Understanding the Role of AI-Enabled Live Tables in Scholarly Sensemaking

Yudong Huang
yhuang143@hawk.illinoistech.edu
Illinois Institute of Technology
Chicago, United States

John Strus
jstrus@hawk.illinoistech.edu
Illinois Institute of Technology
Chicago, United States

Mark Roman Miller
mmiller30@illinoistech.edu
Illinois Institute of Technology
Chicago, United States

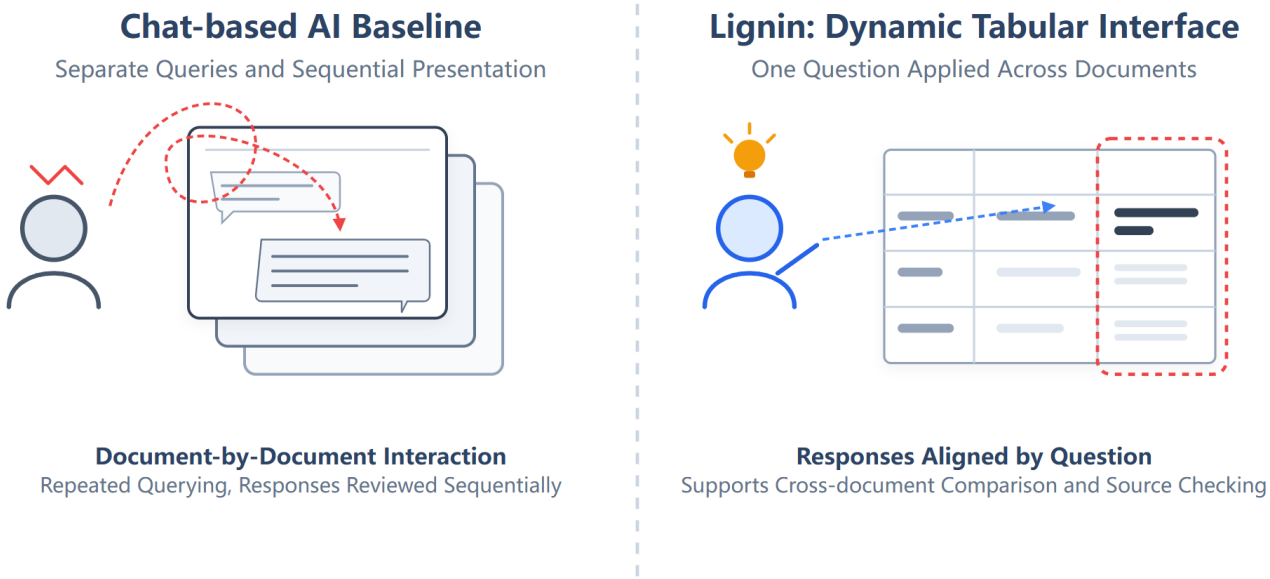


Figure 1: Interaction structures compared in the study. In the chat-based baseline, users query documents in separate conversational contexts and review the resulting responses sequentially. In Lignin, a user-defined question is applied across multiple documents, and the responses are aligned in a table to support cross-document comparison and inspection of source evidence.

Abstract

AI-assisted tools are increasingly used to support information extraction and multi-document reading in research workflows. However, conventional conversational LLM interfaces often mismatch researchers' specific needs, potentially inducing unnecessary cognitive load. To explore more effective interaction paradigms, we developed *Lignin*, an academic literature reading tool centered on tabular information representation and real-time AI querying. Building on existing table-based reading tools, Lignin combines dynamically configurable query columns, per-cell regeneration and editing, and source-level traceability. This design allows users to revise query dimensions as their questions evolve and supports cross-document comparison during scholarly sensemaking.

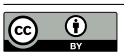
A controlled experiment was conducted to examine the impact of this tool versus traditional AI interaction modes on cognitive activities during exploratory tasks in unfamiliar domains. Through observations and interviews, we identified instances in which participants appeared to benefit from particular design features, clarifying how different information formats and interaction styles influence such tasks. These findings offer valuable insights for the design of future AI-assisted reading tools and the optimization of AI-mediated research practices.

CCS Concepts

• **Human-centered computing** → **Interaction paradigms**.

Keywords

Large Language Models, AI-assisted Research, Schema Theory, Cognitive Processes



UIST '26, Detroit, MI, USA

ACM ISBN 979-8-4007-2856-3/2026/11
https://doi.org/10.1145/3830398.3830689

ACM Reference Format:
Yudong Huang, John Strus, and Mark Roman Miller. 2026. Understanding the Role of AI-Enabled Live Tables in Scholarly Sensemaking. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26), November 02–05, 2026, Detroit, MI, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3830398.3830689

1 Introduction

With the increasing adoption of Large Language Models (LLMs), AI-assisted tools are becoming more common in academic literature search, information extraction, and exploratory reading. LLMs can lower the barrier to extracting information from documents and support query-driven interaction with research materials. These capabilities may reduce some of the effort involved in initial information retrieval. However, their suitability for more complex, multi-document reading activities remains an open question.

In early-stage academic literature review, literature reading is not an isolated extraction task, but rather a complex process of sensemaking and a prerequisite for the internalization of knowledge schemas. Current LLM interactions are predominantly based on linear, chat-based interfaces. While effective for certain tasks, this paradigm exhibits limitations in multi-document synthesis: linear information flows struggle to support non-linear comparison and cross-validation, leading to increased spatial memory burden and cognitive load when researchers handle conflicting sources or subtle terminological differences. Moreover, the immediacy of conversational interfaces often encourages users to focus on local details, at the expense of constructing a global perspective.

To address these challenges, this paper draws on sensemaking and schema theory to design and implement *Lignin*, a multi-document reading tool centered on dynamic tabular representations. Applications utilizing LLMs to transform unstructured or natural language information into structured data are not uncommon [1, 5, 15]. Building on existing table-based approaches, Lignin combines dynamically configurable query columns, per-cell regeneration and editing, and source-level traceability. Each row represents a document, while each column represents a user-defined question. Users can add, remove, and revise query dimensions as their understanding of the literature develops. The resulting spatial organization is intended to support cross-document comparison, iterative question formulation, and source verification while reducing repeated dialogue-management operations.

To evaluate the effectiveness of this interaction paradigm and examine its impact on cognitive processes, we conducted a controlled experiment (N=14). The study employed materials from a fictional domain to eliminate interference from participants' prior knowledge. The tasks included terminology filtering, integrative comprehension, and the identification of potentially misleading information.

The results showed no statistically significant differences in final task scores between the Lignin and control conditions. Participants using Lignin issued significantly fewer manual queries and were presented with substantially less LLM-generated text. Mental and physical demand ratings on the NASA-TLX scale [16] were descriptively lower in the Lignin group and showed relatively large effect sizes, although these differences did not reach statistical significance in the current sample. Qualitative observations further suggest that parallel tabular comparison and source traceability may help some users notice cross-document discrepancies and inspect supporting evidence. Together, these findings provide preliminary evidence that dynamic tabular interaction may reduce repeated interaction demands and support selected activities involved in early-stage scholarly sensemaking.

2 Related Work

2.1 AI Support for Reading Research Literature and Extracting Information

Before the emergence of large language models (LLMs), researchers had already explored automated methods to support document reading and organization. However, methods that could efficiently locate and extract relevant text based on the semantics of natural-language queries were still limited.

With the development of LLMs, AI-assisted reading has shown advantages in information extraction. Gartlehner et al. [13] reported that, for the information extraction tasks examined in their study, an LLM-assisted method was more accurate and efficient than a fully manual method. Sercombe et al. [30] also found that both using an LLM alone and manually verifying LLM-generated results were faster than fully manual processing for extracting basic information.

However, general-purpose chat-based interfaces with linear conversational structures still have limitations when used for comprehensive document or literature reading. Previous work has noted a possible mismatch between linear conversational processes and nonlinear information-processing tasks [14, 31]. In addition, long responses generated in chat interfaces may lack a clear focus and increase the cognitive effort required to identify and understand relevant information [8, 18, 23]. One common response to these limitations is to develop dedicated tools or platforms that organize the materials being analyzed, support interaction between users and LLMs, and present the extracted information in more suitable forms.

2.2 Information Organization and Interaction in AI-Assisted Reading

Existing AI-assisted reading tools use various approaches to document organization, information presentation, and interaction. These approaches include canvases [34], flashcards [29], graph structures [22], and interactive maps [35].

Lignin, the system proposed in this paper, is a tool for literature research and reading that combines the information extraction capabilities of LLMs with table-based information organization. Specifically, each row in Lignin represents one paper, each column represents one user question, and the information extracted by the LLM is presented in the corresponding cells. Users can issue a new query for an individual cell without regenerating the entire table, or directly edit generated content that they consider unsatisfactory.

Tables have already been used as an information organization format in existing AI-assisted reading tools. For example, commercial tools such as Consensus [7] and Elicit [10] organize information from research papers in tables and, to some extent, allow users to determine which questions the tables should answer. Some research systems also use table-based organization and presentation as a central feature.

However, the topic examined in this paper differs from that of existing work. Marco [11], which is relatively close to Lignin in its system design, primarily focuses on specific business and workplace tasks, such as resume screening and contract review. These tasks usually have relatively clear inputs, expected outputs, and evaluation criteria. Users generally understand the task process,

know what materials to provide to the system and what questions to ask, and can determine whether the system's responses meet their needs. This is also the type of usage context assumed by many commercial tools.

In contrast, this paper focuses on the early stage of a literature review. At this stage, users may not yet understand the key terminology of a field or have a complete view of the research problem and the structure of the field. Their actual interactions with a tool, as well as the cognitive activities involved in those interactions, may therefore differ from the task conditions assumed by the systems described above.

We refer to this process as scholarly sensemaking. Existing work has provided relatively limited discussion of this stage. This paper therefore examines whether AI-assisted reading, table-based interaction, and structured information presentation can support scholarly sensemaking, as well as how these designs affect users' cognitive activities during the process.

2.3 Characteristics of Scholarly Sensemaking Tasks

As discussed above, scholarly sensemaking differs from general information extraction tasks. Users may also exhibit different cognitive activities and interaction patterns during this process. Scholarly sensemaking can be described as a process in which researchers gradually form a provisional mental model of a field through continued searching, comparison across papers, and repeated revision, while the research question remains unclear, the structure of the field has not yet been established, and the criteria for relevance continue to change.

Previous research suggests that scholarly sensemaking has the following characteristics.

2.3.1 The Research Question Is Still Being Formed. Kuhlthau's Information Search Process (ISP) model [21] describes the early stage of information search as a period associated with uncertainty, ambiguity, and anxiety. General information extraction tasks usually assume that the questions and required fields have already been defined. Early-stage literature research, however, first requires researchers to determine what questions should be asked. Scholarly sensemaking therefore cannot be reduced to using AI to answer a set of predefined questions.

2.3.2 The Process Is Nonlinear. Foster [12] observed that researchers repeatedly return to earlier activities during information seeking. They redefine their questions, add new clues, and revise their understanding of the field. Pirolli and Card's information foraging and sensemaking models [26, 27], together with Bates's berrypicking model [2], also indicate that reading and search results continue to shape subsequent queries, leading researchers to revise their search directions.

Sensemaking also depends on reading and understanding information across multiple documents [4]. Even when information is extracted accurately from an individual paper, an isolated excerpt may still be insufficient for identifying contradictions, differences, or implicit assumptions across papers.

2.3.3 The Outcome Has a Subjective Component. The outcome of sensemaking includes the researcher's own interpretations, positions, and argument structure, rather than merely a compressed or restated version of the source material [33]. Consequently, the process may not produce a single standard answer, and its outcome can be difficult to evaluate solely against one set of factual criteria.

These characteristics distinguish the research questions of this paper from those of related systems. They also affect the experimental design, evaluation methods, and interpretation of the observed phenomena. The following subsection discusses how theories from cognitive science inform the system design.

2.4 Implications of Cognitive Science for Interaction Design in Scholarly Sensemaking

Schema theory is an important reference for the design of Lignin. In psychology, a schema is generally understood as a cognitive structure that organizes information and guides behavior. We consider scholarly sensemaking to have similarities with schema formation: researchers gradually develop an understanding of a field by comparing information across papers, continually revising their search directions, and correcting their existing understanding.

Piaget described assimilation and accommodation as two ways in which schemas change [25]. This account emphasizes that new knowledge is generally understood and integrated on the basis of existing schemas. Once a schema has been formed, revising it may therefore require additional cognitive effort.

The Knowledge Revision Components (KReC) Framework [19] provides a related account. The framework argues that inaccurate knowledge cannot simply be overwritten or removed by new knowledge. It also introduces the concept of coactivation: when conflicting old and new knowledge are activated at the same time, individuals may find it easier to revise their existing cognitive structures.

Cognitive load theory [32] also informs the design of Lignin. Excessive extraneous cognitive load may divert attention and negatively affect learning. Traditional conversational LLM interfaces may require users to ask similar questions repeatedly and identify relevant information within long textual responses. These additional operations may contribute to extraneous cognitive load.

Based on the characteristics of scholarly sensemaking and the theories discussed above, we propose the following design expectations. Whether these expectations are supported will be examined through subsequent experiments:

(1) An interaction format that allows users to ask a question once, receive responses across multiple papers, and edit individual cells may reduce repeated information extraction operations and unnecessary reading, thereby reducing extraneous cognitive load.
(2) Comparing multiple papers in relation to the same question may help users identify contradictions, differences, or missing information across papers and use this information to form or revise their schemas of the research field.
(3) Presenting potentially misleading content together with information that refutes or corrects it may support coactivation and help users revise their existing knowledge.

3 System Design

3.1 Design Goals

To address the cognitive demands of researchers during exploratory reading, the design of *Lignin* is guided by the following three core goals:

- **DG1: Supporting Evolving Information Needs.** Rather than imposing predefined templates or static summary dimensions, the system should allow users to actively construct and refine their cognitive schemata. By enabling the dynamic addition, removal, and modification of "columns" (query dimensions) as their understanding deepens, *Lignin* transforms a fixed automated task into a user-driven, fluid sensemaking process.
- **DG2: Reducing Repetitive Interaction.** To mitigate the overhead associated with frequent switching between documents or chat windows in traditional tools, *Lignin* implements a "one query, multiple responses" mechanism. This design is intended to minimize the physical effort of interface navigation and, more critically, reduces the cognitive load incurred by constant context switching between questioning, reading, and synthesizing information.
- **DG3: Supporting Comparison and Verification.** Leveraging the inherent advantages of tabular layouts, *Lignin* presents answers to the same query across multiple documents in parallel. This spatial alignment may help users identify source-ungrounded or inconsistent responses through lateral comparison. Furthermore, the system supports direct intervention in cell content, fostering human-AI collaboration in knowledge calibration.

3.2 Interface and Interaction

The interface of *Lignin* centers around a dynamic matrix designed to transform complex literature comparison into intuitive spatial operations. A screenshot of the interface is shown in Figure 2

3.2.1 Table-based Information Organization. The core UI consists of a dynamic table where **rows** represent individual documents and **columns** represent user-defined query dimensions (schemata). Initially empty, the table populates as users upload files and define queries. This layout facilitates lateral comparison, enabling researchers to identify commonalities and discrepancies across multiple sources at a glance.

3.2.2 Core Operations and Table Management. Three primary control buttons are positioned above the table to manage the workflow:

- **Upload New Paper**: Supports batch uploading of multiple PDF documents. Each new file automatically generates a corresponding row in the matrix.
- **Create New Column**: Opens a configuration modal with two fields: *Column Name (Summary)* for concise display in the table header, and *Description (The Question)* for the actual prompt sent to the LLM. This separation ensures that complex, detailed queries do not compromise the visual density and scannability of the matrix.

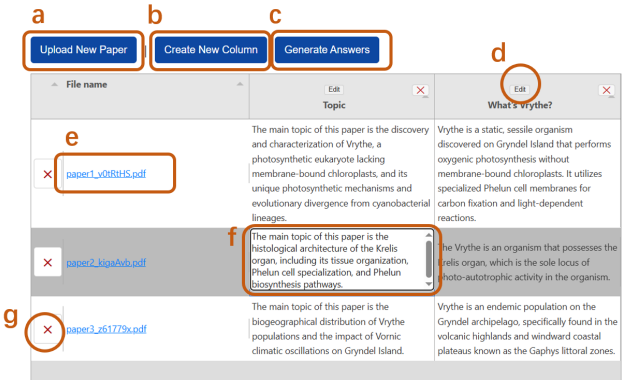


Figure 2: The main user interface of Lignin. (a) "Upload New Paper" for adding documents; (b) "Create New Column" for defining new query dimensions; (c) "Generate Answers" for manual regeneration; (d) Editable column headers (questions); (e) Interactive filenames that open the source verification modal; (f) Cells in edit mode (triggered by a single click) for manual refinement; (g) Delete buttons.

- **Structural Adjustment**: Users can delete obsolete rows or columns and reorder columns via drag-and-drop to align the layout with their current cognitive focus.

3.2.3 Automated Generation and Manual Intervention. To minimize interaction overhead, *Lignin* employs an automated response mechanism. The creation of a new row triggers queries across all existing columns for that document; conversely, a new column prompts the system to generate answers for all rows. The system also supports granular manual control:

- **Regeneration**: Users can clear a cell's content and trigger "Generate Answers" for specific updates. Modifying a column's *Description* field will automatically re-trigger generation for the entire column.
- **Direct Editing and Persistence**: Cells enter an edit mode upon a single click, allowing users to manually refine or correct AI-generated responses. Changes are saved upon clicking outside the cell. Crucially, non-empty cells are protected from being overwritten during automated batch updates, preserving users' manual edits during subsequent batch generation.

3.2.4 Traceability and Source Verification. The first column of the matrix features interactive filenames. Clicking a filename opens a detail modal that facilitates source verification (referencing DG3). The interface after clicking a filename is shown in Figure 3.

- **Multidimensional Q&A View**: The left pane of the modal displays a consolidated list of all queries and responses associated with the specific document.
- **Contextual Sourcing**: Each response is accompanied by a "Show Source Text" button. Clicking this button triggers the integrated PDF viewer in the right pane to jump to the exact location of the evidence, with the source text highlighted. This "Summary-to-Source" link allows users to inspect the

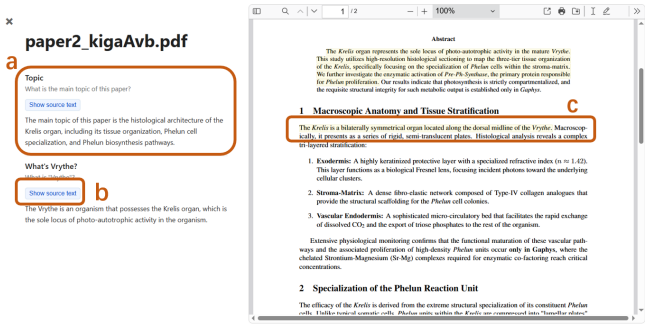


Figure 3: The source verification modal triggered by clicking a document's filename. (a) The left pane displays all queries and responses for the selected paper; (b) The "Show source text" button; (c) The integrated PDF viewer highlights the exact evidence text.

evidence associated with each response and assess whether the generated content is grounded in the source document.

3.3 Implementation

Lignin employs a decoupled architecture: a Django (Python) backend manages data logic, while a Tabulator.js frontend provides a responsive dynamic matrix. Document rendering and coordinate-based highlighting are integrated via Mozilla PDF.js. For core reasoning, the system utilizes the Mistral-7B-Instruct-v0.3 model [24], API provided by DeepInfra [9]. To optimize costs and reduce the risk of cross-document context contamination, we implement a dual-stage strategy: (1) isolated per-document prompting and (2) dynamic query filtering that transmits only empty cells to the model.

Traceability is achieved by extracting physical bounding boxes of evidence text using PyMuPDF and a fuzzy matching algorithm, which are then mapped to viewport pixels for smooth-scrolling navigation. Prompt engineering mandates the LLM to output structured JSON with verbatim citations. The model is explicitly instructed to return a "NOTFOUND" flag when relevant information is absent from the source text.

4 Evaluation

In this section, we evaluate the interactive performance of *Lignin* in multi-document sensemaking tasks. We conducted a between-subjects experiment in which participants were randomly assigned to either the experimental group (using *Lignin*) or the control group. To establish a comparable baseline, we implemented a traditional chat-based LLM interface as the interaction tool for the control group. The experimental materials consisted of a set of synthetic academic literature and three corresponding research tasks, designed to eliminate the influence of participants' prior knowledge. Furthermore, this section details the protocols for data collection, storage, and processing.

4.1 Participants

A total of 16 participants were initially recruited for this study. Data from one participant were excluded due to severe network latency that compromised task performance, and another participant withdrew for personal reasons. Consequently, the final analytical sample consisted of 14 participants ($N = 14$). Regarding educational background, 7 participants were current or graduated Master's students, and 7 were current Ph.D. candidates. The average age of the participants was 27.36 years (range: 23–36), with a gender distribution of 9 males, 4 females, and 1 non-binary individual. The study employed a hybrid participation mode, with 9 participants joining online and 5 attending in person.

All participants provided informed consent prior to the experiment. The study protocol was approved by the Institutional Review Board (IRB) and conducted in accordance with ethical guidelines. Upon completion, eligible participants received $20 in compensation. To mitigate potential carryover effects, a between-subjects design was adopted, with participants randomly assigned to either the experimental or control group, ensuring each individual experienced only one interaction paradigm.

4.2 Materials and Experimental Tasks

To evaluate the impact of interaction paradigms on the sensemaking process, we designed a controlled set of experimental materials and tasks.

4.2.1 Fictional Literature Design. The experimental stimuli consisted of six fully synthetic academic papers designed to eliminate bias from participants' prior knowledge. The content revolved around a fictional biological entity, "Vrythe," covering its physiological mechanisms, ecological habitat, and associated socio-cultural phenomena.

To simulate the cognitive challenges inherent in real-world research—specifically to trigger schema transitions such as Piaget's *accommodation* or Rumelhart's *restructuring* [28]—we intentionally embedded conceptual ambiguities, polysemy, and implicit logical links across the documents. The specific content of the literature is summarized in Table 1.

4.2.2 Experimental Tasks. Participants were required to complete three sequential tasks within 40 minutes. The order was fixed to prevent downstream tasks from priming the participants' initial sensemaking strategies:

(1) **Keyword Discovery**: Identifying terms related to Vrythe in each individual article.
(2) **Cross-document Synthesis**: Answering "Is Vrythe currently facing difficulties? Why?" This required constructing a causal chain across ecological and physiological data.
(3) **Term Definition**: Providing definitions for eight key terms (e.g., *Vrythe, Gaphys, Mawv*). This task tested the ability to disambiguate polysemous and visually similar terms.

4.2.3 Criteria and Instructions. Participants were instructed to aim for "confident relational understanding" rather than absolute accuracy, prioritizing the capture of mental models during the exploratory phase of sensemaking. To ensure authentic cognitive engagement, participants were instructed to avoid the "blind" copying

Table 1: Summary of Fictional Literature Materials

ID	Topic	Key Content
Paper 1	Physiology of Vrythe	Details non-chloroplast photosynthesis and the dependence of Phelun cells on intense light.
Paper 2	Organ Function	Discusses the relationship between the Krelis organ and Phelun cells, mentioning dependence on the "Gaphys" environment.
Paper 3	Ecology and Crisis	Describes the Vornic climate of the Gryndel Islands and the population decline caused by climatic shifts.
Paper 4	Agricultural Technique	Introduces the Zelwa crop and a pruning method called "Gaphis" (orthographically similar to Gaphys).
Paper 5	Sociological Impact	Defines the "Gaphys Effect" in education and introduces the term "Mawv" as innate talent.
Paper 6	Developmental Biology	Reveals Vrythe's metamorphic stages (Aolys, Ventys, Gaphys); in the Gaphys stage, the Mawv organ is shed to grow the Krelis.

of LLM outputs without prior reading or comprehension, ensuring that their responses reflected a genuine sensemaking process.

4.3 Experimental Apparatus and Baseline

To ensure the comparability of experimental results, we developed two distinct interaction interfaces while maintaining consistency in the underlying model configuration.

4.3.1 The Answer Editor. To prevent the answering process from interfering with the literature reading task, a standalone interface called the Answer Editor was developed. Implemented using the Python Tkinter library, this GUI operated independently of both Lignin and the control group interface. Participants were required to complete the three research tasks sequentially within this editor. The system automatically logged all response content, task-switching behaviors, and edit timestamps for subsequent cognitive process analysis. The interface is shown in Figure 4.

4.3.2 Control Group Baseline Tool. The control group utilized an interface that adheres to the interaction paradigms of prevailing conversational LLMs (e.g., ChatGPT):

- **Interaction Logic**: Responses were presented in a linear, bubble-based chat format with Markdown support. Participants interacted via a standard input field at the bottom of the page.
- **Document Management**: The interface supported the uploading of multiple PDF documents, allowing for multi-document querying within a single session.
- **Multitasking Support**: Participants could create, switch between, and delete multiple conversation tabs. Each tab maintained its own contextual memory, remaining isolated from others.

4.3.3 Model Consistency and Prompting Strategy. Both experimental and control groups utilized the same underlying model. However, the prompting strategies differed: the control group did not enforce strict JSON formatting or mandate the inclusion of verbatim source citations. This configuration was designed to simulate the typical user experience when using general-purpose conversational AI for academic reading.

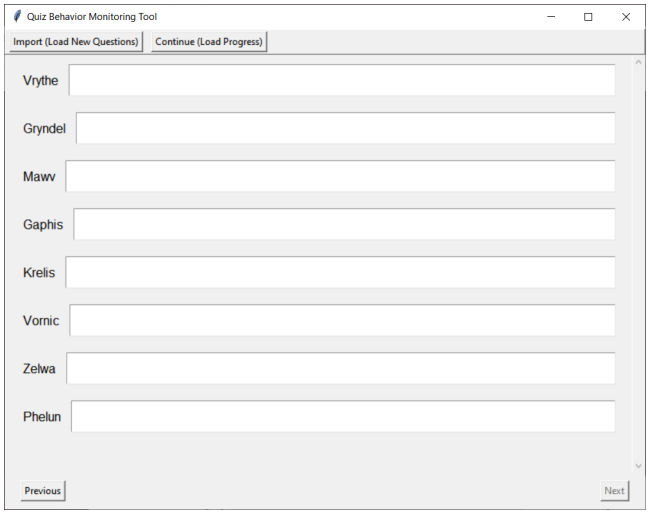


Figure 4: The graphical user interface of the standalone Answer Editor. Participants used this independent tool to complete the experimental tasks (shown here is Task 3: Term Definition).

4.4 Procedure

The experimental procedure was structured into three core phases: training, the main task, and a post-task questionnaire with debriefing. To accommodate different participant needs, the study supported both in-person and online participation modes. In-person participants completed the setup and informed consent in a laboratory environment, whereas online participants accessed and controlled the experimental apparatus remotely using Zoom and TeamViewer. Prior to the main task, participants underwent a 5-minute training session. During this phase, the experimenter demonstrated the core functionalities of the assigned reading tool (either Lignin or the control interface) as well as the independent Answer Editor. Participants were then provided with practice time to familiarize themselves with the interaction logic.

Following the training, participants proceeded to the 40-minute main task. They were instructed to read the fictional literature and continuously update their insights in the Answer Editor as their understanding evolved, rather than waiting until they achieved full comprehension. Throughout this phase, participants were required to employ the "Think Aloud" protocol to continuously verbalize their cognitive processes and real-time reflections. All on-screen interactions and audio were recorded using OBS Studio for subsequent qualitative analysis. Upon completion of the reading task, the recording was stopped, and participants were directed to the Qualtrics platform to complete a subjective questionnaire. The session concluded with a brief debriefing, during which the experimenter conducted a retrospective interview to explore the participants' in-depth experiences and clarified the overarching objectives of the study.

4.5 Data Collection and Analysis

To evaluate the impact of the interaction paradigms on cognitive processes and task performance, we employed a multi-modal data collection strategy and established a scoring rubric alongside a qualitative analytical framework.

4.5.1 Data Collection. Experimental data were collected across three primary dimensions, with all telemetry logs timestamped to the millisecond:

- **Interaction Logs**: For the *Lignin* group, the system recorded document uploads, dynamic column creation (queries), cell edits, and LLM-generated responses. For the control group, logs captured the management of conversation tabs (creation, deletion, switching), file uploads, user prompts, and corresponding LLM replies.
- **Answer and Behavioral Logs**: The Answer Editor logged the participants' final submitted responses, iterative text-editing histories, and task-switching behaviors, enabling us to trace the trajectory of their sensemaking processes.
- **Subjective and Qualitative Data**: Subjective cognitive load was measured post-task using the NASA-TLX scale. Additionally, we collected OBS screen recordings, the experimenter's on-site observation notes, and transcripts from the semi-structured debriefing interviews.

4.5.2 Scoring Rubric. To objectively quantify the quality of participants' sensemaking, two researchers developed a standardized grading rubric based on the ground truth embedded within the fictional materials:

- **Task 1 (Keyword Discovery)**: Grounded in Semantic Network Theory [6], we categorized the extracted terms into four weighted layers:
  - *Core Layer (+2 pts)*: Terms directly associated with the central subject, Vrythe.
  - *Explanatory Layer (+1 pt)*: Terms that explain core concepts and have an indirect relationship with Vrythe.
  - *Common Sense Layer (0 pts)*: General background vocabulary lacking domain specificity.
  - *Noise Layer (-1 pt)*: Irrelevant or misleading terms that deviate from the core logic and unnecessarily increase cognitive load.
- **Task 2 (Cross-document Synthesis)**: The ground truth encompassed two potential causal chains (e.g., climatic warming → nutrient disruption → maturation failure). Points were awarded additively: correctly identifying the predicament of population decline (+1 pt); extracting a causal factor from a single document (+1 pt per factor); and successfully establishing a cross-document causal relationship (+2 pts per link). Incorrect answers were not penalized.
- **Task 3 (Term Definition)**: Given the synthetic nature of the materials, absolute ground truth was available. Participants received 1 point for each correct definition. For polysemous terms, 1 point was awarded for each accurately disambiguated semantic dimension. Incorrect or blank answers received no penalty.

4.5.3 Privacy and Data Management. All collected qualitative and quantitative data were de-identified to remove any Personally Identifiable Information. The anonymized datasets are securely stored on encrypted institutional servers and are accessible solely to authorized research personnel for analytical purposes.

5 Results

5.1 Task Performance

To evaluate the effectiveness of Lignin in supporting sensemaking tasks, we analyzed participants' performance across three categories of questions and their overall scores. Quantitative results indicate that while Lignin did not lead to a statistically significant difference in final task scores compared to the traditional chat-based interface ($p > 0.05$ for all metrics), specific task categories revealed divergent trends and medium effect sizes.

For Question 1 (Keyword Discovery), the control group achieved a higher mean score ($M = 13.71$) than the Lignin group ($M = 10.57$). An independent samples t-test showed no significant difference ($t(12) = -0.97, p = 0.351$), though a medium effect size was observed (Cohen's $d = -0.52$). Qualitative observation suggests this may be attributed to a "greedy" strategy adopted by control group participants, who tended to list all perceived terms regardless of relevance to the core entity. Although the rubric penalized specific misleading terms (Noise Layer), the occasional point deductions were easily offset by the gains from inadvertently capturing explanatory terms. Furthermore, the inclusion of general background terms (Common Sense Layer) carried no penalty. Consequently, this "greedy" strategy yielded a net positive outcome, contributing to the control group's higher scores.

Conversely, for Question 2 (Synthesis), which required cross-document causal reasoning, the Lignin group had a higher mean score ($M_{Lignin} = 3.29$ vs. $M_{Control} = 2.00$). A Mann-Whitney U test confirmed no significant difference ($U = 35.5, p = 0.150$), yet indicated a medium positive effect for the Lignin paradigm (Rank-Biserial $r = 0.45$). Performance on **Question 3 (Definition Search)** was identical across both groups ($M = 4.71, U = 25.0, p = 1.00$).

In terms of Total Score, no significant difference was found between the Lignin group ($M = 18.57$) and the control group ($M = 20.43$; $t(12) = -0.52, p = 0.611, d = -0.28$). Figure 5 presents a comparison of task performance between the experimental and control groups.

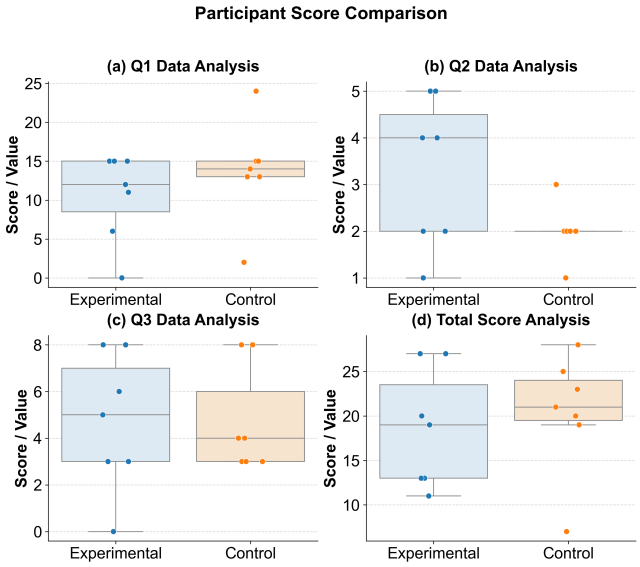


Figure 5: Comparison of participant task performance scores between the Experimental group (Lignin) and the Control group across three specific tasks and the total score. (a) Keyword Discovery; (b) Cross-document Synthesis; (c) Term Definition; (d) Total Score Analysis.

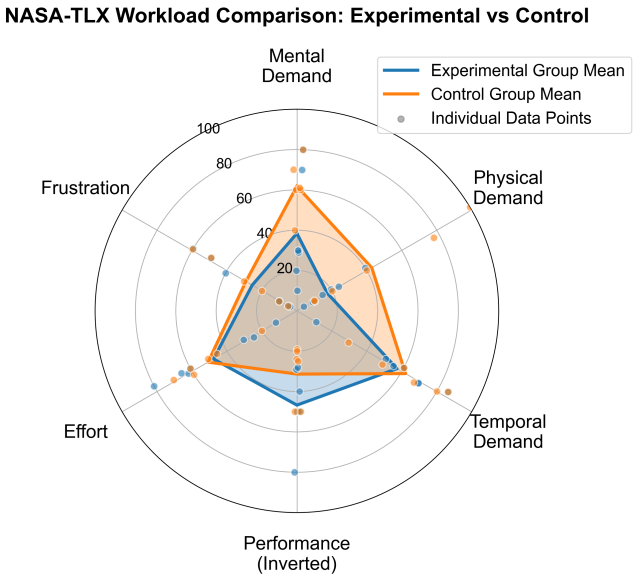


Figure 6: Subjective workload comparison using the NASA-TLX scale. The radar chart illustrates the mean scores and individual data points across six dimensions for both groups. Lower values represent lower workload, with the Performance dimension inverted for consistency.

5.2 Subjective Cognitive Load Analysis

An analysis of the NASA-TLX dimensions provides further insight into the impact of Lignin on the user's cognitive process. Quantitative data suggest that the dynamic table paradigm offers potential advantages in reducing cognitive resource consumption, particularly in terms of mental and physical workload. The results are shown in Figure 6.

In the dimensions of Mental Demand (Q1) and Physical Demand (Q2), the subjective load scores for the experimental group were substantially lower than those of the control group ($M_{Lignin} = 38.43, M_{Ctrl} = 61.83$ for Q1; $M_{Lignin} = 17.43, M_{Ctrl} = 42.83$ for Q2). Although the p-values from the independent samples t-test did not reach statistical significance at the current sample size ($p_{Q1} = 0.074, p_{Q2} = 0.118$), both dimensions exhibited large effect sizes (Cohen's $d_{Q1} = -1.099, d_{Q2} = -0.942$). This suggests that by transforming linear dialogues into spatial comparison tasks, Lignin has the potential to mitigate the pressure users face during information retrieval and mental processing—a finding that aligns with the reduction in interaction frequency discussed in the subsequent section. However, the descriptive differences require further investigation.

Notably, in the Performance (Q4) dimension, the experimental group reported higher load scores than the control group ($M_{Lignin} = 46.71, M_{Ctrl} = 31.33, p = 0.091$), with a large effect size (Rank-Biserial $r = 0.571$). This suggests that participants in the Lignin group were subjectively less satisfied with their own performance. However, this subjective "uncertainty" did not manifest in the actual task scores. The experimental group had a descriptively higher mean score in the synthesis task (Question 2), although the difference was not statistically significant. This discrepancy may stem from two factors. First, the tabular interface may have encouraged cross-document source verification, making participants aware of information conflicts and thus more cautious about their own conclusions. Second, considering the novelty of the dynamic matrix compared to the familiar chat-based paradigm, the higher perceived workload in performance might reflect a learning curve effect, where participants felt less confident navigating a new interaction logic despite achieving higher objective accuracy in synthesis tasks.

Regarding the Total Subjective Workload (RTLX), the experimental group ($M = 233.14$) showed a medium-level reduction compared to the control group ($M = 278.17, d = -0.523$). The experimental group reported descriptively lower total workload, with no statistically significant difference in task scores was observed, indicating a trend toward better cognitive economy.

5.3 Interaction Frequency and Text Processing Volume

While the final task performance remained comparable across conditions, significant disparities were observed in terms of interaction cost. The experimental results provide early evidence that Lignin's tabular paradigm can achieve similar observed task scores while reducing operational overhead.

Regarding interaction frequency, the average questioning rate of the control group ($M = 0.744, SD = 0.339$ queries/min) was approximately 3.78 times higher than that of the experimental group ($M = 0.197, SD = 0.096$ queries/min). An independent samples

t-test revealed a highly significant difference between the groups ($t(12) = 4.115, p = 0.0014$), with a very large effect size (Cohen's $d = -2.1998$). This suggests that in traditional chat-based interfaces, users are compelled to interact frequently to retrieve information, whereas Lignin's "one-query, multi-response" mechanism dramatically optimizes retrieval efficiency. Results are shown in Figure 7.

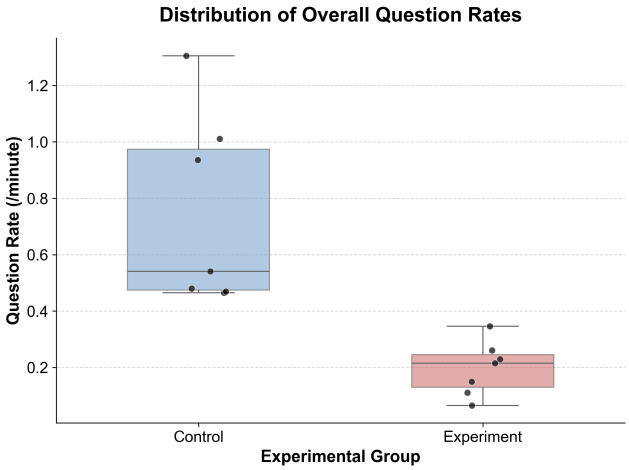


Figure 7: Distribution of overall question rates (queries per minute) between the Control and Experimental (Lignin) groups.

As shown in Figure 8, the discrepancy in text processing volume was also significant. Participants in the control group were presented with an average of 10202.86 words ($SD = 3287.00$) generated by the LLM, while those in the experimental group received significantly less text ($M = 1599.29, SD = 720.13$), representing a 6.38-fold difference. A Welch's t-test was conducted due to unequal variances, showing a significant difference ($t = 6.7647, p = 0.0003$) and a large effect size (Cohen's $d = -3.6159$).

These findings underscore that the control group incurred significantly higher interaction and reading costs, with no observed significant performance improvement in the current experiment. These results align with the descriptive trends observed in the NASA-TLX dimensions of mental and physical demand.

5.4 Qualitative Observations and Behavioral Patterns

In this section, we identify and analyze key behavioral differences between the two groups during the sensemaking process, based on real-time observations and retrospective interviews. Participant quotes cited in the following analysis are sourced from experimental observation notes and interview transcripts, with individuals identified using the notation "P + [ID]".

5.4.1 Spontaneous "Parallel Questioning" Behavior. Through real-time observations, we identified a spontaneous "parallel questioning" pattern among high-performing participants in the control group. These users manually opened multiple chat windows to

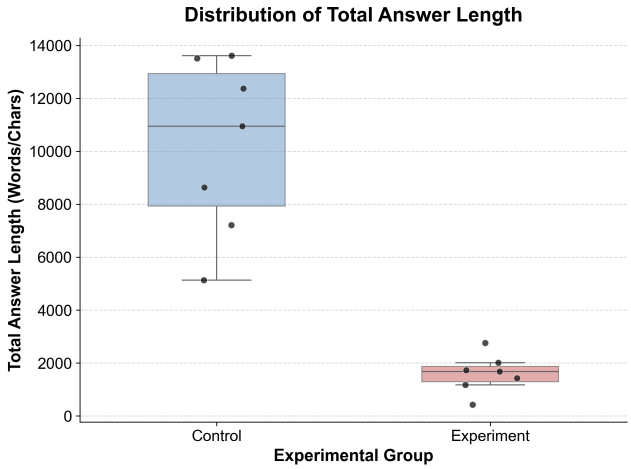


Figure 8: Distribution of total answer text volume processed by participants.

repeat the same queries across different documents. Participants exhibiting this behavior (e.g., P3, P7, and P13) stated in follow-up interviews that this was an empirical strategy developed through their long-term usage of AI tools. This behavior aligns closely with the core interaction logic of Lignin, suggesting that "multi-document synchronous comparison" is an intuitive demand for users during complex reading tasks. Furthermore, it validates the necessity of automating such parallel workflows to reduce manual effort.

5.4.2 Depth-First vs. Breadth-First Exploration Strategies. We observed two distinct reading paths during the experiment. We use the terms depth-first and breadth-first descriptively, following prior distinctions between immediately pursuing one information item and first scanning multiple documents before selectively deepening. These terms have also been used in prior HCI and information-seeking research [17, 20]. Several participants in the control group, influenced by the linear chat interface, tended to adopt a "depth-first" strategy, often becoming bogged down by complex, non-essential terminology in the very first document. This issue was most pronounced with participant P1, who experienced significant friction early on, remarking, "I cannot understand the terms", and even asking the researcher, "Can this experiment really be completed?" Subsequently, P1 exhibited a decline in judgment; despite reminders from the researcher, the participant continued to copy LLM responses directly. In the post-task interview, P1 admitted to not understanding the meaning of the copied text. We hypothesize that this was not a lack of cooperation, but rather a state of cognitive overload that hindered effective reasoning.

In contrast, Lignin users primarily demonstrated "breadth-first" characteristics. By scanning horizontally across the table, participants could quickly locate target information based on the frequency of keywords across different documents, allowing them to prioritize core concepts. This layout helped participants avoid becoming prematurely mired in the idiosyncratic details of a single paper, thereby facilitating a more balanced and global progression of sensemaking.

5.4.3 Limitations of Chat-based AI and Improvements in Lignin.

Detecting Source-Ungrounded Responses through Cross-Document Comparison. In both conditions, the LLM occasionally produced responses that were not supported by the source documents. In one instance, P4 asked, "If Vrythe faces difficulties, what would be the reasons?" Rather than returning NOTFOUND, the model interpreted the question as an open-ended hypothetical and treated Vrythe as a person, generating a scenario in which "Vrythe is currently facing employment difficulties." This response was not an arbitrary fabrication: it was locally plausible within the scenario inferred by the model. However, it departed from the document-grounded task because neither the interpretation of Vrythe as a person nor the employment-related explanation was supported by the source material. In this sense, the episode is better understood as a failure to maintain source grounding and to abstain when documentary evidence was insufficient. By comparing the response with answers derived from other documents, P4 identified the mismatch and discarded it.

Information Gaps as Active Signals. According to Information Foraging Theory, users often abandon materials with a weak "information scent." Control group participants were observed to stop consulting documents they initially deemed irrelevant, leading to the omission of key definitions such as the polysemous term *Mawv*. In Lignin, cells that were initially marked as "NOTFOUND" but later produced responses to subsequent questions provided a clear visual signal. This guided users to re-examine previously ignored documents, offering a "remedy opportunity." We observed that when participants in the experimental group, such as P4 and P8, saw a document which was previously regarded as irrelevant by them because it had repeatedly returned "NOTFOUND" suddenly respond to a subsequent question, they almost immediately fixated on the unusual cell and expressed surprise at this phenomenon. The interview confirmed that the participants' attention was indeed recaptured by these unexpected responses.

Discrepancy-induced source processing. When processing visually similar terms, such as *Gaphys* and *Gaphis*, LLMs sometimes silently "correct" users' spelling errors without notification, for example, by interpreting *Gaphis* as *Gaphys*. Failing to detect such silent corrections leads to recording incorrect definitions.

Multiple participants in the experimental group noticed inconsistencies in the table. For example, four responses referred to *Gaphys*, whereas only one referred to *Gaphis*. This anomaly prompted them to consult the original texts and revise their initial interpretations. A greater proportion of participants detected this issue in the Lignin group than in the control group (4/7 vs. 2/7).

This observation is consistent with the cognitive phenomenon described by the Discrepancy-Induced Source Comprehension hypothesis [3]. According to this account, perceived inconsistencies increase attention to sources and encourage readers to process the information more carefully. The coactivation mechanism proposed in the KReC framework [19] further supports the potentially constructive role of such discrepancies in cognitive processing. Lignin provides favorable conditions for making these inconsistencies visible.

6 Discussion

6.1 Lignin's Response to Cognitive Science Theories and Its Potential Applications

Compared with the conventional interface, participants using Lignin submitted fewer queries and received substantially less response text. Meanwhile, the difference in final task performance between the two groups was not statistically significant. This suggests that, under the current experimental conditions, the reduced interaction frequency did not cause an evident loss in performance. Participants in the experimental group may have spent more time comparing and reflecting on information outside their direct interactions with the system. However, the nature of these cognitive activities and their contribution to scholarly sensemaking cannot be determined from the current experiment.

Qualitative observations indicate that Lignin's design was consistent with the breadth-first exploration strategies adopted by some higher-performing participants. Its table-based interface presented responses from multiple documents to the same question simultaneously, making differences across documents more visible and creating favorable conditions for Coactivation in the KReC framework. These observations provide preliminary evidence that table-based information organization may support multiple-document reading and scholarly sensemaking.

6.2 Influence of Individual Strategies

In addition to the limited sample size, participants' experience with AI and their learning strategies may have contributed to the absence of statistically significant differences. Although Lignin supports comparison across documents, it cannot ensure that users adopt effective questioning strategies or workflows. Its advantages may be reduced when participants continue to focus on individual documents or rely mainly on reading the original text because they do not trust AI-generated responses.

For example, P6 and P15 in the experimental group spent considerable time checking the original documents because they distrusted the AI output. In contrast, experienced users in the control group, including P3, P7, and P13, partially compensated for the limitations of the conversational interface by manually simulating a "parallel interaction" workflow through repeated queries. This suggests that tool performance depends partly on individual strategies. Future systems should provide clearer guidance and communicate their evidence and performance more transparently, helping users develop appropriate trust and adopt more effective strategies.

6.3 Deeper Sensemaking and the Limitation of Evaluation

Interview data show that some participants in the experimental group made further inferences about relationships that were not directly addressed by the task questions. For example, they identified etymological relationships between visually similar terms or recognized conceptual borrowing across disciplinary contexts. These observations may suggest that Lignin supported the formation of connections across documents and the development of cognitive schemas. However, the current experiment cannot directly attribute these outcomes to the system.

The current scoring method mainly evaluates the accuracy of factual retrieval and reasoning. It does not capture connections and inferences that extend beyond the task questions. This indicates that the present evaluation framework cannot fully represent the cognitive activities involved in scholarly sensemaking.

6.4 Limitations and Future Work

The main limitations of this study are its small sample size and the inability of the current evaluation criteria to capture the full range of cognitive activities involved in scholarly sensemaking. The findings have limited statistical robustness and provide only limited insight into the mechanisms underlying cognitive activities in scholarly sensemaking. Future studies should include larger samples and develop more comprehensive evaluation methods that encompass factual extraction, relationship discovery, knowledge revision, and question formation.

In addition, the current experiment was conducted using a set of fictional documents. Future research could conduct experiments in authentic academic settings to further investigate the role of this interaction paradigm in literature reading and scholarly sensemaking.

7 Conclusion

To support early-stage, multi-document scholarly sensemaking, this study developed *Lignin*, an AI-assisted reading tool that combines dynamic tabular organization, incremental querying, per-cell editing, and source traceability. We evaluated *Lignin* against a chat-based LLM baseline in a controlled study using fictional academic materials. No statistically significant differences in final task scores were observed between the two conditions. Participants using *Lignin*, however, issued significantly fewer manual queries and were presented with substantially less LLM-generated text. Mental and physical demand ratings were descriptively lower in the *Lignin* condition, although these differences did not reach statistical significance. Qualitative observations further suggest that parallel comparison and source-linked responses may help some users notice cross-document discrepancies and assess whether generated content is grounded in the source materials. These findings provide preliminary evidence that dynamic tabular interaction can reduce repeated interaction demands during scholarly sensemaking, while larger studies in authentic academic settings are needed to confirm its effects and evaluate the underlying mechanisms.

References

[1] Qianxiang Ai, Fanwang Meng, Jiale Shi, Brenden Pelkie, and Connor W Coley. 2024. Extracting structured data from organic synthesis procedures using a fine-tuned large language model. *Digital discovery* 3, 9 (2024), 1822–1831.

[2] Marcia J Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online review* 13, 5 (1989), 407–424.

[3] Jason LG Braasch, Jean-François Rouet, Nicolas Vibert, and M Anne Britt. 2012. Readers' use of source information in text comprehension. *Memory & cognition* 40, 3 (2012), 450–465.

[4] M. Anne Britt and Jean Rouet. [n. d.]. Multiple Document Comprehension. In *Oxford Research Encyclopedia of Education*, Kathy Hytten (Ed.). Oxford University Press. doi:10.1093/acrefore/9780190264093.013.867

[5] Andry Castro, João Pinto, Luís Reino, Pavel Pipek, and César Capinha. 2024. Large language models overcome the challenges of unstructured text data in ecology. *Ecological informatics* 82 (2024), 102742.

[6] Allan M Collins and M Ross Quillian. 1969. Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior* 8, 2 (1969), 240–247.

[7] Consensus. 2026. Consensus: Evidence-Based Answers, Faster. https://consensus.app/. Accessed: 2026-03-31.

[8] Adam J Coscia, Shunan Guo, Eunyee Koh, and Alex Endert. 2025. OnGoal: Tracking and Visualizing Conversational Goals in Multi-Turn Dialogue with Large Language Models. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–18.

[9] DeepInfra. 2026. DeepInfra | Fast, scalable, and cost-effective AI inference. https://deepinfra.com/ Accessed: 2026-03-31.

[10] Elicit. 2026. Elicit: The AI Research Assistant. https://elicit.com/. Accessed: 2026-03-31.

[11] Raymond Fok, Nedim Lipka, Tong Sun, and Alexa F Siu. 2024. Marco: Supporting Business Document Workflows via Collection-Centric Information Foraging with Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 842, 20 pages. doi:10.1145/3613904.3641969

[12] Allen Foster. 2005. A Non-Linear Model of Information Seeking Behaviour. *Information Research: an international electronic journal* 10, 2 (2005), n2.

[13] Gerald Gartlehner, Shannon Kugley, Karen Crotty, Meera Viswanathan, Andreea Dobrescu, Barbara Nussbaumer-Streit, Graham Booth, Jonathan R Treadwell, Jung Min Han, Jesse Wagner, et al. 2025. Artificial Intelligence–Assisted Data Extraction With a Large Language Model: A Study Within Reviews. *Annals of Internal Medicine* 178, 12 (2025), 1763–1771.

[14] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K Kummerfeld, and Elena L Glassman. 2024. Supporting sensemaking of large language model outputs at scale. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.

[15] Wu Hao Ran, Xi Xi, Furong Li, Jingyi Lu, Jian Jiang, Hui Huang, Yuzhuan Zhang, and Shi Li. 2025. Structured Semantics from Unstructured Notes: Language Model Approaches to EHR-Based Decision Support. *arXiv e-prints* (2025), arXiv–2506.

[16] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[17] Christine Jenkins, Cynthia L Corritore, and Susan Wiedenbeck. 2003. Patterns of information seeking on the Web: A qualitative study of domain expertise and Web expertise. *IT & society* 1, 3 (2003), 64–89.

[18] Peiling Jiang, Jude Rayan, Steven P Dow, and Haijun Xia. 2023. Graphologue: Exploring large language model responses with interactive diagrams. In *Proceedings of the 36th annual ACM symposium on user interface software and technology*. 1–20.

[19] Panayiota Kendeou and Edward J O'Brien. 2014. 16 the knowledge revision components (KReC) framework: processes and mechanisms. In *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences*. MIT Press Cambridge, MA, 353–377.

[20] Kerstin Klöckner, Nadine Wirschum, and Anthony Jameson. 2004. Depth-and breadth-first processing of search result lists. In *CHI'04 extended abstracts on Human factors in computing systems*. 1539–1539.

[21] Carol C Kuhlthau. 1991. Inside the search process: Information seeking from the user's perspective. *Journal of the American society for information science* 42, 5 (1991), 361–371.

[22] Xiang Li, Cara Yuejia Li, Emily Kuang, Can Liu, and Jian Zhao. 2026. MindTrellis: Co-Creating Knowledge Structures with AI through Interactive Visual Exploration. In *Proceedings of the 2026 Designing Interactive Systems Conference*. 4892–4912.

[23] Xiao Ma, Swaroop Mishra, Ariel Liu, Sophie Ying Su, Jilin Chen, Chinmay Kulkarni, Heng-Tze Cheng, Quoc Le, and Ed Chi. 2024. Beyond chatbots: Explorellm for structured thoughts and personalized model responses. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–12.

[24] Mistral AI. 2026. Mistral AI | Frontier AI in your hands. https://mistral.ai/ Accessed: 2026-03-31.

[25] Jean Piaget and Margaret Trans Cook. 1952. The origins of intelligence in children. (1952).

[26] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.

[27] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.

[28] David E Rumelhart and Donald A Norman. 1976. *Accretion, tuning and restructuring: Three modes of learning*. Technical Report.

[29] Scholarcy Ltd. 2026. Scholarcy: An AI-Powered Research Assistant and Article Summarizer. https://www.scholarcy.com/. Accessed: 2026-07-20.

[30] Jayden Sercombe, Zachary Bryant, and Jack Wilson. 2025. Evaluating a Customized Version of ChatGPT for Systematic Review Data Extraction in Health Research: Development and Usability Study. *JMIR formative research* 9, 1 (2025), e68666.

[31] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling multilevel exploration and sensemaking with large language models. In

Proceedings of the 36th annual ACM symposium on user interface software and technology. 1–18.

[32] John Sweller, Jeroen JG Van Merrienboer, and Fred GWC Paas. 1998. Cognitive architecture and instructional design. *Educational psychology review* 10, 3 (1998), 251–296.

[33] Victoria Uren, Simon Buckingham Shum, Michelle Bachler, and Gangmin Li. 2006. Sensemaking tools for understanding research literatures: Design, implementation and user evaluation. *International journal of human-computer studies* 64, 5 (2006), 420–445.

[34] Runlong Ye, Patrick Lee, Matthew Varona, Oliver Huang, and Carolina Nobre. 2025. Scholarmate: A mixed-initiative tool for qualitative knowledge work and information sensemaking. In *Adjunct Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*. 1–7.

[35] Sam Yu-Te Lee and Kwan-Liu Ma. 2024. Hints: Sensemaking on large collections of documents with hypergraph visualization and intelligent agents. *IEEE Transactions on Visualization and Computer Graphics* 31, 9 (2024), 5532–5546.