kaur. 2026. PaperTrail: A Claim-Evidence Interface for Grounding Provenance in LLM-based Scholarly Q&A. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3772318.3791101

1 Introduction

The accelerating growth rate of scientific literature presents exciting opportunities for advancing knowledge, yet creates significant challenges as domain experts face escalating cognitive demands in monitoring and synthesizing knowledge within their fields [16, 36, 43, 44, 64]. Large language models (LLMs) have emerged as promising solutions for this information overload crisis [27, 45, 58]. LLMs are leveraged by end-users for a variety of scholarly tasks [51, 74], and are being integrated into scholarly question-answering (QA) systems like Semantic Scholar's "Ask This Paper"[111], JSTOR's AI research tool [48], and Elicit AI [126], as well as general search applications [121]. These tools promise to transform scholarly settings by automating synthesis, accelerating systematic reviews from months to hours, performing intelligent citation analysis, and identifying research patterns and gaps [91, 126]. Moreover, emerging LLM-based research agents like DeepResearch [4] and Google's AI Co-Scientist [37] are presented as able to autonomously conduct literature reviews and even generate research hypotheses.

The promise of automated research processes remains critically undermined by persistent limitations in the reliability of LLM-based systems. While LLM-augmented scholarly tools generate fluent and authoritative-sounding outputs, they inherit the same ethical issues and harms as their base models [12, 60, 125]. They particularly risk introducing errors due to their propensity to hallucinate information [42, 46, 75, 85]. These errors are difficult for scientists to detect [9], and persist even in Retrieval-Augmented Generation (RAG) systems, designed to be more reliable by grounding their generations in a corpus [87]. These errors carry particularly high stakes in academic contexts where domain experts require precise information for literature reviews, peer assessment, and research design. Finally, LLMs can spread misinformation [59, 127], spurring concerns that relying on ungrounded information systems will result in error propagation through scholarly discourse [29].

Predicting erroneous outputs and understanding the full capabilities of LLMs is difficult [33]. Current mechanisms for trust and reliance calibration in LLM outputs are limited, and offer insufficient affordances for a setting like scholarly QA. For example, source citations for attribution in LLM responses can improve perceptions of trust, but have been found to sometimes be hallucinated or inaccurate in representing the source material [19, 89, 120]. Similarly, uncertainty visualization methods like confidence highlighting can reduce over-reliance [15], but rich explanatory information often increases cognitive load [1] without providing actionable paths for evidence validation . Approaches combining sources with explanations are promising but remain vulnerable to an "illusion of explanatory depth" where users overestimate their understanding without actually verifying evidentiary support [20, 55]. However, fostering appropriate trust is not merely a matter of providing better information. Scholarly work occurs under time pressure and cognitive load—conditions that may prevent researchers from acting on their skepticism even when they recognize potential problems. Understanding how designs for trust and reliance calibration interact with these practical constraints is essential for developing effective scholarly AI tools.

In this work, we design a novel provenance mechanism for LLM outputs in scholarly QA settings, grounded in the argumentation structures inherent to scholarly work. We build on research that shows how structured representations of claims and evidence can improve interpretability [86, 104]. Argumentation structures offer promise for scholarly QA because they mirror academic discourse structures, where claims are supported by evidence connected through warrants; scientific papers inherently follow these argumentation patterns [38, 65]. Our system, PaperTrail, makes these implicit structures explicit through claim-evidence matching between a source document corpus and LLM responses to user queries for scholarly QA. It presents this provenance information via interface indicators that provide immediate visual feedback to the user by showing unsupported answer claims and omitted paper claims. Our design leverages domain experts' existing mental models of how scholarly arguments work, which has the potential to enable more efficient verification than generic explanation approaches.

We evaluate PaperTrail and our argument-grounded source provenance approach through a within-subjects user study involving 26 domain experts recruited from a research organization. We compare our setup against a baseline of source citations for attribution, common in commercial LLM-based search tools. Our participants complete two scholarly tasks where they are asked to edit LLM-generated text after a QA session with the LLM in question. We measure three outcomes of provenance information presentation: subjective trust perception [41], self-confidence in participants' revised outputs, and behavioral reliance quantified via normalized Levenshtein edit distance between the original LLM-generated text and the participants' edited versions.

Our results show that granular claim-evidence provenance information encourages more caution towards LLM outputs in scholarly settings. People trust in LLMs is significantly lower after using PaperTrail compared to baseline. However, this change does not translate to differences in perceived confidence or, importantly, changes in reliance behaviors. Experiential measures of usability and qualitative feedback suggest that, while helpful, this additional argument-grounding information clutters the interface and reduces usability, especially given the time constraints of a study setup. However, participants consistently appreciate the ethos of receiving this detailed breakdown of LLM arguments. We discuss the value of our argument-grounded source provenance approach for establishing trustworthiness of LLMs in this scholarly context, and discuss design implications for further reducing the cognitive load of presenting this additional information.

This work makes the following contributions to human-AI collaboration in information-intensive contexts:

- The design and implementation of PaperTrail, a novel system that operationalizes argumentation structures for source provenance by decomposing and linking claims and evidence between LLM-generated answers and source documents.
- A flexible backend architecture for claim-evidence extraction that can serve as an interactive tool and as a framework for evaluating LLM trustworthiness in scholarly settings.
- Empirical evidence from a within-subjects study showing the value of claim-evidence provenance in calibrating trust compared to standard source-citations from commercial LLMs.
- Empirical evidence of a trust-behavior gap in scholarly LLM use, showing that reduced trust alone is insufficient to change

reliance behaviors without addressing systemic constraints of time, usability, and cognitive resources.

2 Related Work

2.1 Evidence-Based Text Generation, Attribution, and Provenance

Recent surveys position LLM sourcing as a critically important design problem. Schreieder et al. [109] survey evidence-based text generation across 134 systems that use source material to ground model outputs in external evidence. Pang et al. [92] distinguish provenance at the levels of model authorship, model structure, training data, and external data, and separate prior-based approaches that embed explicit markers from posterior-based approaches that infer provenance from observed behavior. Within this broader sourcing landscape, Li et al. [71] focus on external-data sourcing for question answering, formalizing attribution as mapping each answer statement to one or more cited passages and evaluating methods in terms of citation coverage (recall) and sufficiency (precision).

Across these surveys, most systems attach provenance at document, paragraph, or sentence granularity and treat attribution primarily as a backend or benchmarking problem: how to retrieve better evidence, assign more accurate citations, or define more faithful automatic metrics. Coarse-grained citations lead to familiar failure modes such as granularity errors, mistaken synthesis across sources, and hallucinated statements when complex answers are supported only by flat sentence-level links [71], and little work studies how experts actually use provenance cues in context. In Pang et al. [92]'s terms, our system is a posterior external-data sourcing system: we do not modify model weights or embed watermarks, but instead infer and expose how an LLM answer relates to a fixed corpus of scholarly articles at interaction time. Building on the attribution formulation of statement-plus-citations [71], we instantiate provenance at a finer granularity by decomposing both papers and answers into discrete claims, aligning answer claims to paper claims and evidence snippets, and making omissions and mismatches explicit.

2.2 LLMs and Explanations

Prior work in human-centered AI asks not only what information LLMs should expose, but also how such cues can shape people's trust perceptions and reliance behaviors. Approaches towards transparency, including model reporting, evaluations, explanations, and communicated uncertainty, can be a means to support people in appropriate trust calibration, particularly when tailored to stakeholder goals and contexts [73]. Recent behavioral studies probe which cues actually foster appropriate reliance. Uncertainty cues were found to help reduce overreliance [54] and confidence highlighting were found to help users catch errors [117] in two recent lab studies.

It is important to consider which kinds of transparency impact behavior. Kim et al. [55] found that explanations overall tend to increase reliance on both correct and incorrect answers, while verifiable sources help reduce overreliance when the model is wrong and support appropriate reliance when it is right; they also identify "inconsistencies" as a distinct unreliability cue in LLM outputs [55]. In controlled reliance interventions, simple, persistent disclaimers can be effective, while token-level uncertainty cues or removing direct answers reduce overreliance but often at time costs and without reliably improving appropriate reliance [15]. For domain experts using retrieval-augmented generation (RAG) systems, surfacing sources and uncertainty interacts with users' verification practices and trust, which shows the importance of provenance cues in expert workflows [101]; such interventions can also assist domain experts in identifying confabulations in RAG-based systems [102].

2.3 Argument Structures in Textual Contexts

Our interface foregrounds claims and the evidence that supports them based on argumentation theory. In predictive-advice settings, structuring justifications with Toulmin [119]'s argument structure components (data, warrants, backings, and rebuttals) selectively strengthens distinct trusting beliefs, suggesting that showing what the claim is and why it might not hold can calibrate trust more effectively than undifferentiated explanations [104]. In scholarly writing specifically, argument structure underlies rhetorical function. Recent work augmented a scientific corpus with argumentative components and find that coupling argument extraction with rhetorical tasks in multi-task machine learning improves performance, and that argument components are most tightly linked to discourse roles [65]. We draw on this finding to propose that claims and evidence are the right unit to expose for scholarly sense-making.

We also build on work that evaluates argument-based explanations as user-facing artifacts. In medical QA, a recent study mined argument components and assessed explanation structure via graph patterns (e.g., missing premises, inconsistent support/attack) [86]. They found that people benefit when explanations are explicitly organized as arguments rather than free-form text. Beyond professional domains, an educational psychology study shows that recomposing arguments with Toulmin elements can measurably improve critical-thinking skills [99].

2.4 Scholarly Question Answering

Work on scholarly question answering (QA) clarifies both the task demands of answering research questions from papers and the interface signals needed for credible use. Scholarly corpora such as PubMedQA [47] and Qasper [23] establish that answering researcher-style questions requires reasoning over long, technical texts rather than factoids. More recent studies broaden the space to knowledge-graph QA [8], large-scale science QA [106], expert-authored long-form questions with attributed answers [81], multi-document / multimodal settings [70, 98], and exam-style free-response evaluation [26]. Scholarly QA evaluation show limitations of LLM-based QA systems. Long-context models still degrade with text distance [40], retrieval-augmented models can fabricate supporting evidence in science tasks [87], and domain experts judge model outputs as coherent yet inconsistently accurate [94]. Martin-Boyle et al. [84] developed an expert-derived schema identifying specific error types in scholarly QA, which goes beyond inaccuracies and hallucinations to describe issues with synthesis, formatting, question interpretation, and completeness. These performance issues motivate attribution-first interfaces that make evidence not only available to users, but also intelligible.

Argument structures are important to several works on Scholarly QA. Scientific claim verification shows the importance of aligning claims with cited evidence in open domains [122]. More recently,

SciClaimHunt introduces large-scale scientific claim-verification resources [61]. Such argumentation-forward resources show that scholarly discourse is naturally structured around claims, supports, and rebuttals [105], and that LLMs evaluated as science communicators can appear persuasive while remaining unreliable [9], which shows the importance of foregrounding verifiable sources. Finally, HCI perspectives urge centering domain experts' values and workflows in NLP tools [114, 128], and QASA [68] contributes a scholarly question taxonomy and full-stack reasoning setup that our work leverages. Our research extends this line of inquiry by looking at how a claim-evidence-based user interface affects trust perceptions and behavioral reliance during scholarly writing and critique.

3 PaperTrail: a Scholarly Source Provenance System

We implemented a novel system to investigate how argument-grounded provenance affects user trust and reliance in LLM-based scholarly QA. The system generates and displays claim-evidence structures from both a corpus of source documents and real-time LLM-generated answers. Below, we first describe our backend architecture, in particular our novel argument extraction and matching engine for determining source provenance; followed by our design goals and frontend interface.

3.1 Backend: Argument Extraction and Matching for Grounding Provenance

To generate the argument-grounded provenance annotations intended to help users appropriately calibrate their trust and reliance, we developed an Argument Extraction Engine that combines three approaches to claim and evidence extraction, each offering different tradeoffs between computational cost and semantic capability. **LLM-based extraction** leverages LLMs' semantic understanding to identify claims and evidence based on few-shot prompting, with their linguistic understanding offering a more nuanced interpretation of scientific discourse. While computationally expensive, this approach produces natural language representations that align with human categorization and can capture semantic meaning beyond surface-level text similarity [95]. **Similarity-based extraction** uses a sentence-transformer model and cosine similarity for rapid filtering and deduplication. This lightweight approach offers speed, interpretability, and reliability for high-volume operations where semantic nuance is less critical. **Retrieval-Augmented Generation (RAG)** combines retrieval efficiency with LLM semantic understanding by first filtering content using similarity search, then applying LLM processing to the reduced set. This hybrid approach reduces computational cost compared to applying LLMs to full documents by limiting the candidate set.

We deploy these methods at different pipeline stages based on stage-specific requirements: whether a user query is available to guide relevance filtering, whether processing occurs offline or in real-time, and the volume of text to be processed. The resulting three-stage pipeline consists of: offline paper-level extraction of claims and evidence for a source document corpus (Section 3.1.1); real-time answer-level extraction for LLM-generated scholarly QA answers (Section 3.1.2); and real-time claim-evidence matching to calculate argument-grounded source provenance between scholarly documents and real-time LLM answers in QA (Section 3.1.3). The pipeline is served by a Flask web server that provides a RESTful API, session management, interaction logging, and static content delivery. See Figure 2 for an illustration of the backend stages.

3.1.1 Stage 1: Paper-Level Claim-Evidence Extraction. We construct a corpus of scholarly documents structured into claims and evidence through one-time offline preprocessing, where computational cost is less constrained than in real-time operations. We preprocess each document by extracting the text using PyMuPDF v1.23.5,[1] followed by manual validation to ensure accuracy. We reviewed the preprocessed text against the original PDFs to check for the correctness of extracted mathematical notation; correct preservation of paragraph boundaries; and accurate handling of hyphenated words across line breaks. The plain text is programmatically segmented into sections and paragraphs to maintain contextual coherence and to allow for detailed claim extraction.

For claim extraction in this offline setting, we use **LLM-based extraction** where each paragraph is provided as context to Gemini 2.5 Pro [31] using a few shot prompt inspired by the work of Kumar et al. [61] and Toulmin [119]'s argumentation model. This prompt provides 10 examples of claims randomly selected from the SciClaimHunt dataset [61] and instructs the model to identify and extract distinct, verifiable scientific claims. Following Kumar et al. [61], we specify that claims should be *atomic*, *faithful*, and *decontextualized*. Drawing on Toulmin [119]'s argumentation framework, we additionally require claims to be *verifiable* (checkable against evidence) and *declarative* (statements rather than questions or method descriptions). Kumar et al. [61] validated these criteria through human annotation of 100 claims, achieving inter-annotator agreement of $\alpha = 0.69–0.78$ across dimensions. We adopted their methodology and qualitatively spot-checked claims extracted from one paragraph from each of the Introduction, Methods, and Results sections from each paper. These sections represent distinct discourse functions, where Introductions contain motivational and background claims, Methods contain procedural claims, and Results contain empirical findings; we wanted to verify that the extraction prompt handled this variation appropriately. This spot-check confirmed that extracted claims generally satisfied the five criteria from [61, 119] noted above. We cover details on our quantitative evaluation of this approach in Section 4 below.

We use LLM-based claim extraction here despite its computational cost for two reasons. First, without a query to inform relevance filtering, we must extract all potentially relevant claims from the corpus comprehensively. Similarity-based filtering or RAG approaches require a query as a reference point for relevance scoring. Second, the offline nature of this preprocessing step makes the computational expense acceptable, as it occurs once per document rather than in real time. While we could have eliminated the preprocessing step and only extracted paper claims in real-time using the user's query to identify relevant paper claims, this would introduce additional latency into every user interaction. Pre-extracting all claims allows us to more quickly identify relevant claims by searching an already-processed corpus. This shifts the computational burden from synchronous user-facing operations to asynchronous

[1] https://pypi.org/project/PyMuPDF/, GNU Affero General Public License

preprocessing. It also makes the interactions more consistent across users, because offline preprocessing creates a single ground truth. Additionally, pre-extraction enables reproducible results and allows the claim-evidence corpus to be versioned and audited independently of the real-time query processing pipeline.

Finally, in the evidence retrieval stage, we use **Similarity-based extraction**, where the claims extracted previously are used as queries to find their supporting evidence within the source text. Sentences exceeding a similarity threshold of 0.75 (based on guidance from Kumar et al. [61]) are considered candidate evidence for a given claim. To improve the readability of the extracted evidence, the preceding and subsequent sentences surrounding each evidence snippet are included as context to reconstruct coherent snippets. We use low-cost similarity-based extraction for evidence retrieval because this stage serves only as an initial filtering step to identify potentially relevant passages. The semantic relevance of this evidence to a user's specific information needs is later determined by LLM-based processing during the real-time claim-evidence matching phase (Section 3.1.3), where the user's query provides the necessary context for selecting the most pertinent evidence. At this preprocessing stage, we simply need to establish which text segments have any topical relationship to each claim—capturing passages that mention the same concepts, entities, or phenomena. This broad initial retrieval ensures comprehensive coverage while deferring the computationally expensive task of determining contextual relevance until it can be informed by the user's actual query.

The output of this preprocessing pipeline is a single JSON file containing a structured list of all paper claims including each claim's associated evidence, citation, and section name. This file is loaded by the backend server at startup and serves as the ground-truth knowledge base for PaperTrail.

3.1.2 Stage 2: Answer-Level Claim-Evidence Extraction. Real-time answer-level extraction involves two distinct models serving different roles: an **answerer LLM** that responds to user questions, and an **extraction LLM** that decomposes these answers into claims and evidence. This separation ensures that the answer generation remains focused on content quality while the extraction process maintains structural consistency with how paper claims were processed.

When a user asks a question during the interactive session, the **answerer LLM** first generates a complete answer based on the query, conversational history, and task context. In our study configuration, the answerer LLM receives the source documents as context alongside the query, similar to document-grounded question answering in commercial LLM interfaces where users upload PDFs and ask questions about their content. While this differs from RAG-based systems that retrieve relevant passages based on the query, PaperTrail's claim-evidence extraction and matching stages are model-agnostic and would function identically with RAG-generated answers. This model operates without additional structural constraints, which aligns the outputs to familiar commercial LLM chat experiences: mimicking the natural question-answering flow users encounter when uploading PDFs to commercial models and asking questions about their content. This separation between answer generation and claim extraction also makes our system model-agnostic; it is capable of analyzing outputs from any LLM rather than requiring a specific architecture or training approach. In standard RAG deployments, PaperTrail would function identically: the claim-evidence extraction and matching stages operate on the generated answer regardless of how that answer was produced. The key difference from typical RAG interfaces is granularity—while RAG systems often display retrieved passages as coarse-grained attribution, PaperTrail decomposes both the answer and source documents into discrete claims, which allows for verification of specific assertions rather than entire passages.

Once the answer is generated, the secondary **extraction LLM** performs the claim-evidence decomposition using the same definition of claims as the prompts used for paper-level extraction. In addition to the prompt, which is modified to also request supporting evidence to be identified for each claim, the LLM receives the complete answer text along with a JSON schema that enforces structured output (Google's Gemini API allows for a schema specification to be passed as an argument that dictates how the output should be structured, which guarantees usable information extraction). Since answers are much shorter than full paper texts, claim and evidence extraction occurs in a single pass without a separate evidence retrieval step. We use **LLM-based extraction** for this stage despite its computational cost because the context is significantly shorter than full papers, making the processing time acceptable for real-time interaction. Additionally, we can extract both claims and evidence in a single LLM call rather than requiring separate passes.

Following extraction, the system performs span annotation using the NLTK punkt tokenizer to segment the text into sentences. Then, a programmatic matching function locates the precise character positions of each claim and evidence piece identified by the extraction LLM, which maps the structured output back to specific text spans in the original answer. This lightweight post-processing step enables precise text highlighting required for the interactive user interface, where users need to see exactly which portions of the answer correspond to specific claims.

3.1.3 Stage 3: Claim-Evidence Matching. Our system matches the claims and evidence across the corpus of source papers and an individual LLM answer in QA using the structured claim-evidence representations extracted from the previous two steps. The goal of this stage is to improve the trustworthiness of the system by creating both global explanations (overall claim coverage) and local explanations (individual claim provenance) for the scholarly setting.

The matching process uses **RAG-based extraction** to identify relevant claims and evidence from the paper. In this RAG pipeline, the corpus consists of all paper-level claims and their supporting evidence extracted offline from source documents. The query is the user's asked question. Similarity-based retrieval first filters the complete paper claim corpus to identify candidates relevant to the user's question (using cosine similarity between the SPECTER embeddings of the question and each paper claim). The LLM is then prompted to select from the list of extracted claims the most relevant ones to the query. The same **RAG approach** is applied to evidence selection: for each relevant claim, its supporting evidence forms a sub-corpus that is searched using the query, with similarity-based retrieval filtering candidates and the LLM performing final selection of the most relevant evidence passages. We use **RAG-based extraction** in this step because the LLM provides

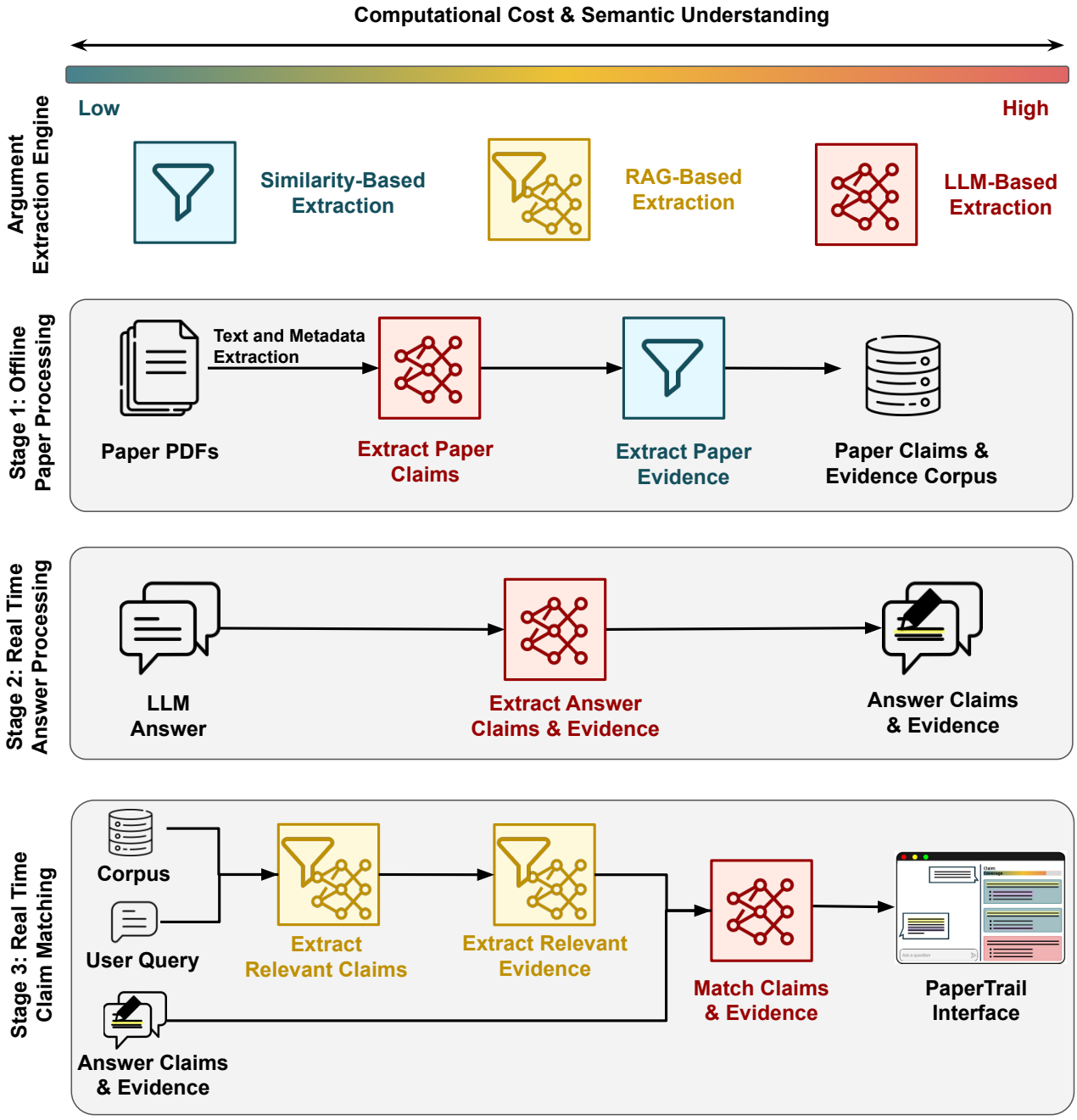

Figure 2: The Argument Extraction Engine (top) provides three extraction methods with different computational cost and accuracy tradeoffs. Colors follow a computational cost gradient: red indicates LLM-based extraction (highest computational cost), gold represents RAG-based extraction (medium cost), and teal denotes similarity-based extraction (lowest cost). We deploy these methods strategically across pipeline stages based on design-time considerations: (1) offline paper-level information extraction that preprocesses research papers into structured claims and evidence; (2) real-time answer-level extraction that decomposes LLM-generated answers into claims and supporting evidence; and (3) real-time claim-evidence matching that uses retrieval-augmented generation (RAG) to filter and align relevant paper claims with answer claims, producing source provenance indicators.

semantic capabilities to disambiguate between superficially similar but conceptually different claims and evidence, while the initial similarity-based retrieval dramatically reduces the search space from hundreds of claims/evidence to a manageable set of candidates.

Next, the LLM is prompted to perform claim-to-claim matching. The model is tasked with comparing the list of answer claims to the filtered list of relevant source papers' claims, and to find the most semantically equivalent pairs. This step benefits from the model's more nuanced language understanding to map the answerer LLM's generated assertions in scholarly QA to their ground-truth counterparts in the source literature, and avoids spurious connections that can occur when using simple cosine-similarity matching.

Finally, the evidence from the answer is verified using cosine similarity. To help users calibrate their trust and reliance on the system's output, we set a permissive cosine similarity threshold of < 0.55 to flag potentially unsupported evidence. This threshold value was selected through iterative testing on five held-out examples to satisfy two requirements: first, to identify answers that diverge substantially from the source material; and second, to avoid alert fatigue and acknowledging that relevant evidence in an answer may not be present in the filtered set of paper evidence. Our goal is to help users form accurate mental models of the system's capabilities [6] to support appropriate reliance and mitigate both overtrust and undertrust of the answerer LLM [67], and to manage user expectations of imperfection [57].

3.1.4 Implementation Details. Our system uses Gemini 2.5 Pro (gemini-2.5-pro-preview-05-06) for all LLM-based extraction and matching operations with temperature set to 1.0. We enforced structured JSON output using Gemini's response_json_schema parameter, which guarantees responses conform to our claim-evidence schema. We used the sentence-transformer model SPECTER [21] for all similarity-based operations, which we selected for its optimization on scientific text. Average end-to-end response latency was around 90 seconds per query, primarily due to sequential LLM calls in Stages 2 and 3. Full prompts are provided in Appendix A.

3.2 Frontend Interface

Our user interface is a three-panel web application designed to support scholarly question answering (QA) tasks (Figure 3). The **Left Panel (A)** contains the user's primary task workspace, the **Middle Panel (B)** contains the LLM-based scholarly QA chat, and the **Right Panel (C)** provides provenance information based on claim-evidence matching. This section presents our four design goals and then describes how these goals are instantiated in the interface components.

3.2.1 Design Goals. Our interface for PaperTrail is based on four design goals informed by challenges of source provenance in scholarly settings. Overall, these goals address tension between providing comprehensive provenance information while maintaining usability for domain experts, who are engaged in complex analytical tasks.

DG1: Support graduated cognitive engagement. Our goal is to structure access to provenance such that users can get a collective sense of the main provenance metric—number of claim-evidence matches—and dive into details as needed. This pattern builds on Shneiderman [113]'s visual information-seeking mantra: "overview first, zoom and filter, then details-on-demand". This graduated cognitive engagement is helpful for scholarly tasks, which often involve what Marchionini [83] calls "exploratory search": the scholar moves beyond fact retrieval to support deeper learning and investigation. In their foundational work on a cognitive task analysis of literature search, Pirolli and Card [96] outline several cognitive processes under foraging and sensemaking loops that help people gradually synthesize information. This design goal for our intended interface reflects their model by connecting their "shoebox" stage (the collected source papers) and "evidence file" stage (the extracted claims and evidence) [50, 96]. Our interface allows users to operate at different levels of the sensemaking process, from a high-level answer (overview) to claim-level annotations (zoom/filter) and finally to the source text itself (details-on-demand).

DG2. Minimize interaction overhead while preserving exploration depth. Source provenance inherently adds more elements to an already interaction-heavy interface, e.g., by providing more links to click and more information modalities to handle. To keep provenance manageable, we follow design principles for Coordinated Multiple Views (CMV) [103, 124]. Specifically, we adhere to Wang Baldonado et al. [124]'s guidelines by optimizing for space and time resources with a multi-pane layout, ensuring self-evidence by using brushing-and-linking to make relationships between claims and sources clear, and supporting attention management with selective information hiding. Following Pirolli and Card [96]'s information foraging cost structures, our aim is to provide comprehensive verification capabilities while reducing engagement burden. In our interface, automatic claim highlighting, structured claim-evidence decomposition, and focused interaction modes are designed to reduce the costs of scanning, recognizing, and selecting information, respectively.

DG3. Enable flexible verification workflows. Experts approach the same data using a variety of strategies [50], and opportunistically combine bottom-up and top-down processes based on emerging insights and verification needs. For example, an expert might "find a clue, and follow the trail" [50], a non-linear process that requires flexibility to support. This also aligns with principles of exploratory search, which emphasize iteration and discovery over linear lookup [83]. Our interface avoids enforcing a single verification pathway. Users can navigate bidirectionally between claims and evidence, which supports both top-down hypothesis checking (starting from answer claims to find supporting evidence) and bottom-up evidence discovery (exploring paper claims to identify gaps in the answer).

DG4. Align with scholarly mental models and human-AI interaction guidance. To be effective, an intelligent system must align with its users' existing knowledge structures and meet their expectations for interaction. Experts develop mental models of their domain that rely on structural and causal features rather than superficial ones [110]. Our system reflects inherent argumentative structures of scientific discourse [61, 65, 105, 122] by decomposing information into claims and evidence, so that it can leverage researchers' existing mental models.

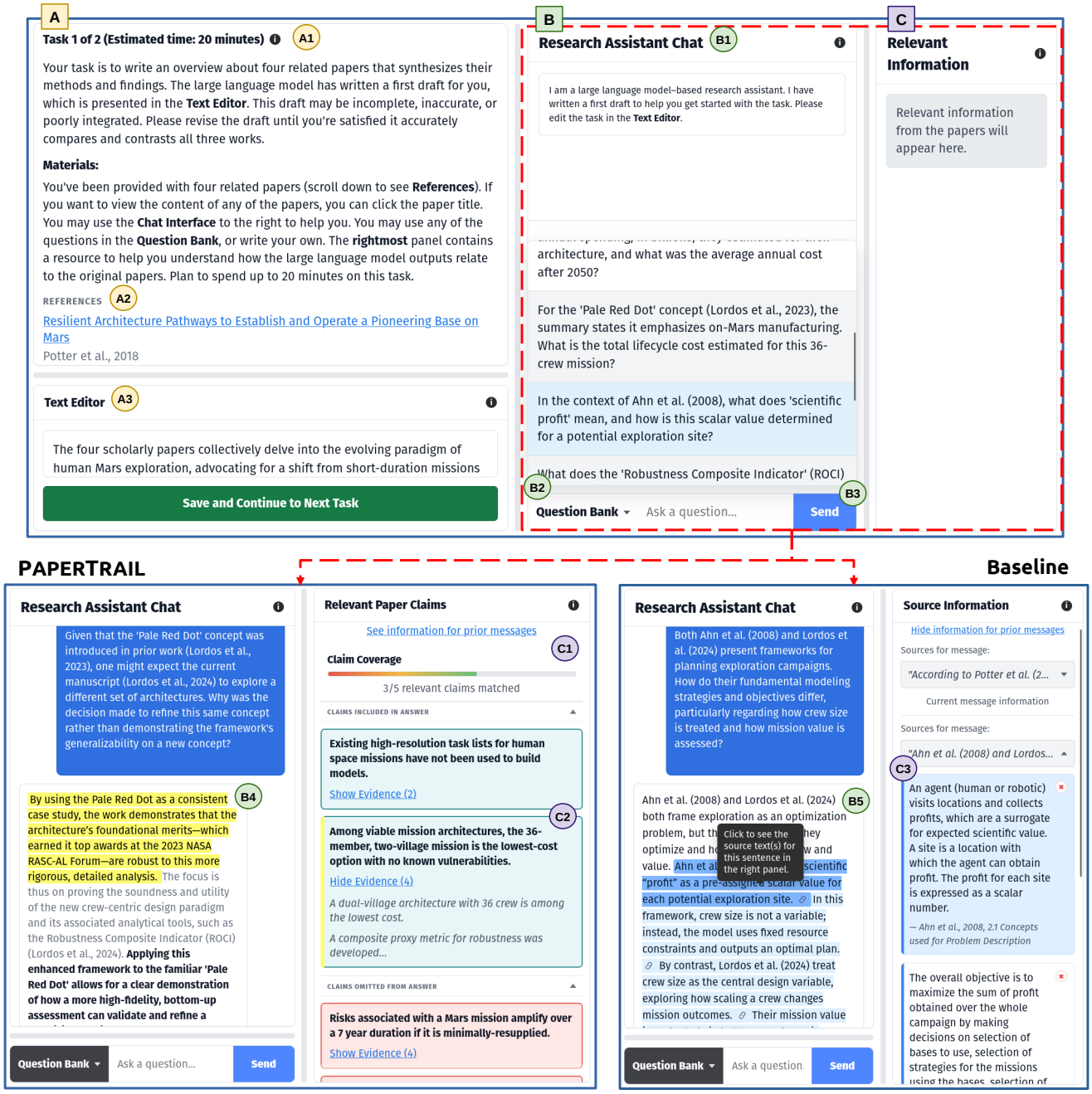

Figure 3: The user interface comprises three main panels. The general layout, shown at the top, consists of the Left Panel (A), the Middle Panel (B), and the Right Panel (C). The Left Panel (A) contains the user's main workspace, including the Task Context (A1), a References List (A2), and the Text Editor (A3). The Middle Panel (B) serves as the Chat Interface (B1), which includes a Question Bank (B2) and Chat Controls (B3). The Right Panel (C) displays information provenance, with its content changing based on the condition. The lower half shows the differences between the conditions. In the PaperTrail interface (left), the Middle Panel (B) shows interactive Answer Claims (B4). When a user clicks a claim, the corresponding Paper Claim (C2) is highlighted in the Right Panel (C), which also shows the overall Claim Coverage (C1) for the LLM's answer. In the baseline interface (right), the Middle Panel (B) contains a sentence-level Source Highlight (B5). Clicking this highlight surfaces the verbatim Paper Source (C3) text in the Right Panel (C).

3.2.2 Interface Layout and Features. The user interface is a single-page web application built with React and Redux Toolkit for state management. It presents a three-panel, resizable layout so users can keep multiple sources of information in view without context switching, a design decision aligned with **DG1** and **DG2** by maintaining continuity of attention across related elements [103, 124]. The panels provide the task context (**Left Panel (A)**), the chat interface (**Middle Panel (B)**), and the source provenance (**Right Panel (C)**). The panels are synchronized through a lightweight event bus, which allows brushing, linking, and coordinated navigation [103, 118]. Panel boundaries are resizable (horizontal and vertical splits where present), supporting DG3's flexible verification workflows. Components use nested scrolling so panes and in-panel sections can be scrolled independently, enabling DG1's graduated cognitive engagement. User selections auto-scroll linked panels to the relevant context, implementing DG2's principle of minimizing interaction overhead through coordinated views. Pervasive tooltips define sections and describe functionality, supporting DG1's details-on-demand pattern.

The **Left Panel (A)** contains the main task environment, which includes the **Task Context (A1)** and **Text Editor (A3)** . The **Task Context (A1)** provides an overview of the task instructions and an explanation of the resources available to the user, including a **References List (A2)** of source documents relevant for the task. This list of references is clickable and opens up a PDF viewer which superimposes the clicked reference PDF over the entire window. The left panel maintains proximity between the task description and the user's own writing space, the **Text Editor (A3)** , which reduces the cognitive burden of switching between context and text editing (**DG1**). The **Text Editor (A3)** contains a placeholder response to the scholarly task at hand generated by an LLM. Users can edit this text as they see fit, especially once they have used the scholarly QA interface in the **Middle Panel (B)** . The components within the panel are vertically resizable so that the user can proportionally enlarge the task description or the editor when needed, which supports flexibility (**DG3**): some users may rely heavily on the task prompt while writing, while others may minimize it entirely after an initial read.

The **Middle Panel (B)** comprises the **Chat Interface (B1)** where users can ask questions about the source documents used for the task (i.e., our scholarly QA setting), and use this information to edit the text in the **Text Editor (A3)** . We provide a **Question Bank (B2)** , a set of pre-written questions relevant for the scholarly context that the user can simply click and select to request a response (**DG1,DG2**). We also include **Chat Controls (B3)** , a Send/Stop control that becomes *Stop* while the system is waiting for the LLM response. This affords user control in deciding when to wait for an annotated LLM response with **Answer Claims (B4)** or switch strategies or queries (**DG3**). Users can read a generated response at a glance, or inspect more deeply using embedded annotations (**DG1,DG3**). Clicking on an answer claim engages the **Right Panel (C)** . All other text in the answer except for supporting evidence is grayed out to help the user focus on the selected claim (**DG2**).

The **Right Panel (C)** presents **Paper Claims (C2)** , which are claims from the paper that are relevant to the question asked; the supporting evidence for each claim is also provided (**DG4**). Paper claims which match with claims made in the answer are under "Claims included in answer" and are presented in teal cards. Paper claims that do not have a match in the answer are under "Claims omitted from answer" and are presented in red cards. Claims whose match has been selected in the **Middle Panel (B)** are highlighted using a vertical yellow bar (**DG2,DG3**). The right panel also contains a global overview **Claim Coverage (C1)** , a horizontal bar indicator that represents the number of relevant claims from the source documents included in the LLM's QA response (**DG1**). The indicator is colored from red to teal, and these colors correspond claim inclusion/exclusion colors of each claim-evidence provenance card. Each provenance card is linked to a specific message from the chatbot, creating a labeled, collapsible section for each turn so information from prior messages persists and can be revisited rather than disappearing (**DG2,DG3**). Within each section, cards can be expanded or removed to manage clutter. The interface applies selective information hiding and coordinated-view synchronization (brushing/linking with auto-scroll) [103, 113, 118, 124].

4 Offline Evaluation for Backend Approach

We conducted an offline evaluation to assess the quality of our claim extraction approach. No gold-standard benchmark exists for evaluating LLM-based claim extraction that generates a complete set of atomic claims (rather than extracting verbatim text spans) from scientific texts—existing datasets target either non-exhaustive sentence-level claim classification [2, 80] or claim verification [5, 61, 122]. Therefore, we adapted two related datasets to approximate an extraction evaluation.

4.1 Evaluation Datasets

4.1.1 SciClaimHunt. Kumar et al. [61] constructed a dataset of scientific claims generated from research paper paragraphs using few-shot prompting with Llama-2-13B, focusing on claims that could be verified against textual evidence rather than exhaustive extraction. Since the dataset provides paper-level but not paragraph-level claim mappings, we matched each claim to its source paragraph by first searching for exact string matches against sentences, then using SPECTER embeddings [21] to find the most similar sentence for remaining claims. We sampled 50 paragraphs with at least 3 associated claims, filtering for samples with sufficient annotation density to enable meaningful comparison with our exhaustive extraction approach. We used an additional 50-sample holdout set from SciClaimHunt while developing our procedure to avoid overfitting.

4.1.2 BioClaimDetect. Achakulvisut et al. [2] released a human-annotated dataset of abstracts in the biomedical domain with sentence-level claim annotations. Annotators labeled entire sentences as claims without decomposing them into atomic units. We randomly sampled 50 abstracts from their test set, and used the full abstract text as input and the annotated claim sentences as reference claims.

4.2 Evaluation Procedure

For each sample, we applied our claim extraction prompt using Gemini-2.5-Pro with 10-shot examples drawn from their respective datasets. We embedded both reference and extracted claims using SPECTER and computed pairwise cosine similarities. A reference claim was considered "matched" if any extracted claim exceeded a similarity threshold of $\tau = 0.9$; likewise for extracted claims matching references. This conservative threshold ensures matched claims are semantically near-equivalent rather than only topically related.

We define recall as the proportion of reference claims matched by at least one extracted claim, measuring coverage of benchmark claims; precision as the proportion of extracted claims matched by at least one reference claim, measuring extraction accuracy; and F1 (the harmonic mean of precision and recall). Note that neither dataset used for this evaluation provides exhaustive atomic claim annotations—no such gold-standard dataset exists. Therefore, we expect precision to be underestimated. That is, valid extracted claims in our more atomic approach may not match any reference claim simply because the reference set is an output of a less granular approach. Recall represents a more realistic metric of comparison, as we intend for our approach to provide coverage beyond the existing methods of claim extraction.

4.3 Validation Results

On SciClaimHunt, our pipeline achieved precision of 0.69, recall of 0.62, and F1 of 0.62. On BioClaimDetect, we observed precision of 0.63, recall of 0.88, and F1 of 0.72.

Our pipeline achieved high recall on BioClaimDetect (0.88), successfully recovering most human-annotated claims. The lower precision (0.63) reflects atomic decomposition: our pipeline extracts multiple fine-grained claims from sentences that annotators labeled as single claims. Manual inspection confirmed that most unmatched extracted claims were valid claims absent from the reference annotations rather than extraction errors. On SciClaimHunt, we observed higher precision (0.69) but moderate recall (0.62). The lower recall likely reflects noise in the SciClaimHunt reference set, which was generated by Llama-2-13B rather than human annotators. We developed our extraction approach using SciClaimHunt examples, so we included BioClaimDetect to test robustness on an out-of-distribution dataset that uses human annotations. The stronger performance on BioClaimDetect suggests our pipeline generalizes beyond its development dataset and aligns well with human judgment of what constitutes a claim.

5 User Study Design

We evaluated the efficacy of our argument-based source provenance system by comparing its use to a baseline interface that represented current LLM design for QA. This was done via a within-subjects user study with people in research-oriented roles in an organization (N=26). Participants experienced both interfaces across two tasks: a multi-paper synthesis (Task 1) and devil's advocate paper review (Task 2). The tasks were presented in a fixed order to control for task complexity progression, cognitive fatigue, and learning effects; interface condition order was randomized and counterbalanced across participants. Our study was classified as exempt from review by our ethics review board, and we pre-registered our hypotheses and methods on AsPredicted.[2]

5.1 Baseline Interface

The baseline interface differs from PaperTrail primarily in how provenance information is presented in the **Right Panel (C)**. While PaperTrail decomposes answers into discrete claims with matched evidence from source papers, the baseline interface uses a source citation approach that mirrors how commercial LLMs currently indicate provenance. In the baseline, when users click on **Source Highlights (B5)** in the LLM's answer, the corresponding verbatim **Paper Source (C3)** text appears in the **Right Panel (C)**. Given the prominent use of LLMs for scholarly QA now, with source links being the main form of explanatory information, our comparison to this baseline helps us measure if scholarly LLM use can be made more deliberate and tempered by introducing provenance details.

5.2 Task Design

5.2.1 Scholarly Writing Tasks. We designed two tasks that reflect common scholarly activities researchers encounter when engaging with literature. Each task requires participants to edit LLM-generated text, while using our scholarly QA system for verification and improvement. The initial texts were generated using Google's Gemini model to ensure realistic outputs with strengths and limitations characteristic of commercial LLMs. Moreover, this same model is used for LLM-based QA, maintaining consistency in content.

Task 1: Multi-Paper Synthesis. This task represents the process of comparing sources for the purpose of literature review. Participants are asked to edit an approximately 300 word LLM-generated draft statement that attempts to synthesize methods and findings from all four papers. The LLM-generated synthesis naturally exhibits common LLM characteristics, including potentially incomplete integration across papers, varying levels of technical detail, and possible gaps in comparative analysis.

Task 2: Devil's Advocate. This task simulates a pre-submission manuscript review, where researchers must anticipate and address potential reviewer critiques. Participants are asked to edit an LLM-generated draft defense of a paper positioned as their own manuscript, with the other three papers representing prior work. The 300-word LLM-generated statement defending the paper's novelty and soundness was a placeholder for the argument that this task required participants to verify and edit.

These tasks are informed by prior work categorizing how LLMs are used in scholarly settings. A recent survey study found that 81% of 816 researchers across multiple domains already incorporate LLMs in their research workflows, with literature review being among the most frequent use cases, aligning with our multi-paper synthesis task [74]. The second task (devil's advocate) reflects practices where researchers use LLMs to get critical feedback on their work and examine logical consistency in their manuscripts [51, p. 9].

5.2.2 Source Document Corpus. Next, we describe our document selection process for the scholarly QA tasks outlined above. We

[2] https://aspredicted.org/wp22-d58z.pdf

chose four peer-reviewed papers on the topic of Mars exploration and planetary surface operations to serve as our corpus of source papers that the participants must ask questions about. These papers were chosen according to several criteria. The topic needed to be broad enough to be accessible across many of the research disciplines at the organization we recruited from, as well as interdisciplinary, to introduce unfamiliar material across participants. Our selected research topic involves multiple domains in engineering, science, and human factors; it is also a topic of contemporary interest [10, 63, 112], which we hoped would make the task more compelling to participants. The topics and themes needed to be coherent across papers to enable synthesis tasks, while methodologies needed to be diverse to support critical comparison across papers.

We achieved this by first selecting two anchor papers based on publication in peer-reviewed venues recognized within the research area. These were recommended by two domain experts at our organization who confirmed the papers represent methodologically sound contributions. These papers had no authorship or acknowledged affiliation with our organization, to minimize the likelihood that participants had prior familiarity with them. From these anchor papers, we selected two additional papers from their reference lists that addressed the same topic but used different methodological approaches to ensure diversity in contribution types. Our final list was curated in collaboration with experts in that research area. The resulting corpus consists of of four papers that look at Mars exploration architectures from different perspectives and scales. "An Optimization Framework for Global Planetary Surface Exploration Campaigns" [3] presents an optimization framework that addresses the problem of selecting landing sites, routing decisions, and maximizing scientific return under resource constraints. "Resilient Architecture Pathways to Establish and Operate a Pioneering Base on Mars" [97] describes an architecture for establishing a Mars base supporting 50 people, including mission timelines, system requirements, and cost estimates. "Pale Red Dot: a Large, Robust Architecture for Human Settlements on Mars" [78] proposes a Mars settlement architecture for 36 crew members distributed across two villages. And "Leveraging Economies of Scale and Gains from Specialization for Robust Crewed Mars Architectures" [77] analyzes Mars missions with crew sizes from 4 to 63 members using a modeling approach that includes economies of scale and specialization effects. The papers were between 11 and 16 pages excluding references and appendices, and contained artifacts typical of scholarly papers and technical reports such as mathematical expressions, figures, and tables of quantitative information.

5.2.3 Question Bank. To reduce task burden and add some standardized interaction potential, we developed a question bank of prewritten questions for each task. We pre-generated answers to these questions, enabling system responses with no lag for provenance annotations. Participants could also write their own questions.

We grounded the development of these questions in the QASA framework [68], which categorizes questions that scholars write about papers into three types: **surface questions**, which "aim to verify and understand basic concepts in the content;" **testing questions**, which focus on "meaning-making and forming alignment with readers' prior knowledge;" and **deep questions**, which "ask about the connections among the concepts in the content and elicit advanced reasoning." Each type of questions is further categorized in subtypes with examples provided in the QASA paper.

We used Gemini to generate the question bank in the form of 1-2 questions for each relevant subtype. The prompt for Gemini included: (1) the QASA question type definitions, subtypes, and examples; (2) our task descriptions; and (3) the pre-generated texts that participants would edit. For Task 1 (Multi-Paper Synthesis), we used questions from the **testing** category: examples, quantitative comparisons, definitions, and compare/contrast questions. These question types support cross-paper synthesis by prompting participants to align information and make connections across sources. For Task 2 (Devil's Advocate), we used questions from the **deep** category: causal relationships, goals and motivations, procedural details, rationales, and expectations. These question types mirror how critical reviewers interrogate a manuscript's argumentation and methodology. We did not pre-write **surface** questions because they trigger basic fact-lookup that participants can easily perform on their own, and are misaligned with our aim to elicit the meaning-making and higher-order reasoning required by synthesis and critique.

5.3 Procedure and Flow

Participants completed a pre-study survey to indicate interest, share demographic information, and rate their familiarity with the source document corpus (too much familiarity was used as an exclusion criteria; Section 5.4). Sessions were then conducted remotely and asynchronously via a web browser, and included the following steps:

- Participants land on an About page that summarizes the study, provides them with contact information for the research team, and lists the study steps and their approximate durations. This page informs them that they may choose not to participate at any time and gathers consent.
- Next, participants complete the Trust in Explainable AI (TXAI) survey instrument [41]. Following Perrig et al. [93]'s recommendation, we exclude the reverse-coded question (item 6: "I am wary of the AI"), as doing so improves internal consistency. They are asked to think of an LLM-based system that they have used recently for scholarly tasks and to answer the questions specifically with that system in mind.
- After completion of the initial trust survey, participants view a tutorial that illustrates the interface they will use for Task 1.
- After viewing the tutorial, the participant is brought to the interface they will use for Task 1. This page includes the task environment, the chatbot, and the intervention specific to the current condition. The participant uses the interface and edits the drafted text until they are satisfied with the result. Our guidance was to aim for 20 mins to complete the task.
- After submission of the edited text, the participant takes four post-task surveys: (1) the TXAI scale [41], only being asked in the context of the LLM used, not the entire interface (with instructions and edits to TXAI scale items to clarify this); (2) a two-item confidence assessment, where they indicate on a scale of 1 to 7 their confidence in the edits they made ("I am confident in the edits I made to the text") and in the final text ("I am confident in the quality of the final text"); (3) the NASA-TLX scale to measure workload [39]; and (4) the SUPR-Q scale to evaluate the usability of the application [108].

- After completion of the post-task surveys, the participant is taken to a tutorial for the interface used for Task 2. They then perform the second task and complete the post-task surveys again with Task 2 details in mind.
- After completion of the second post-task surveys, the participants land on a final set of questions that ask them to indicate which system they preferred, and provide an optional free-response box for feedback on the systems. This is followed by a debrief page that explains the study in more detail.

5.4 Participants

We recruited people in research-oriented roles (both scientists and research engineers) across National Aeronautics and Space Administration centers—they were asked to express interest in participating in our study by completing a pre-screen survey. Participation in this study was voluntary, i.e., there was no monetary incentive provided for completing the study. 78 people completed our pre-screen, of which 74 met our inclusion criteria. We excluded participants who rated their familiarity with any of the papers in the corpus as greater than 3 out of 7, as these people would have an unfair advantage given their prior knowledge of the work. Of these 74 people who met our inclusion criteria, 38 ultimately participated in our study. The remaining 36 did not complete the study due to scheduling conflicts, time constraints, or lack of response to follow-up communications.

We report results from 26 participants after excluding data from 12 who did complete the study but their data was not valid. These 12 participants were excluded for the following reasons: (1) they took less than 10 mins to complete each task, without any interaction with the components of our systems; (2) they added gibberish text in their responses; and (3) their qualitative responses indicated that system latency had prevented them from engaging with the information (caused by load balancing issues in the backend). This exclusion process was conducted based only on feedback and task duration, without referencing our primary outcome measures, to avoid biasing the results.

The sample included 20 men and 6 women. Most participants (24 of 26) were between the ages of 25 and 54. The majority were highly experienced, with 18 participants having between 6 and 20 years of professional experience, and all but two held an advanced degree (14 Master's, 11 Doctoral). Regarding their familiarity with LLMs, 21 participants reported using LLM-based tools either daily or weekly. All participants had expertise in research and engineering in the sciences, including domains like chemistry, physics, materials science, and aerospace.

5.5 Measures

We use three primary measures to understand the outcomes of using PaperTrail compared to the baseline interface for scholarly tasks, as well as additional exploratory measures to understand participants' experiences with our system.

5.5.1 Primary Measures. Our system is intended to help people appropriately use LLMs for scholarly tasks. We measure aspects of this usage via three primary metrics: people's trust in LLMs after using the interfaces, their reliance on the placeholder LLM-generated text for the scholarly tasks, and their confidence in the LLM and their output.

Trust. We measure subjective trust using the validated TXAI scale [41] at three points: baseline (pre-study) and after each task. Items are rated on a 1-7 Likert scale and averaged. We use the pre-study trust value descriptively for contextualizing initial behaviors; only the trust values for individual conditions are used for comparisons. The scale items are asked in the context of the LLM used for the Q&A, rather than the other design features of the interfaces.

Reliance. We operationalize reliance based on changes to the LLM-generated placeholder text provided for the scholarly task. We calculate a token-level edit similarity based on Levenshtein distance [69] between the original LLM-generated draft (x) and the participant's final edited text (y), normalized by the token count of the longer token sequence:

$$\text{Reliance} \;=\; 1 \;-\; \frac{\mathrm{LD}\big(W(x),\, W(y)\big)}{\max\{|W(x)|,\, |W(y)|\}} \tag{1}$$

where LD is Levenshtein distance over tokens and $W(\cdot)$ maps a string to a sequence of lowercased lemmas after removing standard English stop words while retaining negations (*no, not, nor, n't*) to preserve polarity.[3] Normalizing by $\max\{|W(x)|, |W(y)|\}$ ensures large expansions or pruning register as reduced reliance. Values for reliance range from $[0, 1]$, with 1 indicating no edits (full reliance) and lower values reflecting greater rewriting. We chose token-level Levenshtein distance as our reliance measure because our research question concerns behavioral engagement with the text—specifically, whether participants physically edited the LLM output—rather than the semantic quality of those edits. Levenshtein distance directly captures editing actions: additions, deletions, and substitutions that participants made to the draft at the word-level, excluding minor edits (e.g., to stop words). Our measure treats all edit operations equally regardless of their semantic impact, which aligns with our goal of understanding whether provenance information changes low-level editing behavior.

Confidence. We measure participants' subjective confidence in each task output on two 7-point items ("I am confident in the edits I made to the text" and "I am confident in the quality of the final text." We average the two items to form a per-condition confidence score given internal consistency in the measures, calculated as Cronbach's $\alpha = 0.89$, 95% CI = [0.81,0.93].

5.5.2 Secondary Measures. We collect additional subjective and behavioral measures to contextualize primary outcomes. **Workload** is assessed post-task for each interface using the NASA-TLX questionnaire [39]; we compute the raw (unweighted) overall score as the mean of the five subscales (omitting the second item, "Physical Demand"). **Perceived usability** is assessed post-task with the SUPR-Q questionnaire [108] and averaged. **Interaction logs** include time-on-task; number of chat questions; and clicks on provenance features (e.g., source cards, claim/evidence items, focus toggles). We also record **interface preference** and **open-ended feedback** at the end of the study.

[3] Lemmatization collapses inflectional variants (e.g., *optimize/optimized/optimization*), and stop-word removal focuses the metric on content-bearing terms; operating at the word level aligns the unit of comparison with typical revision actions and reduces sensitivity to punctuation and formatting noise.

Measure	Baseline Mean	PaperTrail Mean	Statistic	Effect Size
		Primary Factors		
Trust	4.22 ±1.22	3.68 ±1.24	$t = 2.61,\ p = .015*$	0.44
Reliance	0.73 ±0.21	0.75 ±0.29	$W = 146.00,\ p = .313$	-0.23
Confidence	4.27 ±1.51	4.05 ±1.58	$t = 0.64,\ p = .525$	0.14
		Experiental Factors		
Cognitive Load	4.10 ±0.83	4.12 ±0.78	$t = -0.21,\ p = 0.838$	0.03
Usability	5.05 ±1.08	4.48 ±1.17	$W = 71.50,\ p = .026*$	0.52
Clicks	8.26 ±6.02	12.47 ±14.96	$W = 97.50,\ p = .137\cdot$	-0.35
Messages	3.67 ±2.57	4.08 ±2.84	$t = -0.79,\ p = .438$	0.16

t = Paired t-test, W = Wilcoxon signed-rank test, $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$, $^{\cdot}p = .1$

Table 1: Comparison of measures between baseline and PaperTrail conditions.

5.6 Analyses

We analyze our primary dependent variables using paired comparison tests between the baseline and PaperTrail data: t-tests when the distribution is normal, Wilcoxon signed-rank tests otherwise. Normality of the data is tested using the Shapiro-Wilk test. All statistical tests are two-tailed, with an alpha level 0.05. We also report comparison testing outputs for our exploratory experiential variables: cognitive load, usability, clicks, and messages; these are intended for descriptive purposes and to understand the differences in our primary measures. We also report the final interface preference as a frequency count. The qualitative data is coded using Braun and Clarke [17]'s inductive approach, with the intention to understand specific outcomes for what people liked, disliked, and wanted to add to our system design. We conduct Spearman correlation analyses between demographic, experiential, and primary outcome variables to understand relationships between measures and how they differ across interface conditions.

6 Results

This section presents the results of our within-subjects study. We first report results from comparisons between PaperTrail and baseline on our primary outcome and secondary experiential measures, then correlation analysis and qualitative themes to contextualize these numbers. Table 1 presents an overview of the descriptive and comparison outputs for all our measures.

6.1 Primary Measures: Trust, Reliance, and Confidence

A paired t-test showed a statistically significant effect of the interface on subjective trust on the LLM. Participants reported significantly lower trust in LLMs when using PaperTrail compared to the baseline ($t(25) = 2.61, p = 0.015$)), with a medium effect size (Cohen's $d = 0.44$). This supports our hypothesis that claim-evidence annotations encourage more caution towards LLM use in scholarly settings. However, despite the reduction in trust, we found no significant difference in behavioral reliance between the two conditions ($W = 146, p = 0.313$); nor in self-reported confidence ($t(25) = 0.64, p = .525$).

We propose three potential explanations for the discrepancy between the drop in trust and the non-significant change in reliance behavior for further consideration. First, the cognitive effort required to learn and navigate the novel PaperTrail interface within the limited time may have diverted participants' attention from the primary writing task, reducing the likelihood of extensive edits. Second, while PaperTrail's granular details made participants more skeptical of the LLM itself, the system's transparent design may have been perceived as more trustworthy overall, shaping reliance in a way that counteracted their caution towards the LLM. Finally, the interface may have had a bimodal effect, where it increased trust for some users who valued the verification features while decreasing it for others who were confronted with the LLM errors. However, coupled with the limited editing behavior of several participants given time constraints, this bimodal effect is lost in regression to the mean. We unpack these further in our qualitative findings below.

6.2 Secondary Experiential Measures

Our results show that the increased critical scrutiny afforded by PaperTrail came at the cost of lower perceived usability. A Wilcoxon signed-rank test indicated that the PaperTrail interface was rated as significantly less usable than the baseline ($W = 71.5, p = 0.026$), with a medium effect size ($r = 0.52$). We consider two possible explanations for these values. First, scholarly synthesis is an inherently demanding task, and this intrinsic difficulty may have been exacerbated by the feature-richness of PaperTrail. Indeed, prior work has often found this to be the case with rich explanatory outputs [11, 53, 100]. Second, our lab study did not afford participants enough time to move past the initial learning curve associated with the novel interface. In a real-world field deployment where users could develop expertise in the system over time, perceptions of usability might be different.

Click counts showed a marginally significant difference between conditions ($W = 97.50, p = 0.137$), with participants clicking more in PaperTrail ($M = 12.47, SD = 14.96$) than in the baseline ($M = 8.26, SD = 6.02$). However, higher click counts are ambiguous as an engagement indicator. More clicks could reflect deeper exploration of provenance information (the intended use), but could also

indicate: interface inefficiency requiring more actions to accomplish equivalent goals; confusion leading to exploratory clicking; or repeated attempts to understand unfamiliar features. The large standard deviation in PaperTrail clicks suggests highly variable engagement patterns across participants. Without click sequence analysis or qualitative observation of navigation patterns, we cannot say whether the additional clicks represent productive verification behavior or interaction overhead. We therefore interpret this finding cautiously. We return to this tradeoff between complex information presentation and usability in the qualitative results and Discussion. Other experiential measures—cognitive load and number of messages sent to the LLM—were similar across the two interfaces.

6.3 Correlation Analysis

To better understand some of our findings above, we used Spearman's coefficient to explore relationships between demographic, experiential, and primary outcome variables for the baseline (Table 2) and PaperTrail (Table 3).

One difference between the interfaces is the relationship between reliance and trust. In the baseline condition, reliance was significantly and positively correlated with trust ($r = 0.57, p < 0.01$) and usability ($r = 0.60, p < 0.05$). This suggests that the participants who trusted the LLM more and found the system more usable also tended to rely on the LLM output more heavily (i.e., edited the text less). In contrast, these correlations disappear in the PaperTrail condition. While the strong relationships between trust, confidence, and usability remained, none of these factors were significantly correlated with how much a participant chose to edit the LLM output. This suggests that the claim-evidence features in PaperTrail may have decoupled the simple relationship between a user's general trust in LLMs and their reliance on LLM-generated text. We suspect this is due to the same bimodal shift in behavior change as a consequence of reduced trust that is described in Section 6.1.

Additionally, we observed a significant negative correlation between education and confidence with PaperTrail ($r = -0.41, p < .05$) that was absent in the baseline. This suggests that more highly educated participants may have been more sensitive to the provenance information shown by PaperTrail, leading to reduced confidence in their outputs. We did not find any other significant correlations between participant demographics and our measures.

6.4 Qualitative Findings

Our qualitative analysis of participants' study feedback revealed three primary themes that help explain the quantitative results: external constraints that shaped reliance behaviors; tensions between information richness and usability; and paradoxical trust behaviors despite recognition of verification needs.

6.4.1 Factors Affecting Reliance Behaviors. **Time Pressure and System Performance.** The most frequently cited barrier to meaningful engagement was time constraints, mentioned explicitly by over half of the participants. Guiding participants to complete the tasks in around twenty minutes meant they had insufficient time to deeply engage with the provenance features or verify claims against source documents. This constraint was exacerbated by system latency, with participants describing response times as "excruciatingly slow," (P1) and "much slower than others [LLMs] that I have used" (P3). This led to many participants having a similar experience to participant 23, who "didn't make any edits to the text," and "[instead] asked a few questions to query and verify the accuracy of the responses." The combination of limited time and slow responses impacted how participants engaged with the provenance features, and participants noted this explicitly as well, like P15: "I feel the time constraints fundamentally changed the way I used and trust the AI output. I depended on the AI more because of the short time constraint and ultimately spent less time revising and checking the results than I otherwise would have."

Task Authenticity and Engagement Strategies. Participants approached the tasks with divergent strategies based on their framing of the study context. Some experts wanted to at least skim the papers on their own first before doing the task, and consequently expressed frustration with the artificial nature of evaluating less familiar papers, though in their research domain: "To produce more informative results, it would've been necessary for the participants to either familiarize themselves with the four papers before the test or to supply their own four papers with which they are thoroughly familiar" (P5). Conversely, some participants explicitly acknowledged the artificial nature of the study context and calibrated their engagement accordingly. As P2 explained, "I did not have to actually read any of the papers. I guess that means I trusted the AI tool to be accurate. I didn't believe everything it was telling me was accurate, but I figured it was close enough to complete the task at hand." This divergence in engagement strategies between those seeking ecological validity and those accepting the study's limitations likely influenced the variation in reliance scores, as participants' editing behaviors reflected their differing interpretations of what the task demanded.

6.4.2 The Information Complexity–Usability Tradeoff. Participants consistently recognized the need for access to complex information for scholarly tasks while struggling with its presentation. The quality of LLM outputs was frequently criticized, with one participant comparing it to "an eighth grader" (P25) and another to "what I would expect from a novice researcher placed in the same bind" (P4). Given their dissatisfaction with the LLM output quality, many participants expressed a desire for deeper engagement with the source materials, suggesting they would have preferred to read all papers thoroughly before attempting the tasks. Yet this desire for comprehensive review reflects an underlying tension in scholarly AI tool evaluation. While thorough paper familiarity might improve task performance in a study setting, such exhaustive pre-reading is specifically what these tools aim to help researchers avoid in practice, where the volume of literature makes reading every paper impractical. Overall, this dissatisfaction drove appreciation for the detailed provenance features, but the implementation created significant usability challenges.

Interface Complexity and Physical Constraints. Given their desire for richer, nuanced outputs, participants appreciated the theoretical value of the argument-based provenance information. In fact, we received some invitations to share our system and findings more broadly in the organization, noted by participants in their feedback responses to the study. However, the implementation revealed gaps between our design goals and user experience.

	Ed.	Exp.	Freq.	Gender	Conf.	Rel.	Trust	Cog. Load	Usability	Clicks	Messages
Demographics											
Age	0.47*	0.79***	0.19	-0.04	-0.03	0.12	-0.08	-0.29	-0.08	-0.09	-0.02
Education		0.53**	0.10	-0.16	0.07	0.18	0.02	-0.11	-0.05	-0.07	0.16
Experience			0.10	-0.20	-0.01	0.02	-0.01	-0.21	-0.10	0.04	0.14
Frequency				0.06	0.05	0.13	0.00	0.27	0.06	-0.06	0.27
Gender					-0.07	0.34	-0.10	0.16	0.03	0.13	0.05
Primary											
Confidence						0.24	0.60**	-0.08	0.47*	-0.13	0.07
Reliance							0.57**	0.21	0.50*	0.02	0.11
Trust								0.31	0.74***	0.11	0.30
Experiential											
Cognitive Load									0.29	0.26	0.47*
Usability										-0.10	0.22
Clicks											0.16

***$p < .001$, **$p < .01$, *$p < .05$

Table 2: Spearman correlations in the baseline condition. Education refers to level of educational attainment, Experience refers to years of experience, Frequency refers to frequency of AI use, and Clicks and Messages refer to the total counts of each.

	Ed.	Exp.	Freq.	Gender	Conf.	Rel.	Trust	Cog. Load	Usability	Clicks	Messages
Demographics											
Age	0.47*	0.79***	0.19	-0.04	-0.20	0.00	0.03	-0.18	0.23	-0.38	-0.28
Education		0.53**	0.10	-0.16	-0.41*	0.25	0.01	-0.15	-0.08	0.04	-0.24
Experience			0.10	-0.20	-0.29	0.22	-0.01	-0.12	0.07	-0.20	-0.23
Frequency				0.06	-0.19	0.11	-0.12	0.17	-0.09	-0.11	0.08
Gender					0.03	0.14	-0.29	0.13	0.05	-0.28	-0.11
Primary											
Confidence						-0.06	0.61***	-0.28	0.62***	0.08	-0.13
Reliance							0.06	-0.11	0.16	-0.11	-0.32
Trust								-0.27	0.73***	-0.03	-0.11
Experiential											
Cognitive Load									-0.28	0.22	0.04
Usability										-0.22	-0.32
Clicks											0.32

***$p < .001$, **$p < .01$, *$p < .05$

Table 3: Spearman correlations in the PaperTrail condition. Education refers to level of educational attainment, Experience refers to years of experience, Frequency refers to frequency of AI use, and Clicks and Messages refer to the total counts of each.

Our design goal DG2 aimed to "minimize interaction overhead while preserving exploration depth," yet participants found the interface to be "cluttered...for a fairly small laptop screen" (P25), with nested windows that allowed reading "only a line or two at a time." This directly contradicted our intention to optimize space/time resources. Similarly, while DG3 sought to "enable flexible verification workflows," the rigid presentation of claim-evidence cards actually constrained participants' verification strategies. The tension between appreciation and frustration can be seen in P24's feedback, who found the interface was "interesting and [having] value," but that it needed to "communicate this value more easily." The question bank partially addressed DG1's goal of "graduated cognitive engagement" by providing an entry point for exploration. As P18 noted: "I could not have done this without the suggested questions." However, this single success could not overcome the broader failure to achieve balance between complexity and usability. We consider this usability feedback in our Discussion of future design implications.

6.4.3 The Ethos of Grounding LLM Outputs in Provenance Information. Despite usability frustrations, most participants understood

and valued the project's underlying goals. While there were exceptions ("this interface/model completely misses how I use an LLM for research or paper writing purposes" (P21)), the majority recognized the critical need for verification capabilities in scholarly LLM tools. Participants articulated various aspects of this need: the importance of "fact checking" (P1); "trying to figure out what was missing or possibly incorrect" (P2); and evaluating "how credible or trustworthy the application's responses were" (P18). P6 framed it as an ethical concern, warning against "the temptation to use AI to write material for you," which they equated with "a general dumbing down of the world."

While usability issues limited positive experiences—only a small subset, like P3, "found the tasks easy to complete with the LLM QA interface provided"—participants nonetheless appreciated the conceptual value of argument-based provenance. P14's feedback articulated precisely what PaperTrail aimed to achieve:

> If more tools could be given to find the location of specific claims within the papers, that would be helpful. I find that errors often occur in LLMs through stripping context. To help with overall accuracy, every effort should be made to help the human user track down the original context to verify claims.

This recognition that LLM errors stem from "stripping context" and that verification requires tracing claims to their original sources supports the premise of claim-evidence provenance that we prioritized in PaperTrail.

7 Discussion

Our evaluation of PaperTrail shows that while argument-based provenance can encourage healthy skepticism toward LLM outputs, translating this into changed reliance behavior requires overcoming substantial barriers related to time, usability, and ingrained patterns of tool use. The relationship between attitudes and actions in human-AI collaboration is complex, particularly in time-constrained, cognitively demanding contexts like scholarly tasks. Significant usability costs of our implementation likely contributed to this trust-behavior gap. Below, we first contextualize our findings within prior work on trust and reliance for AI-assisted decision-making. Grounded in that argument, we present theoretical and design implications towards high-utility paths forward for our new scholarly AI assistance setting; and end with computational opportunities for improvement.

7.1 Trust-Behavior Gap: Why Didn't Lower Trust Change Reliance?

While our methodological and implementation constraints likely played a role in the trust-behavior change gap (details in Limitations below), our methods and findings share similarities with prior work on explainable AI (XAI) and AI-assisted decision making. These prior work contexts similarly show that explanations can increase reliance on both correct and incorrect predictions [55, 62]; enforced engagement with them can reduce user satisfaction [18, 24]; and their presentation does not often cultivate behavior change as people continue to defer to system outputs under time pressure or cognitive load [1, 53, 100]. We hypothesize three reasons for the consistent results across these contexts and ours.

First, people default to System 1 thinking—fast, automatic, and heuristic-based reasoning—even when they intellectually recognize the need for deliberation. The Dual Process theory of cognition distinguishes between System 1's efficient but error-prone processing and System 2's effortful analytical reasoning [32, 49]; also termed "bounded rationality" [115]. As XAI work has shown, making AI systems more explainable can sometimes exacerbate this problem by providing convenient narratives and analogic thinking devices that feel like comprehension without requiring genuine verification [22, 53]. Similarly, our claim-evidence interface attempted to engage System 2 thinking by requiring users to evaluate logical connections between claims and sources, but the cognitive cost of verification remained too high relative to the perceived benefits within the constrained study context.

Second, the intrinsic incentives that drive people away from System 1 thinking are missing in AI-assisted settings. Klein [56]'s reconciliation of naturalistic decision-making vs. heuristics-based outcomes identifies two necessary conditions for genuine intuitive expertise: high-validity environments (with stable regularities to learn) and opportunities to learn these regularities through feedback. AI-assisted settings may violate both conditions. In our case, LLMs produce errors unpredictably—plausible-sounding synthesis that may contain subtle omissions, over-generalizations, or unsupported leaps—making it difficult to develop stable heuristics for spotting problems. Moreover, people receive limited feedback on whether their edits successfully improved accuracy. Without this feedback loop, people cannot not learn when their skepticism should translate into action versus when selective reliance was appropriate, giving them no motivation to engage deliberately.

Third, designing systems that successfully shift cognitive behavior is a hard problem. PaperTrail represents one approach—introducing what has been called "design friction" [53, 88], "seamful design" [30, 52] or a "cognitive forcing function" [18] in XAI work to scholarly contexts, by requiring that people engage with claim-evidence structures before accepting LLM outputs. However, as both prior work and our findings demonstrate, such interventions face a difficult balancing act. If the friction is too burdensome, it hurts usability and people circumvent or abandon the system entirely [35, 53, 107]; our study feedback suggests we erred in this direction. If the intervention is too lightweight, it fails to engage System 2 thinking and people maintain their original behaviors. Prior work has even framed this as a cost-benefit analysis, seeking to understand when to prioritize different types of thinking: a challenge that persists [100].

If we anticipate some continued consistency in these two types of AI-assisted settings (prior work on XAI and AI-assisted decision making and our scholarly context), we can ground our implications in what would be different from the ideas explored before. While there remains tremendous potential for translating successes across these two contexts, we consider how to make progress on what is uniquely challenging for scholarly QA and writing via theoretical, design, and computational implications below.

7.2 Theoretical Implication: From Trust to Trustworthiness

What does it mean to study appropriate trust and reliance on LLMs? Is the goal to always lower trust? As we note above, this would not be ideal from a cognitive standpoint, and people might become so skeptical that they under-utilize a system. Appropriate trust is particularly challenging in a setting when the model is inherently black-box with no faithful representation of interpretability. We consider an alternate framing: making an LLM trustworthy rather than changing user trust.

7.2.1 Argument-Grounded Provenance as a Metric of LLM Trustworthiness. The backend architecture of PaperTrail can be extended beyond a user-facing tool into a formal trustworthiness benchmark for scholarly QA systems. Our method of using both local (claim-level matching) and global (claim coverage) information structures is a detailed way to evaluate LLM outputs that goes beyond surface-level metrics. This approach complements expert-derived error schemas [84] that categorize LLM failures in scholarly QA across dimensions including correctness, completeness, hallucinations, interpretation, and synthesis quality. Comparing answer claims against relevant source claims enables quantitative measure of both faithfulness and completeness. This approach is particularly effective for highlighting critical omissions, which is a difficult-to-detect failure mode in current systems. Such a benchmarking framework could enable systematic comparison of commercial LLMs to identify which are best suited for scholarly QA, support development of specialized scholarly LLMs through granular feedback, and allow researchers to audit LLMs before deployment.We present this argument-grounded provenance matching approach as an "application-grounded evaluation" [13, 28] for establishing LLM trustworthiness in scholarly settings, complementing existing methods while addressing the specific needs of scholarly users.

7.3 Implications for Design

Based on design lessons from prior work and our study, we consider the following implications to address how future systems might better balance the competing demands of information completeness and interaction fluidity.

7.3.1 Managing the Cost of Information Granularity with Adaptive Provenance. Our results show that detailed provenance encourages caution but can overwhelm users. Future systems should investigate adaptive provenance that dynamically adjusts detail levels based on context; for example, providing simple source links for straightforward questions but surfacing full claim-evidence interfaces for complex or controversial queries. The interface might default to simpler cues even in settings requiring greater complexity (perhaps just the claim coverage bar/indicator), and allow progressive disclosure of deeper structures on demand. Such adaptive systems would need to intelligently categorize query complexity, perhaps using the QASA framework's distinction between surface, testing, and deep questions [68], or borrowing from user-centered question banks of prior XAI work [7, 72, 116].

7.3.2 Scaffolding Critical Thinking Through Progressive Engagement. Claim-evidence interfaces can potentially act as effective scaffolds for critical thinking when users have sufficient cognitive resources. Scholarly information systems should progressively develop users' verification skills over time through tutorial modes that guide users through verification on simpler tasks before expecting independent critical evaluation on complex syntheses. This learning-based approach has shown promise in improving appropriate reliance for AI-assisted decision making [34]. Systems could also provide verification templates tailored to specific scholarly tasks—another approach with successful artifacts in improving responsible AI practices [25, 79]. Finally, ludic design elements [76] could reward thorough verification behaviors through what Nguyen [90] calls "epistemic playfulness," engaging with provenance through game-like challenges rather than purely for accuracy assessment. However, ludic elements must carefully encourage genuine exploration rather than superficial point-scoring.

A key difference between scholarly QA and prior work on AI-assisted decision-making is that appropriate reliance is conceptually different for these settings. Decision-making offers binary or multi-class constraints, while scholarly QA is open-ended—lacking ground truth and depending on question and user context. This necessitates combining designs from prior work with richer workflow analyses, which we hope future work will tackle.

7.4 Computational Considerations

Performing deep semantic analysis across large documents in real-time for every user query is computationally expensive, leading to the high latency that hurt user experience in our study. We consider some ways in which this can be improved in future iterations of our computational approach. Our backend demonstrates load-balancing through offline preprocessing, but we retained LLM processing in real-time for answer generation and claim matching (Stages 2 and 3), creating sequential bottlenecks. Each query required multiple LLM calls that could not be parallelized due to dependencies—answers must be generated before claims can be extracted, and paper claims must be filtered before matching. Future implementations could address these through aggressive caching (pre-computing common query patterns), parallel processing (extracting claims from answer chunks simultaneously), or faster inference infrastructure. The flexible extraction approach allows strategic selection of methods: similarity-based extraction for high-volume filtering, LLM-based extraction for user-selected claims requiring deeper analysis. This could extend to federated architectures where institutions pre-process document collections offline, sharing only lightweight claim indices for real-time matching [14, 82].

8 Limitations and Future Work

Task Realism. This took two forms: (1) participants engaged with unfamiliar papers within artificial time constraints, rather than conducting authentic literature searches or working with materials from their own research; and (2) the 20-30 minute task duration guidance, that prevented the deep engagement that characterizes scholarly work in practice. While this standardization was necessary for measuring reliance consistently across participants, it may have fundamentally altered engagement patterns. Future field studies should examine PaperTrail's effectiveness when researchers

use it for their actual work, with familiar literature and self-directed questions. Extensions of this experimental setting to include the writing elements of research reading (e.g., as done to generate workflow datasets in [66]), also deserve more attention, but was out of scope for our user study.

Metrics and Measurement. First, our operationalization of reliance through edit distance may not fully capture the nuanced ways participants engaged with LLM assistance. The measure conflates various behaviors (from wholesale acceptance to strategic delegation), and creates a potential confound. Participants might maintain high textual similarity not from overtrust in the LLM, but from trust that PaperTrail would alert them to problems requiring intervention. Future work should develop more sophisticated behavioral measures that distinguish between passive acceptance and informed delegation. Relatedly, when reporting subjective trust using the TXAI scale, participants may not even have distinguished between the interface and the LLM. Component-specific trust measures will be important for future verification. Finally, we did not evaluate the quality of participants' edited texts. While our focus was to capture behavioral differences, a quality assessment would help identify whether unsupported claims were removed, omitted information was added, and if the overall argumentation improved. Future work should include blind expert evaluation of output quality to complement behavioral measures.

Design Generalizability. PaperTrail instantiates one approach to claim-evidence provenance within a specific interface paradigm (three-panel layout with coordinated views). Alternative designs, such as inline annotations, progressive disclosure, or conversational verification, might produce different trust-reliance dynamics. Our findings speak to the value of argument-grounded provenance as a concept but not to the optimality of our particular implementation. Similarly, the basis of this design may be approached differently. Domains within and outside of STEM may have varying argumentation conventions, and other scholarly tasks may benefit from different provenance approaches. We do not capture rebuttals, warrants, backings, or qualifiers [119]; or other argumentation styles [123]. Future work should examine how argument-based provenance may be designed across diverse academic disciplines, argumentation details, and task types. Additionally, our system assumes source documents follow a clear claim-evidence structure. If source papers contain unsupported claims or lack explicit evidentiary reasoning, the extraction pipeline may produce incomplete or misleading provenance information. Users cannot easily distinguish whether an answer claim is flagged as unsupported due to a genuine LLM error or due to insufficient coverage in the source corpus—a limitation that may contribute to the lack of behavior change we observed.

Participant Pool. Our sample of researchers at a single organization represents a narrow slice of potential users. Scholars in different disciplines may have varying familiarity with structured argumentation (e.g., legal scholars vs. bench scientists), different verification norms, and different time pressures. Students, who are increasingly using LLMs for literature review [51], face distinct challenges around domain knowledge that our expert sample did not capture. Our consistency-oriented setup and findings can speak to the internal validity of our approach, but we leave this kind of evaluation of external validity to future work.

Long-term Adaptation. Our single-session study captured initial reactions to a novel interface. Trust and reliance patterns likely evolve as users develop mental models of system capabilities and limitations. Longitudinal deployment might reveal whether the trust-behavior gap narrows as verification workflows become habitual, or whether users develop stable patterns of selective engagement with provenance features.

9 Conclusion

We present a novel system, PaperTrail, that provides argument-grounded provenance information comparing LLM responses in a scholarly QA setting with claims and evidence from source documents. Through a within-subjects user study with 26 researchers, we found that PaperTrail significantly reduced participants' trust in LLM outputs compared to standard citation-based provenance. However, this did not translate into different editing behavior change—people edited LLM-generated text similarly across conditions. While people valued claim-level verification in principle, time pressure, system latency, and interface complexity prevented meaningful engagement with provenance features. Beyond the interface, our backend architecture for claim-evidence extraction shows promise as an evaluation framework for the trustworthiness of LLM scholarly outputs. The gap between recognizing verification needs and performing verification actions remains a challenge. Our findings indicate that granular provenance information alone is insufficient to change behavior when users face the time pressures and cognitive constraints typical of research settings.

Acknowledgments

We would like to thank our reviewers for their helpful comments. We are grateful to numerous colleagues at NASA including Braxton VanGundy and his team for their guidance on PDF text extraction, and Michael Steele and Charles Liles for their support in deploying PaperTrail, and Shanel Smith for facilitating participant recruitment. We are also grateful to Malik Khadar, Joel Markley, Matthew Zent, and all of the faculty and students in the GroupLens research lab for their feedback and support. We also want to thank everyone who participated in our study.

References

[1] Ashraf Abdul, Christian Von Der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[2] Titipat Achakulvisut, Chandra Bhagavatula, Daniel Acuna, and Konrad Kording. 2020. Claim Extraction in Biomedical Publications using Deep Discourse Model and Transfer Learning. arXiv:1907.00962 [cs.CL] https://arxiv.org/abs/1907.00962

[3] Jaemyung Ahn, Olivier De Weck, and Jeffrey Hoffman. 2008. An optimization framework for global planetary surface exploration campaigns. *Journal of the British Interplanetary Society* 61, 12 (2008), 487.

[4] Open AI. 2025. Introducing deep research. https://openai.com/index/introducing-deep-research/

[5] Carlos Alvarez, Maxwell Bennett, and Lucy Wang. 2024. Zero-shot Scientific Claim Verification Using LLMs and Citation Text. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, Tirthankar Ghosal, Amanpreet Singh, Anita Waard, Philipp Mayr, Aakanksha Naik, Orion Weller, Yoonjoo Lee, Shannon Shen, and Yanxia Qin (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 269–276. https://aclanthology.org/2024.sdp-1.25/

[6] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen,

Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300233

[7] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).

[8] Sören Auer, Dante A. C. Barone, Cassiano Bartz, Eduardo G. Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, Ivan Shilin, Markus Stocker, and Eleni Tsalapati. 2023. The SciQA Scientific Question Answering Benchmark for Scholarly Knowledge. *Scientific Reports* 13, 1 (04 May 2023), 7240. doi:10.1038/s41598-023-33607-z

[9] Prasoon Bajpai, Niladri Chatterjee, Subhabrata Dutta, and Tanmoy Chakraborty. 2024. Can LLMs replace Neil deGrasse Tyson? Evaluating the Reliability of LLMs as Science Communicators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 15895–15912. doi:10.18653/v1/2024.emnlp-main.889

[10] David Baneke. 2023. Big Astronomy: large telescopes and the dual narrative of impact. In *Big Science in the 21st Century*. IOP Publishing, 28–1 to 28–21. doi:10.1088/978-0-7503-3631-4ch28

[11] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. doi:10.1145/3411764.3445717

[12] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922

[13] Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for AI safety–a review. *arXiv preprint arXiv:2404.14082* (2024).

[14] Geoffrey Bilder, Jennifer Lin, and Cameron Neylon. 2015. Principles for Open Scholarly Infrastructures-v1. (2 2015). doi:10.6084/m9.figshare.1314859.v1

[15] Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 905, 23 pages. doi:10.1145/3706598.3714097

[16] Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications* 8, 1 (07 Oct 2021), 224. doi:10.1057/s41599-021-00903-w

[17] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa

[18] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (April 2021), 21 pages. doi:10.1145/3449287

[19] Courtni Byun, Piper Vasicek, and Kevin Seppi. 2024. This Reference Does Not Exist: An Exploration of LLM Citation Accuracy and Relevance. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, Su Lin Blodgett, Amanda Cercas Curry, Sunipa Dev, Michael Madaio, Ani Nenkova, Diyi Yang, and Ziang Xiao (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 28–39. doi:10.18653/v1/2024.hcinlp-1.3

[20] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 307–317. doi:10.1145/3397481.3450644

[21] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 2270–2282. doi:10.18653/v1/2020.acl-main.207

[22] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 352, 13 pages. doi:10.1145/3544548.3580672

[23] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 4599–4610. doi:10.18653/v1/2021.naacl-main.365

[24] Sander de Jong, Ville Paananen, Benjamin Tag, and Niels van Berkel. 2025. Cognitive Forcing for Better Decision-Making: Reducing Overreliance on AI Systems Through Partial Explanations. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–30.

[25] Wesley Hanwen Deng, Wang Claire, Howard Ziyu Han, Jason I. Hong, Kenneth Holstein, and Motahhare Eslami. 2025. WeAudit: Scaffolding User Auditors and AI Practitioners in Auditing Generative AI. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW521 (Oct. 2025), 35 pages. doi:10.1145/3757702

[26] Tu Anh Dinh, Carlos Mullov, Leonard Bärmann, Zhaolin Li, Danni Liu, Simon Reiß, Jueun Lee, Nathan Lerzer, Jianfeng Gao, Fabian Peller-Konrad, Tobias Röddiger, Alexander Waibel, Tamim Asfour, Michael Beigl, Rainer Stiefelhagen, Carsten Dachsbacher, Klemens Böhm, and Jan Niehues. 2024. SciEx: Benchmarking Large Language Models on Scientific Exams with Human Expert Grading and Automatic Grading. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 11592–11610. doi:10.18653/v1/2024.emnlp-main.647

[27] Simona Emilova Doneva, Sijing Qin, Beate Sick, Tilia Ellendorff, Jean-Philippe Goldman, Gerold Schneider, and Benjamin Victor Ineichen. 2024. Large language models to process, analyze, and synthesize biomedical texts: a scoping review. *Discover Artificial Intelligence* 4, 1 (19 Dec 2024), 107. doi:10.1007/s44163-024-00197-2

[28] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[29] Subhabrata Dutta and Tanmoy Chakraborty. 2023. Thus Spake ChatGPT. *Commun. ACM* 66, 12 (Nov. 2023), 16–19. doi:10.1145/3616863

[30] Upol Ehsan, Q Vera Liao, Samir Passi, Mark O Riedl, and Hal Daumé III. 2024. Seamful xai: Operationalizing seamful design in explainable ai. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–29.

[31] Gheorghe Comanici et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261 [cs.CL] https://arxiv.org/abs/2507.06261

[32] Jonathan St BT Evans and Keith E Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science* 8, 3 (2013), 223–241.

[33] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. 2022. Predictability and Surprise in Large Generative Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1747–1764. doi:10.1145/3531146.3533229

[34] Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. 2024. Going beyond XAI: A systematic survey for explanation-guided learning. *Comput. Surveys* 56, 7 (2024), 1–39.

[35] Susanne Gaube, Ekaterina Jussupow, Eesha Kokje, Jowaria Khan, Elizabeth Bondi-Kelly, Andreas Schicho, Felipe Campos Kitamura, Timo Koch, Timur Ezer, Jürgen Mottok, et al. 2024. Underreliance Harms Human-AI Collaboration More Than Overreliance in Medical Imaging. (2024).

[36] Ian D. Gordon, Debbie Chaves, Dylanne Dearborn, Shawn Hendrikx, Rebecca Hutchinson, Christopher Popovich, and Michael White. 2022. Information Seeking Behaviors, Attitudes, and Choices of Academic Physicists. *Science & Technology Libraries* 41, 3 (2022), 288–318. doi:10.1080/0194262X.2021.1991546 arXiv:https://doi.org/10.1080/0194262X.2021.1991546

[37] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. 2025. Towards an AI co-scientist. arXiv:2502.18864 [cs.AI] https://arxiv.org/abs/2502.18864

[38] Nancy Green. 2014. Towards Creation of a Corpus for Argumentation Mining the Biomedical Genetics Research Literature. In *Proceedings of the First Workshop on Argumentation Mining*, Nancy Green, Kevin Ashley, Diane Litman, Chris Reed, and Vern Walker (Eds.). Association for Computational Linguistics, Baltimore,

Maryland, 11–18. doi:10.3115/v1/W14-2102

[39] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. doi:10.1016/S0166-4115(08)62386-9

[40] Lukas Hilgert, Danni Liu, and Jan Niehues. 2024. Evaluating and Training Long-Context Large Language Models for Question Answering on Scientific Papers. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, Sachin Kumar, Vidhisha Balachandran, Chan Young Park, Weijia Shi, Shirley Anugrah Hayati, Yulia Tsvetkov, Noah Smith, Hannaneh Hajishirzi, Dongyeop Kang, and David Jurgens (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 220–236. doi:10.18653/v1/2024.customnlp4u-1.17

[41] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* Volume 5 - 2023 (2023). doi:10.3389/fcomp.2023.1096257

[42] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. doi:10.1145/3703155

[43] Debbie Chaves Ian D. Gordon, Brian D. Cameron and Rebecca Hutchinson. 2020. Information Seeking Behaviors, Attitudes, and Choices of Academic Mathematicians. *Science & Technology Libraries* 39, 3 (2020), 253–280. doi:10.1080/0194262X.2020.1758284 arXiv:https://doi.org/10.1080/0194262X.2020.1758284

[44] Michael White Ian D. Gordon, Patricia Meindl and Kathy Szigeti. 2018. Information Seeking Behaviors, Attitudes, and Choices of Academic Chemists. *Science & Technology Libraries* 37, 2 (2018), 130–151. doi:10.1080/0194262X.2018.1445063 arXiv:https://doi.org/10.1080/0194262X.2018.1445063

[45] Benjamin V. Ineichen, Marianna Rosso, and Malcolm R. Macleod. 2023. From data deluge to publomics: How AI can transform animal research. *Lab Animal* 52, 10 (01 Oct 2023), 213–214. doi:10.1038/s41684-023-01256-4

[46] 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. *ACM Comput. Surv.* 55, 12, Article 248 (March 2023), 38 pages. doi:10.1145/3571730

[47] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 2567–2577. doi:10.18653/v1/D19-1259

[48] JSTOR. 2024. JSTOR's AI research tool. https://www.about.jstor.org/research-tool/#about

[49] Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux* (2011).

[50] Youn-ah Kang, Carsten Gorg, and John Stasko. 2009. Evaluating visual analytics systems for investigative analysis: Deriving design principles from a case study. In *2009 IEEE Symposium on Visual Analytics Science and Technology*. 139–146. doi:10.1109/VAST.2009.5333878

[51] Shivani Kapania, Ruiyi Wang, Toby Jia-Jun Li, Tianshi Li, and Hong Shen. 2025. 'I'm Categorizing LLM as a Productivity Tool': Examining Ethics of LLM Use in HCI Research Practices. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW102 (May 2025), 26 pages. doi:10.1145/3711000

[52] Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. 2022. Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 702–714. doi:10.1145/3531146.3533135

[53] Harmanpreet Kaur, Matthew R Conrad, Davis Rule, Cliff Lampe, and Eric Gilbert. 2024. Interpretability gone bad: The role of bounded rationality in how practitioners understand machine learning. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–34.

[54] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 822–835. doi:10.1145/3630106.3658941

[55] Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 420, 19 pages. doi:10.1145/3706598.3714020

[56] Gary Klein. 2009. Conditions for Intuitive Expertise. *American Psychologist* 64 (09 2009), 515–526. doi:10.1037/a0016755

[57] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3290605.3300641

[58] Mario Krenn, Lorenzo Buffoni, Bruno Coutinho, Sagi Eppel, Jacob Gates Foster, Andrew Gritsevskiy, Harlin Lee, Yichao Lu, João P. Moutinho, Nima Sanjabi, Rishi Sonthalia, Ngoc Mai Tran, Francisco Valente, Yangxinyu Xie, Rose Yu, and Michael Kopp. 2023. Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. *Nature Machine Intelligence* (16 Oct 2023). doi:10.1038/s42256-023-00735-0

[59] Sarah Kreps, R. Miles McCain, and Miles Brundage. 2022. All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science* 9, 1 (2022), 104–117. doi:10.1017/XPS.2020.37

[60] Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 3299–3321. doi:10.18653/v1/2023.eacl-main.241

[61] Sujit Kumar, Anshul Sharma, Siddharth Hemant Khincha, Gargi Shroff, Sanasam Ranbir Singh, and Rahul Mishra. 2025. SciClaimHunt: A Large Dataset for Evidence-based Scientific Claim Verification. arXiv:2502.10003 [cs.CL] https://arxiv.org/abs/2502.10003

[62] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[63] W.H. Lambright. 2014. *Why Mars: NASA and the Politics of Space Exploration*. Johns Hopkins University Press. https://books.google.com/books?id=5id8AwAAQBAJ

[64] Esther Landhuis. 2016. Scientific literature: Information overload. *Nature* 535, 7612 (01 Jul 2016), 457–458. doi:10.1038/nj7612-457a

[65] Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Kai Eckert. 2018. Investigating the Role of Argumentation in the Rhetorical Analysis of Scientific Publications with Neural Multi-Task Learning Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 3326–3338. doi:10.18653/v1/D18-1370

[66] Khanh Chi Le, Linghe Wang, Minhwa Lee, Ross Volkov, Luan Tuyen Chau, and Dongyeop Kang. 2025. Scholawrite: A dataset of end-to-end scholarly writing process. *arXiv preprint arXiv:2502.02904* (2025).

[67] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. doi:10.1518/hfes.46.1.50_30392 PMID: 15151155.

[68] Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. 2023. QASA: advanced question answering on scientific articles. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 787, 17 pages.

[69] V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* 10 (Feb. 1966), 707.

[70] Chuhan Li, Ziyao Shangguan, Yilun Zhao, Deyuan Li, Yixin Liu, and Arman Cohan. 2024. M3SciQA: A Multi-Modal Multi-Document Scientific QA Benchmark for Evaluating Foundation Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 15419–15446. doi:10.18653/v1/2024.findings-emnlp.904

[71] Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023. A Survey of Large Language Models Attribution. arXiv:2311.03731 [cs.CL] https://arxiv.org/abs/2311.03731

[72] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376590

[73] Q. Vera Liao and Jennifer Wortman Vaughan. 2024. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review* (February 2024). https://www.microsoft.com/en-us/research/publication/ai-transparency-in-the-age-of-llms-a-human-centered-research-roadmap/

[74] Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X. Zhang. 2024. LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions. arXiv:2411.05025 [cs.CL] https://arxiv.org/abs/2411.05025

[75] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3214–3252. doi:10.18653/v1/2022.acl-long.229

[76] Craig A. Lindley. 2004. Ludic Engagement and Immersion as a Generic Paradigm for Human-Computer Interaction Design. In *Entertainment Computing – ICEC 2004*, Matthias Rauterberg (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 3–13.

[77] George Lordos, Madelyn Hoying, Lanie McKinney, Olivier de Weck, and Jeffrey Hoffman. 2024. Leveraging Economies of Scale and Gains from Specialization for Robust Crewed Mars Architectures. In *2024 IEEE Aerospace Conference*. 1–18. doi:10.1109/AERO58975.2024.10521341

[78] George C. Lordos, Madelyn Hoying, Yousif AlSadah, Liliana Arias, Ignacio Arzuaga Garcia, H Azzouz, John Beilstein, Wing Lam Chan, Ezra Eyre, Dane Gleason, Meltem Ikinci, Divya Iyer, Yuying Lin, Estelle Martin, Lanie G. McKinney, Duncan Miller, Cormac O'Neill, Omar Orozco, Palak B. Patel, Elizabeth Romero, Francisco Sepulveda, David Villegas, Alisa N. Webb, Kir Latyshev, Chloe Gentgen, Alexandros C. Lordos, Olivier L. De Weck, and Jeffrey A. Hoffman. 2023. *Pale Red Dot: a Large, Robust Architecture for Human Settlements on Mars*. doi:10.2514/6.2023-4776 arXiv:https://arc.aiaa.org/doi/pdf/10.2514/6.2023-4776

[79] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[80] Ian Magnusson and Scott Friedman. 2021. Extracting Fine-Grained Knowledge Graphs of Scientific Claims: Dataset and Transformer-Based Results. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 4651–4658. doi:10.18653/v1/2021.emnlp-main.381

[81] Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-Curated Questions and Attributed Answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 3025–3045. doi:10.18653/v1/2024.naacl-long.167

[82] Paolo Manghi, Alessia Bardi, Claudio Atzori, Miriam Baglioni, Natalia Manola, Jochen Schirrwagen, Pedro Principe, Michele Artini, Amelie Becker, Michele De Bonis, et al. 2019. The OpenAIRE research graph data model. *Zenodo* (2019).

[83] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (April 2006), 41–46. doi:10.1145/1121949.1121979

[84] Anna Martin-Boyle, William Humphreys, Martha Brown, Cara Leckey, and Harmanpreet Kaur. 2026. An expert schema for evaluating large language model errors in scholarly question-answering systems. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (Barcelona, Spain) *(CHI '26)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3772318.3791843

[85] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 1906–1919. doi:10.18653/v1/2020.acl-main.173

[86] Benjamin Molinet, Elena Cabrio, and Serena Villata. 2025. Assessing Argument-based Natural Language Explanations in Medical Text. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing* (Catania International Airport, Catania, Italy) *(SAC '25)*. Association for Computing Machinery, New York, NY, USA, 1007–1008. doi:10.1145/3672608.3707811

[87] Sai Munikoti, Anurag Acharya, Sridevi Wagle, and Sameera Horawalavithana. 2024. Evaluating the Effectiveness of Retrieval-Augmented Large Language Models in Scientific Document Reasoning. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, Tirthankar Ghosal, Amanpreet Singh, Anita Waard, Philipp Mayr, Aakanksha Naik, Orion Weller, Yoonjoo Lee, Shannon Shen, and Yanxia Qin (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 84–89. https://aclanthology.org/2024.sdp-1.8/

[88] Mohammad Naiseh, Reem S Al-Mansoori, Dena Al-Thani, Nan Jiang, and Raian Ali. 2021. Nudging through friction: an approach for calibrating trust in explainable AI. In *2021 8th International Conference on Behavioral and Social Computing (BESC)*. IEEE, 1–5.

[89] 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 (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 1325–1340. doi:10.1145/3715275.3732089

[90] C Thi Nguyen. 2022. Playfulness versus epistemic traps. In *Social virtue epistemology*. Routledge, 269–290.

[91] 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 (11 2021), 882–898. doi:10.1162/qss_a_00146 arXiv:https://direct.mit.edu/qss/article-pdf/2/3/882/1970740/qss_a_00146.pdf

[92] Liang Pang, Kangxi Wu, Sunhao Dai, Zihao Wei, Zenghao Duan, Jia Gu, Xiang Li, Zhiyi Yin, Jun Xu, Huawei Shen, and Xueqi Cheng. 2025. Large Language Model Sourcing: A Survey. arXiv:2510.10161 [cs.CL] https://arxiv.org/abs/2510.10161

[93] Sebastian A. C. Perrig, Nicolas Scharowski, and Florian Brühlmann. 2023. Trust Issues with Trust Scales: Examining the Psychometric Quality of Trust Measures in the Context of AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 297, 7 pages. doi:10.1145/3544549.3585808

[94] Denis Peskoff and Brandon Stewart. 2023. Credible without Credit: Domain Experts Assess Generative Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 427–438. doi:10.18653/v1/2023.acl-short.37

[95] Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A Prompt-based Topic Modeling Framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 2956–2984. doi:10.18653/v1/2024.naacl-long.164

[96] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis.

[97] Robert Potter, Sarag Saikia, and James Longuski. 2018. Resilient architecture pathways to establish and operate a pioneering base on Mars. In *2018 IEEE Aerospace Conference*. 1–18. doi:10.1109/AERO.2018.8396506

[98] Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2025. SPIQA: A Dataset for Multimodal Question Answering on Scientific Papers. arXiv:2407.09413 [cs.CL] https://arxiv.org/abs/2407.09413

[99] Hori Rashid, Nawras Khudhur, Yusuke Hayashi, and Tsukasa Hirashima. 2023. The Effect of Logical Argument Recomposition using Triangular Logic Model on Critical Thinking Compared to Conventional Method. In *Proceedings of the 2022 6th International Conference on Education and E-Learning* (Yamanashi, Japan) *(ICEEL '22)*. Association for Computing Machinery, New York, NY, USA, 220–226. doi:10.1145/3578837.3578869

[100] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 6, CSCW1 (2022), 1–22.

[101] Divya Ravi and Renuka Sindhgatta. 2025. Exploring Trust and Transparency in Retrieval-Augmented Generation for Domain Experts. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 254, 7 pages. doi:10.1145/3706599.3719985

[102] Philipp Reinhard, Mahei Manhai Li, Matteo Fina, and Jan Marco Leimeister. 2025. Fact or Fiction? Exploring Explanations to Identify Factual Confabulations in RAG-Based LLM Systems. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 274, 13 pages. doi:10.1145/3706599.3720249

[103] Jonathan C. Roberts. 2007. State of the Art: Coordinated & Multiple Views in Exploratory Visualization. In *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)*. 61–71. doi:10.1109/CMV.2007.20

[104] Eran Rubin and Izak Benbasat. 2023. Using Toulmin's Argumentation Model to Enhance Trust in Analytics-Based Advice Giving Systems. *ACM Trans. Manage. Inf. Syst.* 14, 3, Article 22 (June 2023), 28 pages. doi:10.1145/3580479

[105] Federico Ruggeri, Mohsen Mesgar, and Iryna Gurevych. 2023. A Dataset of Argumentative Dialogues on Scientific Papers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 7684–7699. doi:10.18653/v1/2023.acl-long.425

[106] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. ScienceQA: a novel resource for question answering on scholarly articles. *International Journal on Digital Libraries* 23, 3 (01 Sep 2022), 289–301. doi:10.1007/s00799-022-00329-y

[107] Lindsay Sanneman, Mycal Tucker, and Julie A Shah. 2024. An information bottleneck characterization of the understanding-workload tradeoff in human-centered explainable AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2175–2198.
[108] Jeff Sauro. 2015. SUPR-Q: A comprehensive measure of the quality of the website user experience. *Journal of usability studies* 10, 2 (2015).
[109] Tobias Schreieder, Tim Schopf, and Michael Färber. 2025. Attribution, Citation, and Quotation: A Survey of Evidence-based Text Generation with Large Language Models. arXiv:2508.15396 [cs.CL] https://arxiv.org/abs/2508.15396
[110] Robert M. Schumacher and Mary P. Czerwinski. 1992. *Mental models and the acquisition of expert knowledge*. Springer-Verlag, Berlin, Heidelberg, 61–79.
[111] Semantic Scholar. 2024. How are questions answered by Ask This Paper? https://www.semanticscholar.org/faq/how-are-questions-answered-by-paper-question-answering
[112] Matthew Shindell. 2023. *Planets: A History of Observing Worlds and Changing Worldviews*. Springer International Publishing, Cham, 1–21. doi:10.1007/978-3-030-92679-3_10-1
[113] B. Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*. 336–343. doi:10.1109/VL.1996.545307
[114] Dilruba Showkat and Eric P. S. Baumer. 2022. "It's Like the Value System in the Loop": Domain Experts' Values Expectations for NLP Automation. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 100–122. doi:10.1145/3532106.3533483
[115] Herbert Alexander Simon. 1997. *Models of bounded rationality: Empirically grounded economic reason*. Vol. 3. MIT press.
[116] Lars Sipos, Ulrike Schäfer, Katrin Glinka, and Claudia Müller-Birn. 2023. Identifying explanation needs of end-users: Applying and extending the XAI question bank. In *Proceedings of Mensch und Computer 2023*. 492–497.
[117] Sofia Eleni Spatharioti, David Rothschild, Daniel G Goldstein, and Jake M Hofman. 2025. Effects of LLM-based Search on Decision Making: Speed, Accuracy, and Overreliance. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1025, 15 pages. doi:10.1145/3706598.3714082
[118] John Stasko, Carsten Gorg, Zhicheng Liu, and Kanupriya Singhal. 2007. Jigsaw: Supporting Investigative Analysis through Interactive Visualization. In *2007 IEEE Symposium on Visual Analytics Science and Technology*. 131–138. doi:10.1109/VAST.2007.4389006
[119] Stephen E. Toulmin. 1958. *The Uses of Argument* (1 ed.). Cambridge University Press.
[120] Sebe Vanbrabant, Gilles Eerlings, Gustavo Alberto Rovelo Ruiz, and Davy Vanacken. 2025. ECHO: Enhancing Conversational Explainable AI through Tool-Augmented Language Models. *Proceedings of the ACM on Human-Computer Interaction* 9, 4 (2025), 1–33.
[121] Srinivasan (Cheenu) Venkatachary. 2024. AI Overviews in Search are coming to more places around the world. https://blog.google/products/search/ai-overviews-search-october-2024/
[122] David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-Open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 4719–4734. doi:10.18653/v1/2022.findings-emnlp.347
[123] Douglas Walton. 2010. Types of dialogue and burdens of proof. In *Computational models of argument*. IOS Press, 13–24.
[124] Michelle Q. Wang Baldonado, Allison Woodruff, and Allan Kuchinsky. 2000. Guidelines for using multiple views in information visualization. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (Palermo, Italy) *(AVI '00)*. Association for Computing Machinery, New York, NY, USA, 110–119. doi:10.1145/345513.345271
[125] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229. doi:10.1145/3531146.3533088
[126] Sharon Whitfield and Melissa A Hofmann. 2023. Elicit: AI literature review research assistant. *Public Services Quarterly* 19, 3 (2023), 201–207.
[127] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 436, 20 pages. doi:10.1145/3544548.3581318
[128] Anastasia Zhukova, Lukas von Sperl, Christian E. Matt, and Bela Gipp. 2024. Generative user-experience research for developing domain-specific natural language processing applications. *Knowl. Inf. Syst.* 66, 12 (Sept. 2024), 7859–7889. doi:10.1007/s10115-024-02212-5

A Prompts

This appendix provides the prompts and JSON schemas used in PaperTrail's Argument Extraction Engine. All LLM-based operations use Gemini 2.5 Pro with structured JSON output via the `response_json_schema` parameter. Similarity-based operations (Stage 1 evidence retrieval, Stage 3 evidence verification) do not require prompts.

A.1 Paper-Level Claim Extraction

Pipeline Stage: Stage 1 (Offline Paper Processing)

Purpose: Extracts atomic scientific claims from individual paragraphs of source documents during offline preprocessing. Applied to each paragraph to build the structured claim database serving as the ground-truth knowledge base.

Prompt:

> You are an expert research assistant specializing in extracting structured information from scientific texts. Your task is to carefully read the provided scientific paragraph and generate one or more core scientific claims based solely on the information present in that paragraph.
>
> A scientific claim must satisfy the following criteria:
>
> (1) **Atomic:** Focus on a single, specific, and indivisible assertion, finding, or conclusion. Avoid compound statements that can be decomposed further.
> (2) **Verifiable:** State something factual whose truthfulness can be checked against evidence or data, either within this paragraph or the broader scientific context.
> (3) **Faithful:** Accurately and precisely reflect the meaning and information given in the source paragraph. Do not introduce outside information or make inferences not directly supported by the text.
> (4) **Decontextualized:** Be understandable as a standalone statement, requiring minimal or no surrounding text from the original paper to grasp its meaning.
> (5) **Declarative:** Be a clear statement or assertion, not a question, hypothesis phrased as a question, or a description of methods or procedures.
>
> Based on these principles, transform the following paragraph into zero or more distinct scientific claims.
> {FEW-SHOT EXAMPLES: 10 paragraph-claims pairs randomly sampled from SciClaimHunt [61]}
> **Paragraph:** {PARAGRAPH}

JSON Schema:

```
{
  "type": "array",
  "items": {
    "type": "object",
    "properties": {
      "claim": {
```

```
        "type": "string",
        "description": "A single, atomic
            scientific claim"
      }
    },
    "required": ["claim"]
  }
}
```

Note: Evidence retrieval in Stage 1 uses similarity-based extraction with a cosine similarity threshold of 0.75 and does not require an LLM prompt.

A.2 Answer Generation

Pipeline Stage: Stage 2 (Real-Time Answer Processing)

Purpose: Generates responses to user questions during the scholarly QA session. The answerer LLM receives source documents as context alongside the query, similar to document-grounded question answering where source documents are provided as context. Citation tags enable sentence-level source attribution for the baseline condition and are removed before displaying responses to users.

Prompt:

> You are an advanced AI research assistant designed to assist users with scholarly literature analysis and question answering. Your primary function is to provide accurate and insightful answers to questions based on one or more scholarly papers provided to you.
> The user is performing an editing task and may refer to text they are editing. This text and a description of the editing task will be provided. The conversation history will also be provided if it exists.
> In your response, use tags around each sentence to indicate which paper(s) are being referenced. These tags will be removed in post-processing before the answer is shown to the user. Format: `<Author et al., year> sentence </Author et al., year>`. For sentences drawing on multiple papers, separate citations with semicolons: `<Author et al., year;Author et al., year>`.
> Provide your answer in 300 words or fewer. Do not use formatting such as bullet points or headers.
> **Papers:** {PAPER_CONTENTS}
> **Task description:** {TASK_DESCRIPTION}
> **Text being edited:** {EDITOR_TEXT}
> **Conversation history:** {CONVERSATION_HISTORY}
> **Question:** {USER_QUESTION}

A.3 Answer-Level Claim and Evidence Extraction

Pipeline Stage: Stage 2 (Real-Time Answer Processing)

Purpose: Decomposes LLM-generated answers into discrete claims and supporting evidence. Uses the same claim criteria as paper-level extraction to ensure consistency in matching. Extracts both claims and evidence in a single pass since answers are shorter than full papers.

Prompt:

> You are an expert research assistant specializing in extracting structured information from scientific texts. Your task is to carefully read the provided text and decompose it into discrete claims and their supporting evidence.
> A claim must satisfy the following criteria:
> (1) **Atomic:** Focus on a single, specific, and indivisible assertion. Avoid compound statements that can be decomposed further.
> (2) **Verifiable:** State something factual whose truthfulness can be checked against evidence or data.
> (3) **Faithful:** Accurately reflect the meaning in the source text. Do not introduce outside information.
> (4) **Decontextualized:** Be understandable as a standalone statement.
> (5) **Declarative:** Be a clear statement or assertion, not a question or description of methods.
> For each claim identified, extract the exact text spans from the input that express the claim and any text spans that serve as supporting evidence. Text spans must match the original input exactly to enable precise highlighting in the user interface.
> {FEW-SHOT EXAMPLES: text-to-claims-and-evidence pairs demonstrating the expected output format}
> **Text:** {ANSWER_TEXT}

JSON Schema:

```
{
  "type": "array",
  "items": {
    "type": "object",
    "properties": {
      "claim": {
        "type": "string",
        "description": "A single, atomic claim
            from the answer"
      },
      "claim_texts": {
        "type": "array",
        "items": {"type": "string"},
        "description": "Exact text spans
            expressing this claim"
      },
      "evidence_texts": {
        "type": "array",
        "items": {"type": "string"},
        "description": "Exact text spans
            supporting this claim"
      }
    },
    "required": ["claim", "claim_texts", "
        evidence_texts"]
  }
}
```

A.4 Relevant Claims Selection

Pipeline Stage: Stage 3 (Real-Time Claim Matching)

Purpose: Filters the corpus of pre-extracted paper claims to identify those relevant to the user's question. Operates on candidates

pre-filtered by similarity-based retrieval (using cosine similarity between SPECTER embeddings), applying LLM-based semantic reasoning to select the most pertinent claims.

Prompt:

> You are provided with a set of scientific claims extracted from scholarly papers and a user's question about those papers. Your task is to identify which claims are relevant to answering the question.
> A claim is relevant if it:
> - Directly addresses the question
> - Provides necessary background information
> - Contains factual information that would contribute to a complete answer
>
> Each claim is accompanied by a numerical ID. Return only the IDs of relevant claims.
> **Question:** {USER_QUESTION}
> **Claims:** {CLAIM_LIST_WITH_IDS}

JSON Schema:

```
{
  "type": "array",
  "items": {
    "type": "integer",
    "description": "ID of a relevant claim"
  }
}
```

A.5 Relevant Evidence Selection

Pipeline Stage: Stage 3 (Real-Time Claim Matching)

Purpose: Selects the most relevant evidence passages for each claim identified in the previous step. For each relevant claim, its supporting evidence forms a sub-corpus that is first filtered by similarity-based retrieval, then the LLM performs final selection of the most pertinent passages.

Prompt:

> You are provided with a set of evidence passages associated with a scientific claim, along with a user's question. Your task is to identify which evidence passages are most relevant in the context of the question.
> Evidence is relevant if it:
> - Directly supports or substantiates the claim
> - Provides data, results, or reasoning that validates the claim
> - Contains contextual information necessary to understand the claim in relation to the question
>
> Each evidence passage is accompanied by a numerical ID. Return only the IDs of the most relevant evidence passages.
> **Question:** {USER_QUESTION}
> **Claim:** {CLAIM_TEXT}
> **Evidence passages:** {EVIDENCE_LIST_WITH_IDS}

JSON Schema:

```
{
  "type": "array",
  "items": {
    "type": "integer",
    "description": "ID of a relevant evidence
        passage"
  }
}
```

A.6 Claim-to-Claim Matching

Pipeline Stage: Stage 3 (Real-Time Claim Matching)

Purpose: Identifies semantic equivalence between claims extracted from the LLM-generated answer and claims from the source papers. The output determines which answer claims are supported by the source literature (displayed as "Claims included in answer") and which lack grounding (flagged for user attention).

Prompt:

> You are provided with two sets of claims: (1) claims extracted from an LLM-generated answer, and (2) claims extracted from source scholarly papers. Your task is to identify which answer claims are semantically equivalent to which paper claims.
> Two claims are semantically equivalent if they express the same core assertion, even if worded differently. Minor differences in phrasing, specificity, or elaboration are acceptable provided the fundamental meaning is preserved. Do not match claims that are merely topically related but make different assertions.
> For each answer claim that has a match in the paper claims, return the answer claim ID paired with the ID(s) of the matching paper claim(s). Omit answer claims that lack a clear match.
> **Answer claims:** {ANSWER_CLAIMS_WITH_IDS}
> **Paper claims:** {PAPER_CLAIMS_WITH_IDS}

JSON Schema:

```
{
  "type": "array",
  "items": {
    "type": "object",
    "properties": {
      "answer_claim_id": {
        "type": "integer",
        "description": "ID of the answer claim
            "
      },
      "paper_claim_ids": {
        "type": "array",
        "items": {"type": "integer"},
        "description": "IDs of semantically
            equivalent paper claims"
      }
    },
    "required": ["answer_claim_id", "
        paper_claim_ids"]
  }
}
```

Note: Evidence verification in Stage 3 uses cosine similarity with a threshold of < 0.55 to flag potentially unsupported evidence and does not require an LLM prompt.