Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking

Hyeonsu B. Kang
hyeonsuk@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Sherry Tongshuang Wu
sherryw@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Joseph Chee Chang
josephc@allenai.org
Allen Institute for AI
Seattle, WA, USA

Aniket Kittur
nkittur@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

ABSTRACT

Efficiently reviewing scholarly literature and synthesizing prior art are crucial for scientific progress. Yet, the growing scale of publications and the burden of knowledge make synthesis of research threads more challenging than ever. While significant research has been devoted to helping scholars interact with individual papers, building research threads scattered across multiple papers remains a challenge. Most top-down synthesis (and LLMs) make it difficult to personalize and iterate on the output, while bottom-up synthesis is costly in time and effort. Here, we explore a new design space of mixed-initiative workflows. In doing so we develop a novel computational pipeline, Synergi, that ties together user input of relevant seed threads with citation graphs and LLMs, to expand and structure them, respectively. Synergi allows scholars to start with an entire threads-and-subthreads structure generated from papers relevant to their interests, and to iterate and customize on it as they wish. In our evaluation, we find that Synergi helps scholars efficiently make sense of relevant threads, broaden their perspectives, and increases their curiosity. We discuss future design implications for thread-based, mixed-initiative scholarly synthesis support tools.

ACM Reference Format:
Hyeonsu B. Kang, Sherry Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. 2023. Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking. In *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), October 29–November 01, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3586183.3606759

1 INTRODUCTION

Scientific and engineering innovations rely on synthesis of prior art: to know what approaches have been tried and identify most promising ideas for new problems; to unlock creative new ideas by combining existing ones; to reason about open challenges and unknown unknowns; and to contextualize one's research in a broader context of literature [26]. At the same time, scholarly synthesis is a cognitively difficult task because it involves many demanding inter-related steps in the process such as discovering relevant literature about a problem, reading and comprehending papers, collecting useful information and organizing it for further distillation, and recording and monitoring progress by developing an outline that summarizes current learning in the space [52]. Furthermore, scholarly synthesis becomes even more challenged by expertise barriers for engaging with scientific literature due to deepening specialization [6, 14, 15], the accelerating rate of growth [8, 17], and its increasingly interdisciplinary nature [34, 48].

In order to synthesize knowledge scattered across multiple papers, scholars often employ iterative workflows that involve multiple inter-related stages. Here, we focus on literature review workflows for high-level exploratory searches and synthesis in less familiar knowledge domains. Such workflows can be characterized by their location on a spectrum of how much of the initiative is automated, between fully *bottom-up* and fully *top-down* workflows. Systems closer to the *bottom-up* end of the spectrum such as Apolo [11] and PaperQuest [38] allow users to explicitly save an interesting paper, and Relatedly [35] allows keyword queries for expanding to additional papers and clips. Threddy [18] and Passages [12] enable highlighting of clips directly from documents for saving and organizing, allowing users to better maintain their context of reading in the process. However, in these workflows users are required to manually and repeatedly collect threads from documents whose cognitive and interaction costs can compound quickly.

In contrast, systems near the *top-down* end of the spectrum such as ConnectedPapers[1] and Metro Maps of Science [41] provide scholars an initial visual overview of the research landscape to help them make sense of the structure of knowledge and discover interesting parts in it which can be especially useful for scholars new to a domain. In addition, recent Large Language Models (LLMs)-based systems such as Galactica [45], ChatGPT[2] and Google Bard[3] enable Q&A-based interactions with knowledge domains which users can iteratively query. However, the responses of such systems are similar to visual overviews described above in the sense that they are complete artifacts, rendering them less penetrable and useful

UIST '23, October 29–November 01, 2023, San Francisco, CA, USA

ACM ISBN 979-8-4007-0132-0/23/10.
https://doi.org/10.1145/3586183.3606759

[1] https://www.connectedpapers.com/
[2] https://chat.openai.com/chat
[3] https://bard.google.com/

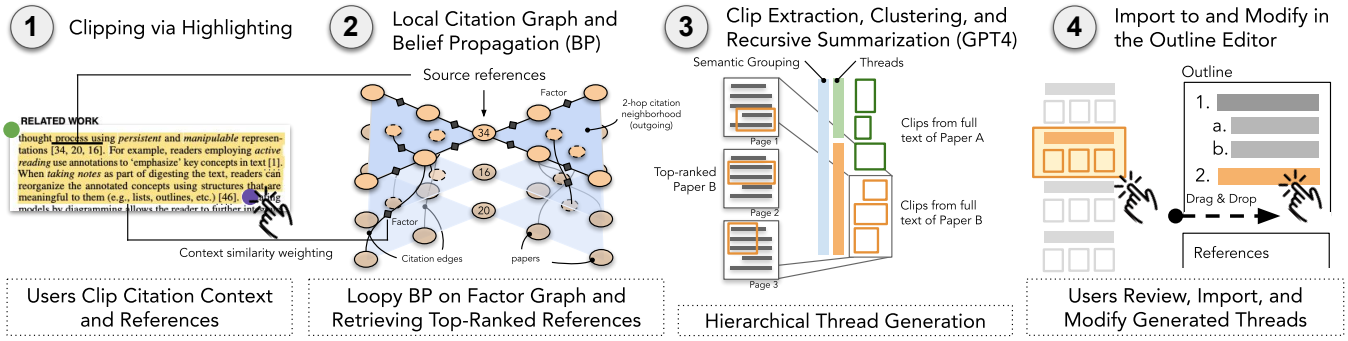

Figure 1: Main stages of Synergi. (A) A scholar highlights a patch of text in a paper PDF that describes an interesting research problem with references. (B) The system retrieves important papers specifically relevant to the highlighted context in terms of how they have been previously cited by other scholars, via Loopy Belief Propagation over a local 2-hop citation graph from the seed references (Section 4.1). (C) Relevant text snippets extracted from top-ranked papers are hierarchically structured and recursively summarized using GPT-4 in the chat interface (Section 4.2). (D) The outline of threads, supporting citation contexts, and references are presented to the scholar for importing, modifying, and refactoring in the editor (Section 4.3 and 4.4).

for learning, iteration, and synthesis. Although chat-based interfaces for LLMs can be helpful in various use case scenarios, they do not support users to easily extract useful parts of the output, to iterate on it by incorporating new information, or to incorporate supporting evidence, which are essential interactions for iteratively synthesizing knowledge from multiple documents. Furthermore, despite great potential for augmenting synthesis workflows, LLMs suffer from hallucination and falsehood (*cf.* [5, 7, 47]), rendering their outputs uncertain, less trustworthy, and needing manual inspection and verification. Instead, in this work we explore using LLMs as a component in a larger computational pipeline that constrains their scope to more tightly bounded summarization and synthesis goals, and enabling an alternative (non-chat) interactive interface that more directly supports users' needs.

Here, we propose a novel mixed initiative workflow, Synergi, that augments scholars' existing synthesis workflows by providing them a structured outline view of research threads, which they can interactively review, curate, and modify. Synergi incrementally expands on user-curated threads that combine rich natural language descriptions and corresponding citations. Threads serve as boundary objects, translating user interests during literature exploration into signals for AI-based outline generation, supporting scholars to move between the *bottom-up* and *top-down* workflows of scholarly synthesis, and help them combine the best of both worlds in the process. Synergi-generated research threads relate specifically to a query clip and seed references, that may match only on a specific citation context within a paper rather than its entirety, and can directly help scholars with making sense of existing threads of research in an area and understanding their relations. Synergi accomplishes this by automatically retrieving a set of important papers from a 2-hop neighborhood on the citation graph and summarizing them in a hierarchical manner with a synthesized label for each parent node that captures the core commonality among its children. In contrast to prior approaches that supported largely manual *bottom-up* synthesis workflows, Synergi synthesizes threads from multiple papers and organizes them into a hierarchy that allows users to quickly discover most relevant threads and understand them through synthesis by other scholars, described in the citation contexts in their papers, that are provided together. Furthermore, in contrast to *top-down* LLM-based workflows that may generate difficult-to-inspect black-box outputs, Synergi-generated threads maintain rich provenance and context to help users relate and inspect them further by following up on the source papers and the specific parts in their body text.

Through case studies and a controlled laboratory experiment where domain experts compared the quality of user-generated outlines from Synergi against those of a baseline system based on Threddy and a GPT4-based approach using the chat interface (henceforth referred to as Chat-GPT4) blind-to-condition, we found that Synergi resulted in the highest overall helpfulness ratings from expert judges. Our quantitative analysis showed that the overall helpfulness of outlines from Synergi was 1.6-point higher compared to Chat-GPT4-generated outlines and 2.6-point higher compared to Threddy-based outlines (on a 7-point Likert scale). In addition, experts judged that threads in the Synergi condition were better-supported with evidence from the literature compared to the Chat-GPT4 condition (+Δ3.3) and the Threddy condition (+Δ2.3; both on a 7-point Likert scale). Through quantitative and qualitative analyses of users' interaction logs, interviews, and responses to experience survey questions, we found that Synergi encouraged higher-level thinking around what existing salient threads of research are and how they divide the space, increased curiosity in them, and boosted confidence in conducting a literature review. We also found that these benefits likely came from efficiency gains over a *bottom-up* Threddy-based baseline, and from gains in coverage of synthesis compared to a *top-down* Chat-GPT4 baseline. We discuss these results and conclude with design implications for future AI-augmented scholarly synthesis systems and workflow designs.

In sum, the contributions of this paper include:

- Synergi, a novel mixed-initiative workflow consisting of retrieval and organizational algorithms and interaction features to support scholarly synthesis.
- The results of a controlled laboratory and case studies involving expert judges and detailed quantitative and qualitative

analyses of user interaction logs, interviews, and surveys uncovering the benefits and challenges of the approach.
- Implications for future workflow designs and relevant research inquiries in this area.

2 RELATED WORK

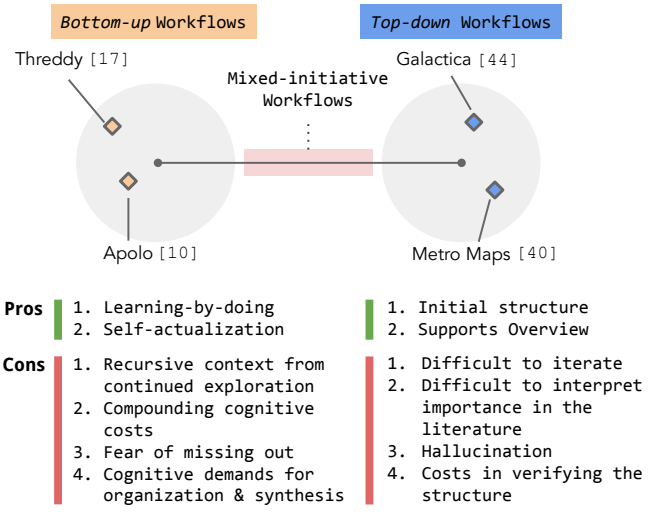

Figure 2: The spectrum of workflows, with pros and cons in *bottom-up* and *top-down* workflows and relevant prior work.

2.1 *Bottom-up* Scholarly Synthesis Workflows

Workflows closer to the *bottom-up* end of the initiative spectrum involve practices such as forward and backward citation chasing and footnote chasing as integral parts for scholars traversing citation graphs to discover important papers related to a research problem [36]. However, these practices often suffer from fragmented information environments and piecemeal tooling [52] that add complexity to their workflows and may take focus away from synthesis. Moreover, while scholars can greatly benefit from reading the related work or introduction sections of a review paper that already synthesizes a relevant domain [46], they are not ideal for iterative exploration of the literature for synthesizing one's own review that may differ in focus, scope, or framing. Scholars would need to read multiple review papers for their synthesis [35], which can quickly increase the cognitive and interaction costs (Fig. 2 left, Con 1 & 2).

Systems such as Apolo [11], CiteSee [10] and Threddy [18] aim to address these challenges by helping scholars iteratively discover or group papers (Fig. 2 left, Con 3). Additionally, systems like Fuse [28], ForSense [39], Passages [12], Mesh [9], and Threddy [18] support clipping and organizing clips on-the-go, reducing the interaction and context-switching costs involved in the process. While helpful, interactions in these systems focused on supporting users with manually saving clips or discovering additional papers, rather than synthesizing knowledge after the early stages of discovery and foraging in sensemaking [37]. This leaves much of synthesis – relating different clips, grouping references, and creating a coherent outline that describes multiple salient threads of research based on the data – as manual work to scholars (Fig. 2 left, Con 4).

2.2 Systems that Support *Top-down* Synthesis

On the other end of the spectrum are systems such as ConnectedPapers[4] and Metro Maps of Science [41] that provide a *top-down* visual overview of the research landscape to aid scholars in making sense of the structure of knowledge space and to discover interesting parts in it. While such representations can serve as a great entry point to a knowledge domain that may be new to the user, they tend to not support additional user interactions beyond overview which limit their utility as a tool for synthesizing knowledge scattered across multiple papers (Fig. 2 right, Con 1).

Furthermore, recent systems' advances in Large Language Models (LLMs) such as Galactica [45], ChatGPT and Google Bard demonstrate impressive capabilities in answering user questions using the knowledge seemingly synthesized from the Web, and tools such as Ask Your PDF[5] show promising avenues for future systems that could support additional personalization and specification based on a set of user-curated documents. While these LLMs show great potential for augmenting scholarly synthesis workflows, they also suffer from challenges such as hallucination and falsehood (*cf.* [5, 7, 47]) that render their outputs uncertain, less trustworthy, and needing manual inspection and verification (Fig. 2 right, Con 3 & 4). Moreover, the process of their computation is obscured [4] and less interpretable to users [29, 53] which further reduces their chance of learning, iterating, and synthesizing based on their outputs (Fig. 2 right, Con 2).

2.3 Systems that Augment Scholarly Discovery

A large body of work exists in scholarly discovery [30], including PaperQuest [38] which allowed users to input query papers to receive other relevant papers based on citation relationships; Sturm [43] which studied requirements for literature search systems and developed LitSonar where users could deploy nested queries to query over multiple sources of document streams; LitSense [44] which included multiple citation relation visualizations and supported filtering and querying for homing in on specific references for further exploration; search and recommender systems that leveraged citation graphs [19, 22, 23] to support relevance features in paper recommendations; diversifying scientific literature search [20, 21]; and Relatedly [35] that developed an approach to recommend relevant unexplored paragraphs in related work sections of papers. Compared to prior work, Synergi's retrieval algorithm simultaneously optimizes for the semantic citation context similarity and the likelihood of a candidate paper building upon related threads of prior research. Synergi uses Loopy Belief Propagation (LBP) and a novel message weighting scheme on a local citation graph to find papers most likely to be important in reviewing related literature, exploring new grounds beyond prior application of LBP in sensemaking over citation graphs [11].

3 USAGE SCENARIO AND DESIGN GOALS

3.1 Synergi Usage Scenario

Consider a research scientist who wants to write a summary of notable threads of research about *how HCI professionals design*

[4]https://www.connectedpapers.com/
[5]https://askyourpdf.com/

human-centered AI systems, a topic she recently started exploring. She uses Synergi to open up the PDF of a paper that she saved earlier on the topic. As she reads through the introduction and related work sections of the paper, she finds several sentences with citations to prior work that describe notable advances in the literature. She clips the sentences by directly highlighting them in the PDF. After saving a few clips from the paper, she is interested in a thread around *challenges designers face with problem setting or ideation with AI*, and wonders what other related research threads there might be. She quickly inputs her saved clips on Synergi.

Based on the input clips, Synergi recommends threads and their high-level grouping that she can quickly scan to understand how different sub-group structure maps to the broader literature. This understanding allows her to orient her attention towards specific areas that align with her interests. Because threads are organized in a hierarchy with rich provenance information such as the source references and exact excerpts from them that support each thread, she can quickly identify threads that look particularly interesting in the literature, and find important references in them. For some of the references she is not sure about or wants to understand in more details, she can examine the relevant sections in the paper that support each thread that are presented together as excerpts.

After reviewing individual threads that were especially interesting to her, she can easily curate useful threads, references, and contexts from the provided hierarchy into an editor using drag-and-drop where she synthesizes an outline and iterates on it. After forming and iterating on her own initial synthesis outline, she has a few new references included in the outline that support individual threads. She can quickly prioritize references that are more frequently cited in her hierarchy of research threads by using the group-by-reference view at the bottom of the editor. This view gives a ranked list of references by the number and context of threads they appear under, which gives her an at-a-glance measure of 'representativeness' a reference has to the threads included in the hierarchy. She clicks on the first ranked reference to explore additional threads of research that may further expand the initial hierarchy she is building. Synergi automatically opens the PDF in its reading interface, and she continues the literature review.

3.2 Design Goals

Motivated by the challenges with existing tools and workflows described in the usage scenario, our design goals are as follows:

- **[D1]** When reading one research paper, allow scholars to clip passages and references of interests, and help them find important papers in the domain for synthesis, specific to query context and seed references.
- **[D2]** Based on clips and references collected by a scholar, the system should provide a structured outline of salient research threads to support their synthesis across multiple papers.
- **[D3]** Help scholars understand the specific research contexts described in each thread in detail, and verify their sources.
- **[D4]** Help scholars review the system-generated threads, curate ones that most interest them into their own outline, and iteratively build upon it.

4 SYSTEM ARCHITECTURE

The system consists of two primary backend algorithms and two sets of interface and interaction features corresponding to the design goals described above.

4.1 Retrieving important papers specific to user's query citation context (D1)

4.1.1 Background: Application of Loopy Belief Propagation for Sensemaking. Loopy Belief Propagation (LBP) [50] is a message-passing algorithm well-suited for iterative sensemaking over graphs that may contain cycles. LBP has previously been applied to sensemaking over citation graphs [11] due in part to its favorable qualities such as simultaneously being able to start from multiple entry points on a graph (e.g., multiple references in a user clipped paper passage), and supporting soft clustering (allowing each paper to belong to more than one research topics; see also Related Work in [11] for additional discussions of the algorithm's advantages over alternatives). While LBP on graphs with cycles may risk non-convergence, in practice the risk is extremely low on citation graphs due to the chronological ordering of citation edges leading to broken cycles and weak correlation [3].

Different from Apolo [11], in our workflow users start by specifying input that consists of *the initial set of seed references* as possible exemplars on the citation graph, along with the *citation context described in natural language in which they were referred to*. This setting does not assume user supervision is provided in an iterative manner throughout the process of discovery to prevent propagation of errors.

While previous use of Loopy BP over citation graphs only considered a set of user-provided seed papers to help discover additional papers [11], users in Synergi clips passages and references as they read a paper to discover relevant research threads and papers. To incorporate this additional context (i.e., text passages) into Loopy BP, we introduced a new multiplicative objective for context-sensitive message weighting (See Appendix A.1 for a detailed description), that goes beyond the constant message weighting scheme used in [11]. Intuitively, each component of the new multiplicative message weighting objective corresponds to the context similarity and reference overlap, respectively, optimization of which prioritizes papers that simultaneously meet the conditions of 1) that they are referred to in semantically related ways by other scholars in their literature reviews (typically appear in the introduction and related work sections of the paper) and 2) that they build upon related threads of research, represented by the overlapping set of references that they cited. Upon LBP convergence, probability distribution ranges from 0 to 1, with higher numbers representing greater relevance likelihood, generating rankings.

4.1.2 Construction of a factor graph using the 2-hop citation neighborhood. We run the LBP algorithm over the local citation graphs sourced starting from the seed references provided in the user clip. In order to construct a candidate set of papers for retrieval (Fig. 1, ②), the system dynamically fetches the 2-hop citation neighborhood using each of the seed references in both directions (*i.e.,* incoming citations and references) using the Semantic Scholar APIs [24]. For each seed paper referenced in a clip, this allowed Synergi to

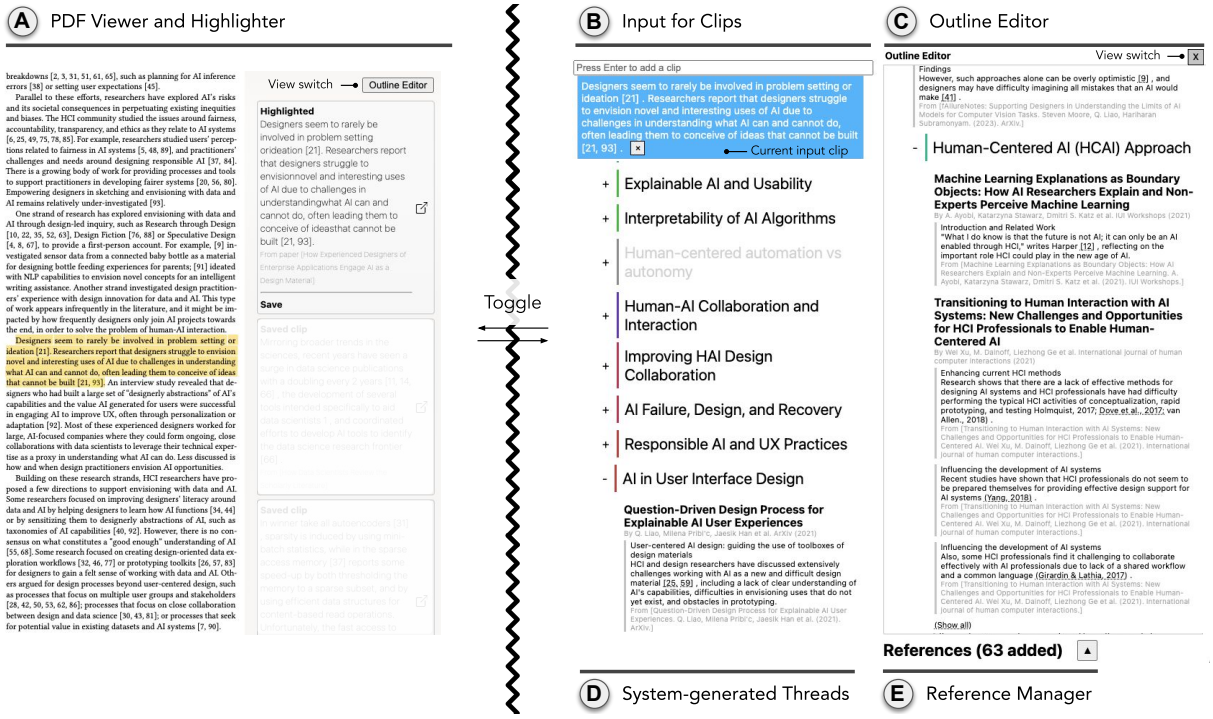

Figure 3: Two main interfaces of Synergi. (A) The PDF viewer and in-text highlighter is similar to that of Threddy [18], with a simplified stream of user-collected clips shown on the right. When the user clicks the 'Outline Editor' button, the view switches to the editor mode. (B) In the top lefthand side corner is an input for user-collected clips where keywords of clips can be typed in to trigger a dropdown menu (not shown). Users can also click on "Try these clips" button to see the most recently saved clips for convenience of their reference (not shown). When the user adds a clip in the input, Synergi kickstarts the pipeline to generate a 3-level hierarchy of salient research threads in the literature specific to input clips. (C) Users can interact with the outline editor to curate interesting threads and citation contexts from the hierarchy (Section 4.4). (D) Synergi-generated threads and grouped citation contexts are made draggable for user curation into the editor (Section 4.3.2). (E) The reference manager automatically updates upon changes in the editor content (Section 4.3.3).

fetch up to 50 most cited incoming or outgoing citations and 50 references for each hop, resulting a total of 50 * 50 * 2 = 5,000 candidate papers. Once the 2-hop citation neighborhood is retrieved for each seed reference, we construct our factor graph with each unique candidate paper as a variable and use the citation edges as factors connecting the variables. To more deeply consider how each candidate paper is semantically relevant to the user clips, we also retrieve from the APIs information about each candidate papers including the titles and citing contexts. These information were stored as annotations on each edge in the factor graph. Since a paper can be cited by the same paper multiple times in different contexts, each edge may end up with multiple citation context annotations. Furthermore, each variable can be connected to multiple papers that have citation connections with it, allowing Synergi to capture different ways a candidate paper had been characterized by other scholars.

4.1.3 Acquiring and parsing top-ranked paper PDFs. Prior work [18] showed that specific citation contexts and synthesis already provided by other scholars (often appear in the related work or introduction sections of a paper) are useful for scholars' sensemaking and literature review. In order to extract them, we developed a full-text PDF acquisition and parse pipeline. First, we ran the LBP algorithm described above until convergence to find 30 top-ranked papers to search for their full text PDFs. Then the pipeline initially searches the S2ORC corpus [31] to see whether a corresponding full text PDF URL is available for each paper. In cases where a PDF URL was not available in the S2ORC corpus, the pipeline uses the Google Custom Search API[6] to search for a matching paper title and its PDF URL using the "filetype:PDF" constraint. After obtaining a PDF file from the URL, the pipeline uses GROBID [1] to parse the PDF and extract the citation contexts along with metadata (*e.g.,* page number that the citation context appeared on, the header of the section containing the citation context, etc.) and the information of the references included in them to render in tooltips. Finally, if a candidate paper fails to fetch its PDF or be parsed, the pipeline defaulted to the paper title and abstract as its content.

4.2 Generating Salient Threads of Research (D2)

Using the top-ranked papers from previous steps, Synergi generates a structured summary of multiple relevant threads of research in the area (Fig. 1, ③). This consisted of steps to home in on specific citation contexts in the papers, structure them into a hierarchy,

[6] https://developers.google.com/custom-search/

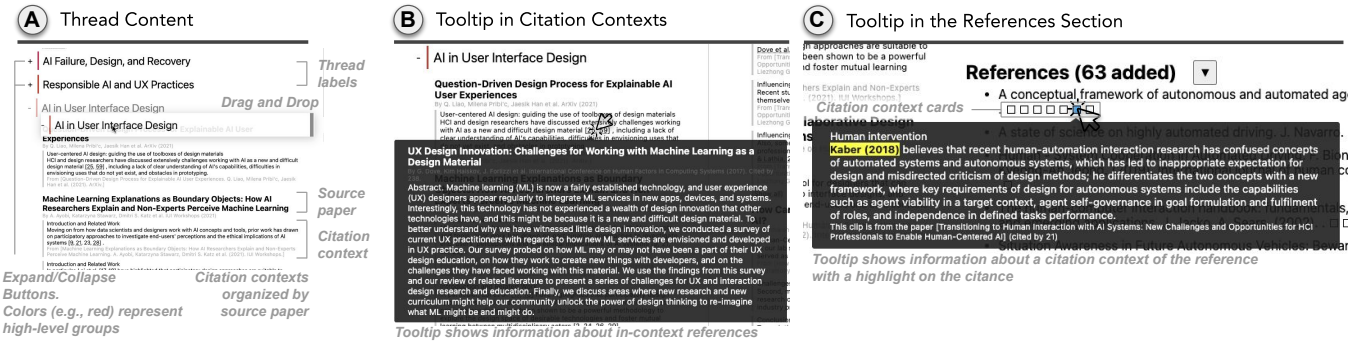

Figure 4: Interface features. (A) The thread outline view is organized using an indented tree visualization. Threads and clips are visually differentiated using colors (the latter always featured a grey bar) as well as information organization. The citation contexts for each thread were grouped by the source papers and presented as a list. By default 3 contexts were shown; clicking on a [show more] button at the end of the list expands the list (not shown). (B) Mouse over each reference in a citation context (dotted and underlined for feature visibility) showed a tooltip that contained information about the reference. (C) The reference section at the bottom of the outline editor was automatically updated with each reference featuring citation cards; mouse over on a card showed a tooltip that contained the citation context information.

and summarize them to capture core commonalities among the lower-level components in the hierarchy.

4.2.1 Filtering citation contexts most relevant to seed clips. To synthesize relevant information scattered across the multiple top-ranked papers identified from the retrieval algorithm (Section 4.1) into a hierarchical structure using relevant text from them, Synergi embedded the extracted citation contexts using `text-ada-002`, and filtered those that have a higher average cosine similarity to seed clips than `0.80`[7]. We used the S2ORC dataset [31] covering multiple citation intents for comparison. While not discerning context type, our pilot demonstrated functionality when combined with the context similarity thresholding.

4.2.2 Agglomerative clustering and tree-cutting. To present the most relevant topical clusters to the users, Synergi first uses the embeddings of the filtered citation contexts to measure how relevant they are to the user clip. For this, Synergi constructs a hierarchical structure from them using a unsupervised agglomerative clustering with the Ward linkage. We perform this using the `fastcluster` package [33]. Agglomerative clustering initializes citation contexts as singleton clusters and computes the ward distance of each pair to successively merge the most similar clusters. The result is a hierarchical binary tree (Fig. 12) where the height of the joint of branches represents the distance at which they were merged (the higher the height of the common ancestor of two leaf nodes on the hierarchy, the more distant they are as neighbors). The resulting binary tree is then converted into a 3-level hierarchy for the user to explore (see Appendix A.2 for a description of the rationale and the method).

4.2.3 Recursively summarizing the children clusters. To help users explore the 3-level hierarchy, Synergi synthesizes labels for each parent thread that succinctly describes the underlying threads or citation contexts. In order to synthesize labels that are simultaneously coherent with the underlying children nodes' texts and are abstractions of them, we traverse the hierarchy in a bottom-up manner to recursively synthesize labels. We use Chat-GPT4 with a prompt (Fig. 13 in Appendix A.4) that instructs it to summarize the underlying text using 6 words or less. In each pass on a parent node, up to 25 text snippets from its children were provided during prompting. Therefore, in the first pass the 25 cluster citation contexts were added to the prompt and in successive runs, the text of the children clusters' synthesized labels were used. We also added a post-processing step to merge similar threads (see Appendix A.3 for a description of the rationale and the method) and assigned a unique color to each top-level thread such that the similarity among the children threads could be visually indicated later on the interface.

Finally, the 3-level tree structure with salient threads and their labels, along with the most relevant citation contexts attached to each, are returned to the front-end to render an overview of the relevant research landscape and salient threads in it.

4.3 Interface Features (D2 & D3)

4.3.1 Walk-through of the interface. Users on Synergi can highlight and clip relevant citation contexts directly from paper PDFs they are reading. Once they have one or more clips they are interested in investigating further, they switch to the editor view by clicking on the 'Outline Editor' button from the PDF viewer (Fig. 3 Ⓐ). In the Outline Editor view, the user can select one or more from the list of saved clips to generate structured research threads related to the citation context and seed references included in selected clips (Fig. 3 Ⓑ).

Once the system finishes processing, the structured thread recommendations appear under the clip input (Fig. 3 Ⓓ). The user can review the content by scrolling through the list and by expanding/collapsing individual threads which contain the detailed information about citation contexts related to the thread, grouped by source papers. The colored bars on the left also provide users with high-level research areas to quickly orient themselves among the surfaced research areas and help guiding their attention to interesting ones. When the user identifies interesting threads, they can curate them into the outline they are building by dragging

[7] determined through a small scaled experiment with five example clips during development

and dropping the threads from the list on the lefthand side into the outline editor (Fig. 3 Ⓒ), into the appropriate location on the hierarchy. The reference section below automatically updates based on the content changes in the editor, providing the users an easy access to information about papers that have been most cited across multiple threads and citation contexts, which help them prioritize what to read next. The user can continue the cycle by opening up a new paper in the PDF viewer and switching between outline editor. The user data persists for iterative development and refinement.

4.3.2 Tree-structured thread recommendations. The tree-structured thread recommendations can be expanded and collapsed to reveal the relevant citation contexts below, which are grouped by source papers (Fig. 4 Ⓐ), to provide users with easy access to the source materials and increase the verifiability. Each thread label also featured a color bar on the left to indicate semantically similar groupings among different threads. Each citation context included the specific context found from the paper, the section header that it appeared in, as well as other metadata about the source paper.

4.3.3 Citation context and reference tooltips. To help scholars quickly gain additional information about the cited references in each citation context, each citation notation (*e.g.,* '[4]') was rendered with a dotted underline (Fig. 4 Ⓑ), with an additional tooltip that reveals information about the reference such as its title, publication year and venue, number of citations, author names, and the abstract over a mouse-over. In the references section under the outline editor, each referenced paper was automatically updated when the content in the editor changes, and pulled in any citation contexts added in the editor that it was cited in. The grouped citation contexts were shown as squares next to the title (denoted as 'citation context cards' in Fig. 4 Ⓒ), which revealed a tooltip that contains information about the citation context with the corresponding reference notation highlighted in the yellow over a mouse-over.

4.4 Drag-and-Drop Outline Editor (D4)

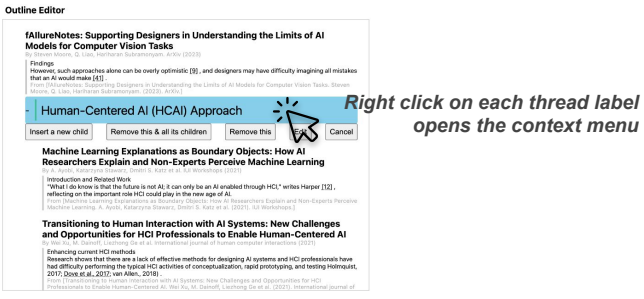

Figure 5: Users could edit the outline either by adding a new thread or citation context into it using drag-and-drop, or by right clicking on each node in the editor.

Threads or individual citation contexts were made draggable into the outline editor. Users could drop the dragged item into any thread node already in the editor or the default top-level thread ('Your Outline'). After the user drops an item to add to the editor, the references section below automatically updated to pull in any new references or new citation contexts for existing references (as shown in Fig. 4 Ⓒ). The added threads and citation contexts in the editor were interactive via right-clicking on them at which point the corresponding context menu was revealed. When a thread was right clicked, the following options were shown (Fig. 5):

Insert a new child: Add a new nested thread node.
Remove this & all its children: Completely remove the sub-tree rooted on this thread.
Remove this: Remove only the clicked thread and moves all its children one level up (equivalent to merging).
Edit: Edit the label of the thread.
Cancel: Close the menu.

Right-clicks on citation contexts showed only the 'Remove this', 'Edit', 'Cancel' options in the menu.

5 EXPERIMENTAL DESIGN

5.1 Objective & Research Questions

Based on a user query as the input, we aimed to study how Synergi-generated threads of research and supporting clips can benefit scholars conducting literature review to cover the broader areas of research. We designed the timed tasks in the experiment to mimic the practice of coming up with a literature review outline for an assigned topic. This is because scholars often craft intermediary outlines before arriving at a fully written article to structure their thoughts, synthesis, and exploration of the literature in earlier stages. We chose two different topics of research based on the papers that our expert judges were lead authors on [16, 51]. To compare different conditions, we measure the quality of the outlines, the efficiency of constructing them, and the participants' perception of Synergi-generated threads and experience. We operationalized the quality of outlines as experts' judgment of the overall helpfulness, and thread-specific relevance, familiarity, and the goodness of the supporting citation context, on a Likert scale from 1 (Strongly disagree) to 7 (Strongly agree). We operationalized efficiency as the number of threads, clips, and references saved in the outline in a fixed amount of time, as well as the number of user actions taken to construct the outline. Our research questions were:

- RQ1. Does Synergi improve the quality of scholars' literature review outlines over the baselines?
- RQ2. Does Synergi improve the efficiency of outline construction over the baseline?
- RQ3. What are perceived benefits and limitations of Synergi-augmented workflows?

5.2 Participants

We recruited 12 people (10M/2F) for the study. Participants' mean age was 26.4 (SD: 2.11) and all actively conducted research at the time of the study (9 Ph.D. students and 3 Pre-doctoral Investigators). Participants' fields of studies included (multiple choices): HCI (10), NLP (4), Information Retrieval (1), Cognitive Science (1). We also recruited two experts (both female) to review participants' outlines. Both experts judges were 5th-year Ph.D. students with multiple first-authored and peer-reviewed publications in HCI venues. Their domains of research were 'cross-functional AI teams in envisioning AI products and experiences' and 'designing and building novel tools to help developers better annotate and share their learning materials'. The expert judges spent 1.5 hours to review participants'

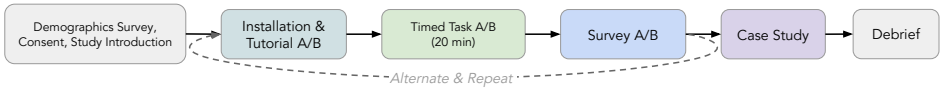

Figure 6: The entire procedure of our study. The order of the middle section of the procedure was swapped based on the assignment (A/B). The order assignment was randomized and counterbalanced across participants (see text).

outlines and were compensated $60 USD. The study lasted for 80 minutes and participants were compensated $40 USD.

5.3 Baseline Implementation

5.3.1 Baseline based on Threddy. The baseline system, based on prior work Threddy [18], supported users in manually curating citation contexts via direct in-text highlighting in the PDF, with persisting clips across papers for increased context awareness, and featured a list of user curated clips on the lefthand side of the editor view that users could drag-and-drop into the editor easily. The user-curated clips replaced the system-generated outline provided in the treatment condition. All other interaction features were identical.

Figure 7: The baseline system was based on Threddy [18] which supported clipping, persistence of clips across multiple papers, and an easy access to the outline editor where users could organize their own outlines using the self-curated clips.

5.3.2 Chat-GPT4 Baseline. We also generated two literature review outlines for each paper used in the main study with Chat-GPT4 on OpenAI Playground[8]. Our prompts requested completion of a literature review that a scholar has started, given the same citation context clip used in the treatment condition. The prompts also included a label of one starter thread. We replaced the citation notations in the input context with actual titles of the references, with clear demarcations, to provide further context about the research topic (see Fig. B.2 in Appendix B.2). The temperature was set to 1 with the maximum generation token length as `2,048`. We repeatedly sampled two outlines for each of the two input paper-clip pairs. The generated outlines were then manually formatted/blinded (*e.g.,* removing auxiliary characters demarcating headers, reference notations, and unifying the style) for expert review.

[8]https://platform.openai.com/playground

5.4 Procedure

Our main study simulates a literature review task using novel domains, with three conditions: a Threddy [18]-based condition (§5.3.1), the Synergi treatment, and a no-human Chat-GPT4 baseline (§5.3.2). The study also triangulates the output quality through third-party expert assessment and case studies in familiar domains where participants can comment on the output quality. The inputs to the Chat-GPT4 baseline were prompts from Fig. 14.

5.4.1 Structure. We employed a within-subjects study design to compare the outline construction process in two human-based conditions – Synergi and Threddy-based baseline. Chat-GPT4-generated outlines were blended to the outlines from the other two conditions for a blind expert quality assessment. We chose two different research areas and topics for timed literature review tasks with a randomized and counterbalanced presentation order, and let individual participants choose personally interesting topic/paper for case studies in the end (thus, three tasks in total per participant). We randomly assigned systems to the topics for the timed tasks. We counterbalanced the order of presentation using 6 Latin Square blocks and randomized rows. Participants followed the following procedure in the study, which took place remotely using Zoom: Introduction, Consent, Demographics survey; Installation and Tutorial (detailed in Appendix B.1) of the first system; Main task for the first system; Survey for the first system; Alternate and repeat for the second system; Case Study based on a personally interesting topic; Debrief. Participants were asked to share their screen during the timed tasks and think-aloud during the case studies.

5.4.2 Timed Literature Review Tasks on Pre-defined Topics (20 mins each). In each of the two timed tasks, participants were instructed to perform a literature review on a randomly assigned topic.The interviewer provided the initial URL to the paper and pointed the participants to the exact location of the clip in each paper that contained the target problem statement. The scenario given to the participants was framed as 'conducting a review of the relevant literature on behalf of their colleague, who is studying a related research question' for motivating participants in unfamiliar domains.

5.4.3 Post-task Surveys. After each task, participants were administered a survey containing questions on their subjective feelings about the experience. Demand (both physical and cognitive) and overall performance were measured using the validated 6-item NASA-TLX scale [13], where a more compact 7-point scale, mapped to the original 21-point scale, was instrumented [40]. In order to probe the compatibility and adoptability of the technology with participants' existing literature review workflows, we included a modified Technology Acceptance Model survey from [49] (4 items). Furthermore, 8 types of benefits around discovery, sensemaking,

outlining, curiosity, confidence, fear of missing out, and organization of clips and references were measured for each system (See Appendix D for details of the questionnaire).

5.4.4 Data Collection. We collected participant-generated literature review outlines at the end of each timed task. The outlines were then transformed into a spreadsheet while preserving the indentation of the original tree structure with additional columns on the left for experts' judgement. Each tree was traversed to tally the number of threads, clips, and references for each participant for analysis. During the experiment, participant's interaction traces (*i.e.,* timestamped action details during timed tasks) on each system were logged. The details of each timestamped action included a unique user ID, time of the action, the type of the action (*i.e.,* clip, import, create, move, edit, remove, merge), and corresponding details. Participants' think-alouds during the case study and debrief were recorded and transcribed.

5.4.5 Experts' Evaluation. The participant-generated literature review outlines were anonymized and blended with two randomly sampled outlines from Chat-GPT4 for each paper (See Appendix B.2 for the details of the prompts used). Therefore outlines were generated from three conditions in total, *Baseline* – the Threddy-based baseline system described in Section 5.3.1, *Treatment*, and the *Chat-GPT4*-based baseline (Section 5.3.2). Experts reviewed each outline independently and blind-to-condition, and evaluated on the basis of the following 7-point Likert-scale (1: Strongly disagree, 7: Strongly agree) questions:

- (Overall Outline Helpfulness) "*I found the outline with supporting context helpful for reviewing the relevant literature.*"
- (Thread Familiarity) "*I found the thread of research familiar.*"
- (Thread Relevance) "*I found the thread of research relevant.*"
- (Thread is Well-Supported by Citation Context) "*I found the thread to be well-supported by the specific citation context(s).*"

The overall helpfulness question was evaluated once per participant resulting in 12 data points in *Baseline* and *Treatment* conditions and 4 data points in the *Chat-GPT4* condition; the three thread-level questions were evaluated once per thread per participant, leading to 108 data points (*i.e.,* 31 in *Baseline*; 10 in *Chat-GPT4*; and 67 in *Treatment*) in total.

5.4.6 Case Studies. At the end of the timed tasks, the interviewer asked participants to find and open the PDF of a paper that they were personally interested in that was also in their domain of research using the treatment system. Each participant highlighted and clipped a patch of text (one sentence or longer) that described a particular research problem that also included at least one citation in it, then generated a list of threads using it in the same way as earlier in the timed task. Once the result has returned, the participants were asked to review the generated list of threads, their semantic grouping, the clips, and the references that the clips had originated from. The interviewer then asked questions about their quality, benefits, and limitations.

5.4.7 Data Analysis. The mappings between the research questions and analyses of collected data are as follows.

- RQ1. We analyzed the quality measures of the outlines, which were on a 7-point Likert scale, using non-parametric tests. For expert-evaluated overall helpfulness of outlines, the Wilcoxon's signed rank test was performed for the paired-samples data (*i.e.,* the Baseline vs. Treatment comparison) and the Mann-Whitney U test was performed for the independent data (*i.e.,* the Chat-GPT4 baseline vs. Treatment comparison). For independent data such as thread-level familiarity and relevance, the Mann-Whitney U test was used.
- RQ2. We analyzed the efficiency measures (*e.g.,* the average number of saved threads/clips/references in 20 minutes and the number of user actions taken to construct the outline) between the conditions using paired Student's t-test.
- RQ3. The Likert-scale and Likert-item responses in the survey data were analyzed using the non-parametric paired-samples Wilcoxon's signed rank test. Participants' comments during the case studies were transcribed and qualitatively analyzed using open coding. Participants' interaction logs were visualized as time graphs and used for triangulating relevant survey responses and qualitative data.

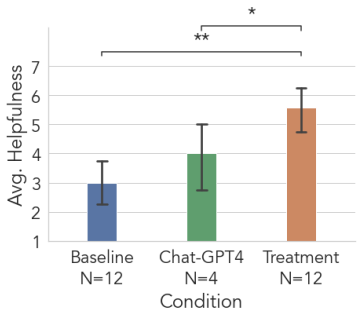

Figure 8: The overall helpfulness judged by experts was the highest in Treatment (M=5.6), followed by Chat-GPT4 (M=4.0) and Baseline (M=3.0) conditions. The pairwise differences between Treatment and others were significant (see text).

6 FINDINGS

6.1 RQ1. Quality of Outlines

6.1.1 Higher quality outlines. Using Synergi , participants were able to generate literature review outlines that were rated as higher quality. The average expert judges' ratings on the overall helpfulness of literature review outlines in the Treatment condition was M=5.6 (SD=1.38), followed by the Chat-GPT4 condition (M=4.0, SD=1.41) and the baseline condition (M=3.0, SD=1.41) (Fig. 8). Both differences between the Treatment and the Chat-GPT4 conditions (two-sided Mann-Whitney U=7, p=0.036) and between the Treatment and the Baseline conditions (Wilcoxon W=4, p=0.003) were significant. The experts were blind to the conditions that each of the outlines were generated under.

6.1.2 Improved support while maintaining relevance and familiarity. We further examined the overall outline helpfulness by comparing between the conditions their component threads' relevance, familiarity, and how well each thread was supported by relevant citation contexts found in the literature (Fig. 9). The results showed that the average thread relevance did not differ between the Treatment (M=5.4, SD=1.32) and the Chat-GPT4 (M=5.7, SD=1.34) conditions, nor between the Treatment and the Baseline (M=5.6, SD=1.74)

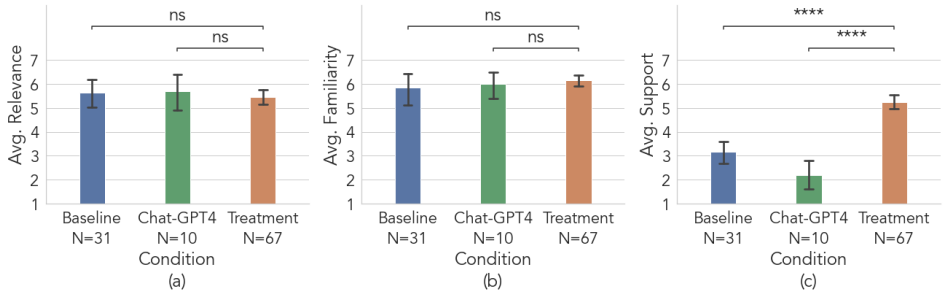

Figure 9: Neither (a) average thread relevance nor (b) familiarity significantly differed between the conditions. (c) However, the average goodness of support from relevant citation context differed significantly, as it was judged higher in the Treatment condition (M=5.5, SD=1.25) than in the Chat-GPT4 (M=2.2, SD=1.03) or the Baseline (M=3.2, SD=1.27) conditions.

conditions. Similarly, the average thread familiarity between the Treatment (M=6.1, SD=1.03) and the Chat-GPT4 (M=6.0, SD=0.94) conditions did not differ significantly, nor did the difference between the Treatment and the Baseline (M=5.8, SD=1.86) conditions. This suggests that while Synergi considered a large set of 2-hop references and citations (more than 5,000 candidate papers), it is able to maintain high relevance to the user query when presenting related research topics.

Further, the average support each thread received from relevant citation contexts differed significantly. Experts' judgement on the goodness of supporting citation contexts was the highest in the Treatment condition (M=5.5, SD=1.25) and positive (between 'slight' (5) and 'moderate' (6) levels), whereas in the Chat-GPT4 (M=2.2, SD=1.03; $p < .0001$) and the Baseline (M=3.2, SD=1.27; $p < .0001$[9]) conditions, it was negative and significantly lower. The goodness of support from relevant citation contexts also seemed to be a differentiating factor of the overall helpfulness of outlines among the conditions; while the relevance and familiarity measures for each thread were highly correlated (Kendall's τ between .45 and .88, $p < .01$ in all cases), support and other measures showed a weak relation at best (relevance-support, $\tau = 0.21$, $p = .04$).

It is notable that despite the lack of supporting citation contexts, both the relevance and familiarity of an average thread generated by Chat-GPT4 tied with those of human-generated threads in the Baseline and Treatment conditions. However, our expert judges noted significant qualitative differences between the Chat-GPT4-generated threads from others, despite not knowing the sources of each outline during the evaluation. The judges proactively offered descriptions of how they differed qualitatively: "[A Chat-GPT4-generated outline was] *Probably the most coherent/thoughtful summarization and distillation of the source paper, but most of the stuff seems like something you could just get from reading only that paper and less of a literature review... no citations in any of the points... although the points are reasonable and feel like informed either by my work or other relevant source.*" (E2); "[After pointing out both Chat-GPT4-generated outlines] *They seem like maybe someone read over some of the citations in my paper and pulled some points from that, but synthesis is generic. Overall, they are both not great as they don't include citations for the points outlined... Numbered lists in both outlines feel as if they were AI-generated, basically too generic to be useful without citations.*" (E1).

6.2 RQ2. Outline Construction Process

6.2.1 Synergi showed significant efficiency gains in the outline construction process. Comparing the outline construction process between the two human conditions[10], the number of research threads, clips, and references saved in the duration of the experiment were all significantly higher in the Treatment than the Baseline condition (Fig. 10a – c). For threads, the average number saved was 6.0 (SD=2.76) in the Treatment condition vs. 3.4 (SD=1.16) in the Baseline condition ($p = .01$). The average number of saved clips was 64.3 (SD=66.27) in the Treatment condition vs. 5.5 (SD=2.81) in the Baseline condition ($p = .01$). The average number of saved references was also significantly higher in the Treatment (M=71.5, SD=63.40) vs. Baseline (M=18.4, SD=9.62) conditions ($p = .01$).

The higher numbers of saved items in the treatment condition could be explained by the overall higher frequency of 'import' actions that users in the treatment condition performed (Fig. 10d) compared to the baseline condition, instead of manually clipping (Fig. 10e). On average, the users in the treatment condition performed 13.3 (SD=9.06) imports vs. 6.3 (SD=2.80) in the baseline ($p = .02$; Fig. 10d) and 0.9 clipping (SD=0.29, Treatment) vs. 7.3 (SD=3.20, Baseline; Fig. 10e) ($p = .00002$). The overall number of refactoring operations (*i.e.,* moving nodes in the outline editor, editing their labels, merging different thread nodes, removing nodes, creating a new parent thread) did not differ significantly between the two conditions (M=12.4, SD=8.44 in Treatment vs. M=12.0, SD=7.75 in Baseline; Fig. 10f, $p = .87$), further suggesting that the efficiency gains originated from replacing the manual clipping of data with examining and importing the system-generated threads and clips in the treatment condition.

6.2.2 Synergi supported both top-down and bottom-up workflows. Interestingly, the users in the Treatment condition exhibited diverging patterns of constructing the outlines. Specifically, some users showed a pattern of top-down construction where they first carefully read through the problem statement and the rest of the source paper to come up with most salient threads of research in their mind before moving on to importing clips that fit those threads, and updating them when a new thread that expands or modifies the initial threads ideated by themselves. Fig. 11 (bottom) demonstrates a prototypical action time-graph which shows a densely populated area of refactoring in the beginning (*e.g.,* in the first 5 minutes in the graph) followed by successive importing. In contrast, Fig. 11

[9]both were tested using two-sided Mann-Whitney

[10]all comparisons in this section were performed using paired t-testing

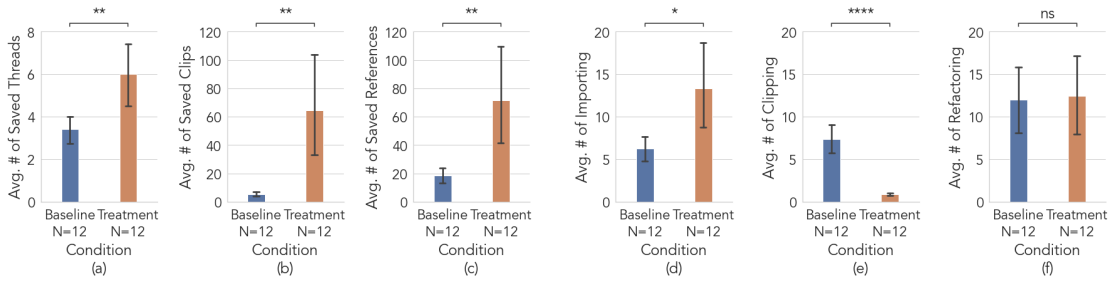

Figure 10: The average (a) number of threads, (b) clips, and (c) references saved during the experiment (fixed length) were significantly higher in the Treatment condition than in the Baseline condition. (d) The differences in the saved numbers could be explained by how much more efficiently users in the Treatment condition imported system-generated outputs, rather than (e) spending time in manually clipping the relevant citation contexts, while (f) performing an overall similar amount of refactoring after adding new items to the outline editor.

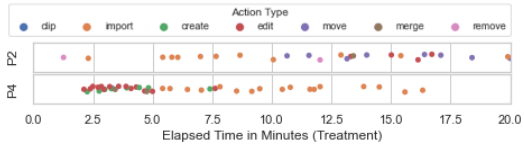

Figure 11: (Top) A prototypical time-graph of user actions demonstrating a bottom-up approach of constructing the outline. (Bottom) Same for a top-down construction approach.

(top) demonstrates a prototypical time-graph for a bottom-up construction approach. In this case, the participant (P2) first imports a number of system-generated threads and clips onto the editor on the right, then moves on to refactor them (*e.g.,* past the 10 minute mark) to work towards a personally interesting outline.

6.3 RQ3. Perceived Benefits and Challenges with Synergi-augmented Workflows

Quantitative analysis of survey results and qualitative analysis of interviews uncovered different types of benefits from Synergi, such as encouraging participants to gain a higher-level perspective about the literature, thinking about threads' relations, and increasing their curiosity. They also uncovered limitations of Synergi-augmented workflows such as additional refinement need related to identifying concepts at a similar level on the conceptual hierarchy, support for probing the relations among threads, and the desire to see explanatory relevance signals for user trust and acceptance.

6.3.1 Reviewing Synergi-generated threads encouraged broader perspectives, sensemaking, and curiosity. Participants commented on how having a list of automatically generated threads of research pushed them to think more broadly about the research space. P1 mentioned that the threads "*help you visualize the literature review outline in your head*" and "*provide better and more context, especially useful for a new topic*" (P1). Relatedly, P4 commented that:

> "*This is giving me a super-power to even begin to think at the level of 'how are different threads of research dividing the space?', which would've been impossible for me to do otherwise.*" – P4

Compared to how they typically conduct a literature in a new domain, they described feeling like saving a lot of time and cognitive effort ("*I usually have to scroll back and forth so many times*" – P2; "*Overhead is significantly reduced... I can now just read, copy-paste, and re-organize stuff*" – P3) that would have otherwise interfered with forming higher-level perspectives. Participants' responses to the survey question: "The system helped me discover relevant threads of research in the literature." also significantly favored the treatment condition (M=6.3, SD=0.75) over the baseline condition (M=3.3, SD=2.14; $p = .009$[11]). Participants also felt as though the "*colors denoted good groupings of threads, for example this brown (color) shows a group about 'Evaluation of toxicity' which was the core question in our research project.*" (P7) and that "*the thread titles are pretty informative. I could easily tell what I should be paying attention to.*" (P8). Interestingly, P1 commented on how "*it's refreshing to find threads on definitions and studies of 'social capital' that may differ in non-western and global south's regional context of use*" (P1) because manually chasing the citations alone tend to get you "*sucked into*" the "*West-dominant*" perspectives in the literature, since "*asymmetry in the citation behaviors exists between the western and non-western bodies of literature*" – P1.

Furthermore, participants' responses to survey questions: "The system helped me make sense of relevant threads of research in the literature." (M=5.3, SD=1.66 in Treatment vs. M=4.3, SD=2.00 in Baseline, $p = .088$) and "The system helped me outline a review of the literature." (M=6.1, SD=0.67 in Treatment vs. M=5.1, SD=2.02 in Baseline, $p = .089$) showed marginal significance between the two conditions at $\alpha = .10$.

Participants commented that the list of papers included in the references section of the outline, automatically extracted from the imported clips, was particularly relevant and contained "*inspiring papers to read in this area*" (P7) and some that one participant wanted to take home ("*Can I get a copy of the list on the left?*" – P11). P10 also described how the list "*Matches the threads and references that I curated for my own on-going literature review of the domain, which is good*" (P10). Participants' responses to the survey questions also showed significant preference for the treatment condition over the baseline condition in terms of boosting their curiosity around different threads of research (M=6.0, SD=0.74 in Treatment vs. M=3.9, SD=1.73 in Baseline; $p = .01$), confidence in conducting the literature review (M=5.8, SD=0.94 in Treatment vs. M=4.0, SD=1.71 in Baseline; $p = .01$), and in reducing the fear of

[11] participants' survey responses were compared using Wilcoxon's signed rank test

missing out on important research (M=5.2, SD=1.22 in Treatment vs. M=3.2, SD=1.64 in Baseline; $p = .01$) (See Appendix D for the details of survey questions).

6.3.2 Trade-offs between Completeness vs. Information Overload. While participants reacted favorably towards the utility of Synergi in the context of the timed literature review outlining task ("*This is a great starting point for a literature review*" – P10), they also commented on limitations that point to future research directions. One of the common concerns for longer-term use of Synergi raised by participants related to how to make sense of the quantity of threads presented to them. On the one hand, "having this many, around 20 or so threads would overwhelm me easily" (P10) and especially "seeing similar threads, even though I like how they are grouped together using the same color, could really overwhelm me" (P4). On the other end of the spectrum, seeing a widely varying number of threads returned for queries made P8 wonder if "the result here is complete in this area because I only got 5 threads for this query. Or am I missing something important?" (P8).

6.3.3 Additional Support for Refining and Relating Threads. Participants also commented on how in some cases the variations among the threads within the same high-level color group may be insignificant yet repeated, leading to visual clutter and information overload: "*[Newcomer Integration in OSS Projects] and [Newcomer barriers] are too similar, they can be merged*" (P10); "*[Prompt engineering in NLP models] and [Prompting in Natural Langugage Processing] feel really similar*" (P7). On the other hand, participants also pointed out threads that were seemingly too narrow in scope for them to be at the same level as other threads that seemed to synthesize across multiple papers: "*The [Skip-thought] thread is kind of weird to have be its own category because it's the name of a specific technique from a single paper.*" (P6); "*[Numeric and logical reasoning] is focused on a very specific aspect of the papers in it, which I appreciate but feels too specific to be included in my review.*" (P7).

P4 described how the threads of research helped him 'lift' his perspective going into the literature review task which was beneficial. However, he also described how he was trying to interpret the relations and the order among different threads within each group and between differently colored high-level groups, and how he wished to "*also be able to reason about what the overlapping spaces are between the threads, for example in a 'Venn diagram' of the research space... which is hard to do with a list of threads.*" (P4).

An interesting sub-thread emerged in this topic when participants examined some of the 'and' conjugated threads and found examples where the phrase before and after the 'and' were at different levels of conceptual abstraction. Often the problematic cases featured one concept that felt too broad to be meaningful in relation to the other concept in the thread. For P10, a thread titled '[Augmenting scientific reading] and [machine learning]' was a clear demonstration of how the 'and'-conjugated concepts could appear at different levels of abstraction, with the second concept in this specific example (*i.e.,* machine learning) being too high-level to be useful. Similarly P6 pointed out two examples, '[Text classification] and [feature weighting]' where the first concept was too broad to be meaningful, and '[Image Captioning] and [Computer Vision]' where the second concept "*did not feel like adding useful information*" (P6).

6.3.4 Scaffolding explanatory relevance information for trust and confidence in recommendations. Last but not least, participants wished to see additional information to understand how each thread was generated, and efficient at-a-glance information around which specific aspect in the query each clip is relevant to, in order to boost their confidence and trust in the recommendations. P10 said that:

> "*Understanding the sourcing mechanisms would help me gauge how much trust I should be lending to the system and stay vigilant for potential failure modes, because there are so many different kinds of relations that could be surfaced, for example 'is it (relation) by authors? venues? publication years? topical similarity?' which makes me want to understand more.*" – P10

For some, being able to group threads by a given paper was desired for helping orient their sensemaking process. P11 commented that "*In my process I move between papers when conducting a literature review... Here, some of the clips look similar to one another and I can see how the same paper is touching on different threads and I appreciate that the system has added clips from the same paper across multiple relevant papers... but it would be nice to be able to see which other threads that this paper has been added to so that I can quickly decide whether to read that paper in more details.*" (P11). P12 commented that "*It would be helpful if I could see the connections between a thread and each clip in the thread because there are a lot of clips in this thread... and I want to quickly go through them, discarding the ones that look tangentially related.*" (P12).

7 DISCUSSION

In this work, we design and develop Synergi as a mixed-initiative system for high-level literature exploration and scholarly synthesis, evaluate its benefits and challenges, and study implications for future systems aimed at augmenting scholars' workflows for synthesizing knowledge from many papers in a domain.

From evaluation studies we found that study participants engaged with Synergi-generated threads of research to broaden their perspectives and free their cognitive bandwidth to focus more on higher-level thinking about salient threads and relations. Interestingly, expert judges found the Chat-GPT4-generated outlines were surprisingly well-synthesized and "thoughtful," distilling key points about the target problem statement. Supporting this observation, the average expert-judged familiarity and relevance of threads did not differ between both human-generated and Chat-GPT4-generated threads. However, expert judges also thought the helpfulness of outlines depended significantly on the scope of its content and the quality of supporting citation context derived from relevant papers in the literature. By examining the outline construction process, we found that the efficiency participants gained in foraging and making sense of research space in Synergi allowed them to broaden their scope of synthesis and to incorporate more relevant papers and supporting citation contexts into their outline. Taken together, these findings suggest that while LLMs such as GPT4 made remarkable advances in condensing scholarly text on demand, synthesis across multiple papers from the broader literature with supporting context remains a uniquely human capability today, albeit human scholars may be challenged by limited cognitive bandwidth while performing literature review and synthesis.

7.1 Thread-focused Workflows and Expansion

Our examination of user interaction logs also revealed two salient behavioral patterns during synthesis around how and when they incorporated the Synergi-generated threads into their own outlines which we labeled as *top-down* and *bottom-up* synthesis workflows (§6.2.2). In the top-down workflow, users often started by processing the problem statement in more depth compared to the bottom-up workflow, and read surrounding contexts in the source paper more deeply to form an initial understanding of their own, and then distill their understanding into an initial outline. In our evaluation participants using this workflow tended to have deeper prior knowledge in the research area that they could draw upon in creating the initial structure. Once appropriate empty threads in their initial structure were identified, they subsequently imported relevant system-generated threads into them.

In contrast, in the bottom-up process participants often started off by iteratively importing system-generated threads into their editor on an individual thread basis, and creating ad-hoc parent threads when they find commonalities among existing threads. Though lacking initial outline structures, this workflow was popular among the participants most of whom were new to the subject domains in the experiment. Bottom-up workflows on Synergi were made possible by using input threads as boundary objects for AI to pre-process other papers along related threads, forming an initial hierarchy with supporting references and context. Their popularity may also have been a result of increased user reliance, caused by Synergi's generation shifting the cost structure of sensemkaing [25], incentivizing user reliance for economic decisions [27].

Centering threads that capture core abstractions of references and citation contexts as first class objects in interaction design also opens up other new design spaces. Possible future work includes thread-based AI-search and self-organization (*e.g.,* autonomously organizing snippets or pulling content from other papers to seamlessly expand the structure) or creativity-increasing retrieval (*e.g.,* targeting threads with generative potential, featuring core thread similarities and peripheral divergence). Another future work direction could explore threads' different use contexts such as augmenting reading interfaces. For example, enabling an ambient 'always on' mode that progressively suggests relevant snippets and threads from the same paper (*e.g.,* synthesized from later sections in the paper, allowing users to quickly scan the rest) or different papers (*e.g.,* supporting user transitions and further building on main threads).

7.2 Implications for Mixed-Initiative Workflows

Our expert evaluation showed that fully AI-generated synthesis was competitive against outlines synthesized by human users in a manual or an AI-augmented workflow in terms of coherence and distillation when the scope of synthesis was limited. Future LLMs with a sufficiently larger context window may overcome this issue via new capabilities in processing many papers at once.

However, even with an improved AI, a fully automated workflow may not be the optimal design for systems aimed at supporting scholarly synthesis. 'Putting in the work' during the literature review may be critical for scholars' learning and building up a necessary repository of knowledge for successful synthesis later on. Rather than adopting a design that may disincentivize self learning and self-actualization [32], successful mixed-initiative systems therefore would need to consider tasks that AI augmentation can be most beneficial *without* interfering with core cognitive tasks and human learning. For example, future workflow designs may selectively delegate tasks involved in synthesis based on their high vs. low importance or the core vs. periphery division. Scholars may specify a subset of research threads deemed peripheral to be further reviewed and summarized by an AI agent taking an initial pass, and manually triage whether newly identified threads from the summary merits a closer look from them, minimizing sunk costs in cases of irrelevant or uninteresting results. Another area is exploring designs to scaffold relevance signals for user comprehension of salient threads. Here, careful designs are needed for capturing users' interests at appropriate levels of thread abstraction initially, and supporting progressive refinement for iteration and better human-AI intent communication. As more relevance signals is not always better, the potential trade-offs between benefits and information overload must also be carefully examined.

7.3 Beyond Chat-based Interfaces for LLMs

Though helpful in various use scenarios, chat-based interfaces for LLMs significantly limit scholars' synthesis workflows. Such interfaces lack support for easy extraction of useful parts in the output and its iteration through incorporating new information or supporting evidence. Despite their lack of support, these interactions were common in study participants' workflows and were also regarded as adding significant value during their sensemaking and synthesis. Our expert evaluations confirmed that literature review outlines when generated on a chat-based interface had lower overall helpfulness ratings and included significantly less supporting evidence. Limited supporting evidence also had second-order downstream implications for reviewers' confidence in the output as well as scholars' further exploration and iteration.

Here, we present an alternative design approach that incorporates LLMs as part of a larger computational pipeline for interactive interfaces, that focused their processing to recursive summarization of relevant snippets from salient research articles on a topic. This approach enabled a mixed-initiative interface design where scholars could easily integrate parts – useful threads – from generated outputs and curate supporting evidence. Summarized threads benefited users by helping them support discover and prioritize new threads and references that they could explore further. Future interface designs may benefit from further exploring the design space of incorporating LLMs as components in computational pipelines, rather than standalone chat interfaces, yielding interaction designs that significantly benefit users in discovering, prioritizing, extracting, organizing, and synthesizing knowledge during sensemaking.

7.4 Limitations

Though our evaluations uncovered new insights into scholarly synthesis workflows and implications for future mixed-initiative synthesis support tools, our experiments were limited to end-to-end evaluations of the pipeline. Additional ablation studies could tease apart contributions from each component in the pipeline (*e.g.,* the algorithms for retrieval based on novel Loopy Belief Propagation;

for formation of a thread-based hierarchy; and for recursive summarization using GPT4). In addition, evaluating against a future baseline that has an expanded prompt context (*e.g.,* using multiple papers' text as input) will contribute to whether GPT4's synthesis capabilities generalize to multiple papers. Furthermore, while our PDF acquisition and parsing was performant in the case studies where participants' personalized queries were used, scaling our approach to real-world scenarios with many users may require significant engineering resources. A notable example here is how our system aimed to acquire and parse the full text PDFs for important papers, but it relied on best effort (by involving use of commercial APIs such as Google's Custom Search; §4.1.3), without a guarantee of coverage. While significant combined research and engineering efforts such as the S2ORC corpus [31] are notable in greatly increasing access to a large paper index with full text PDFs, we note that a significant portion of human knowledge remains locked in non-accessible PDFs, and concerted legal and institutional efforts may be required to make a significant step forward in this area.

Finally, we believe that future empirical evaluations that go beyond the short duration for studies reported here, and in a more ecologically valid use context (*e.g.,* in a field deployment study) may uncover exciting new opportunities and challenges in this space.

8 CONCLUSION

In this paper we develop Synergi, a mixed-initiative system that supports scholarly synthesis and sensemaking of the scientific literature. In contrast to prior approaches that cater to either ends of the initiative spectrum (*i.e., bottom-up* or *top-down* workflows), here we develop a novel approach to help scholars iteratively review the structure of literature related to a specific query context, curate important threads and references, and outline a useful review. Our evaluation that involved 12 participants and domain experts found that Synergi allowed users to create a higher-quality outline for a literature review, compared to a baseline based on a prior system, Threddy [18] and GPT4. We also found that Synergi achieves this through efficiency gains over the Threddy baseline. Moreover, we show that Synergi increased the coverage of synthesis while also enabling effective curation of supporting evidence from multiple papers over GPT4. Participants of the user studies found Synergi to be useful in broadening their perspectives about the literature, increasing curiosity while decreasing the fear of missing out on important research in the area. Finally, we conclude with implications for future mixed-initiative workflow designs for scholarly synthesis and interesting inquiries for research in the space. We believe more work is needed in this area to uncover new mixed-initiative workflow models and to envision improved systems that can help accelerate scientific innovation for all.

ACKNOWLEDGMENTS

This work was supported by the Carnegie Mellon Center for Knowledge Acceleration, National Science Foundation (FW-HTF-RL, grant no. 1928631), the Allen Institute for Artificial Intelligence (Semantic Scholar), and the Office of Naval Research. We also thank the anonymous reviewers for their constructive feedback. In addition, we extend heartfelt thanks to our study participants, without whom this work would not have been possible.

REFERENCES

[1] 2008–2021. GROBID. https://github.com/kermitt2/grobid. swh:1:dir:dab86b296e3c3216e2241968f0d63b68e8209d3c

[2] Rediet Abebe, Nicole Immorlica, Jon Kleinberg, Brendan Lucier, and Ali Shirali. 2022. On the effect of triadic closure on network segregation. *Economic Review* 97, 3 (2022), 890–915.

[3] Yuan An, Jeannette Janssen, and Evangelos E Milios. 2004. Characterizing and mining the citation graph of the computer science literature. *Knowledge and Information Systems* 6 (2004), 664–678.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[5] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).

[6] Charles Bazerman. 1985. Physicists reading physics: Schema-laden purposes and purpose-laden schema. *Written communication* 2, 1 (1985), 3–23.

[7] Michael J. Black. 2022. *Michael J. Black on Twitter.* https://twitter.com/Michael_J_Black/status/1593133722316189696 Accessed: 2023-03-28.

[8] Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology* 66, 11 (2015), 2215–2222.

[9] Joseph Chee Chang, Nathan Hahn, and Aniket Kittur. 2020. Mesh: Scaffolding Comparison Tables for Online Decision Making. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 391–405. https://doi.org/10.1145/3379337.3415865

[10] Joseph Chee Chang, Amy X Zhang, Jonathan Bragg, Andrew Head, Kyle Lo, Doug Downey, and Daniel S Weld. 2023. CiteSee: Augmenting Citations in Scientific Papers with Persistent and Personalized Historical Context. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[11] Duen Horng Chau, Aniket Kittur, Jason I. Hong, and Christos Faloutsos. 2011. Apolo: Interactive Large Graph Sensemaking by Combining Machine Learning and Visualization. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Diego, California, USA) *(KDD '11)*. Association for Computing Machinery, New York, NY, USA, 739–742. https://doi.org/10.1145/2020408.2020524

[12] Han L Han, Junhang Yu, Raphael Bournet, Alexandre Ciorascu, Wendy E Mackay, and Michel Beaudouin-Lafon. 2022. Passages: Interacting with Text Across Documents. In *CHI Conference on Human Factors in Computing Systems*. 1–17.

[13] 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.

[14] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S Weld, and Marti A Hearst. 2021. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[15] Terje Hillesund. 2010. Digital reading spaces: How expert readers handle books, the Web and electronic paper. (2010).

[16] Amber Horvath, Brad Myers, Andrew Macvean, and Imtiaz Rahman. 2022. Using Annotations for Sensemaking About Code. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 61, 16 pages. https://doi.org/10.1145/3526113.3545667

[17] Arif E Jinha. 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing* 23, 3 (2010), 258–263.

[18] Hyeonsu Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An Interactive System for Personalized Thread-Based Exploration and Organization of Scientific Literature. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 94, 15 pages. https://doi.org/10.1145/3526113.3545660

[19] Hyeonsu B Kang, Rafal Kocielnik, Andrew Head, Jiangjiang Yang, Matt Latzke, Aniket Kittur, Daniel S Weld, Doug Downey, and Jonathan Bragg. 2022. From Who You Know to What You Read: Augmenting Scientific Recommendations with Implicit Social Networks. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 302, 23 pages. https://doi.org/10.1145/3491102.3517470

[20] Hyeonsu B Kang, Sheshera Mysore, Kevin J Huang, Haw-Shiuan Chang, Thorben Prein, Andrew McCallum, Aniket Kittur, and Elsa Olivetti. 2022. Augmenting Scientific Creativity with Retrieval across Knowledge Domains. In *Second Workshop on Bridging Human-Computer Interaction and Natural Language Processing*

at NAACL 2022. arXiv. https://doi.org/10.48550/ARXIV.2206.01328
[21] Hyeonsu B. Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. 2022. Augmenting Scientific Creativity with an Analogical Search Engine. *ACM Trans. Comput.-Hum. Interact.* (mar 2022). https://doi.org/10.1145/3530013 Just Accepted.
[22] Hyeonsu B Kang, Nouran Soliman, Matt Latzke, Joseph Chee Chang, and Jonathan Bragg. 2023. ComLittee: Literature Discovery with Personal Elected Author Committees. 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 738, 20 pages. https://doi.org/10.1145/3544548.3581371
[23] Harmanpreet Kaur, Doug Downey, Amanpreet Singh, Evie Yu-Yen Cheng, Daniel S. Weld, and Jonathan Bragg. 2022. FeedLens: Polymorphic Lenses for Personalizing Exploratory Search over Knowledge Graphs *(UIST '22)*.
[24] Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. 2023. The Semantic Scholar Open Data Platform. *arXiv preprint arXiv:2301.10140* (2023).
[25] Aniket Kittur, Andrew M. Peters, Abdigani Diriye, Trupti Telang, and Michael R. Bove. 2013. Costs and Benefits of Structured Information Foraging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2989–2998. https://doi.org/10.1145/2470654.2481415
[26] Jeffrey W Knopf. 2006. Doing a literature review. *PS: Political Science & Politics* 39, 1 (2006), 127–132.
[27] Wouter Kool and Matthew Botvinick. 2018. Mental labour. *Nature human behaviour* 2, 12 (2018), 899–908.
[28] Andrew Kuznetsov, Joseph Chee Chang, Nathan Hahn, Napol Rachatasumrit, Bradley Breneisen, Julina Coupland, and Aniket Kittur. 2022. Fuse: In-Situ Sensemaking Support in the Browser. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 34. https://doi.org/10.1145/3526113.3545693
[29] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy* 23, 1 (2020), 18.
[30] Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, Fangzhou Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Kinney, Aniket Kittur, Hyeonsu Kang, Egor Klevak, Bailey Kuehl, Michael Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita Rao, Paul Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. 2023. The Semantic Reader Project: Augmenting Scholarly Documents through AI-Powered Interactive Reading Interfaces. arXiv:2303.14334 [cs.HC]
[31] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S. Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of ACL*. https://arxiv.org/abs/1911.02782
[32] Abraham Maslow. 1965. Self-actualization and beyond. (1965).
[33] Daniel Müllner. 2013. fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software* 53 (2013), 1–18.
[34] Keisuke Okamura. 2019. Interdisciplinarity revisited: evidence for research impact and dynamism. *Palgrave Communications* 5, 1 (2019).
[35] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding Literature Reviews with Existing Related Work Sections. *arXiv preprint arXiv:2302.06754* (2023).
[36] Carole L Palmer, Lauren C Teffeau, and Carrie M Pirmann. 2009. Scholarly information practices in the online environment. *Report commissioned by OCLC Research. Published online at: www. oclc. org/programs/publications/reports/2009-02. pdf* (2009).
[37] 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.
[38] Antoine Ponsard, Francisco Escalona, and Tamara Munzner. 2016. PaperQuest: A visualization tool to support literature review. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2264–2271.
[39] Napol Rachatasumrit, Gonzalo Ramos, Jina Suh, Rachel Ng, and Christopher Meek. 2021. ForSense: Accelerating Online Research Through Sensemaking Integration and Machine Research Support. In *26th International Conference on Intelligent User Interfaces*. 608–618.
[40] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. 2018. Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 159 (jan 2018), 23 pages.
[41] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. 2012. Metro maps of science. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1122–1130.
[42] Amit Sharma and Dan Cosley. 2013. Do Social Explanations Work? Studying and Modeling the Effects of Social Explanations in Recommender Systems. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13)*. Association for Computing Machinery, New York, NY, USA, 1133–1144. https://doi.org/10.1145/2488388.2488487
[43] Benjamin Sturm and Ali Sunyaev. 2019. Design principles for systematic search systems: a holistic synthesis of a rigorous multi-cycle design science research journey. *Business & Information Systems Engineering* 61, 1 (2019), 91–111.
[44] Nicole Sultanum, Christine Murad, and Daniel Wigdor. 2020. Understanding and supporting academic literature review workflows with litsense. In *Proceedings of the International Conference on Advanced Visual Interfaces*. 1–5.
[45] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085* (2022).
[46] Jaime Teevan. 2014. A formula for academic papers: Related work. http://slowsearching.blogspot.com/2014/11/a-formula-for-academic-papers-related.html
[47] H Holden Thorp. 2023. ChatGPT is fun, but not an author. , 313–313 pages.
[48] Richard Van Noorden et al. 2015. Interdisciplinary research by the numbers. *Nature* 525, 7569 (2015), 306–307.
[49] Jen-Her Wu and Shu-Ching Wang. 2005. What drives mobile commerce?: An empirical evaluation of the revised technology acceptance model. *Information & management* 42, 5 (2005), 719–729.
[50] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. 2003. *Understanding Belief Propagation and Its Generalizations*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 239–269.
[51] Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James M O'Neill, Rudi O'Reilly Meehan, Eoin Ó Loideáin, Azzurra Pini, Medb Corcoran, Jeremiah Hayes, Diarmuid J Cahalane, Gaurav Shivhare, Luigi Castoro, Giovanni Caruso, Changhoon Oh, James McCann, Jodi Forlizzi, and John Zimmerman. 2022. How Experienced Designers of Enterprise Applications Engage AI as a Design Material. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 483, 13 pages. https://doi.org/10.1145/3491102.3517491
[52] Xiaolong Zhang, Yan Qu, C. Lee Giles, and Piyou Song. 2008. CiteSense: Supporting Sensemaking of Research Literature. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 677–680. https://doi.org/10.1145/1357054.1357161
[53] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. 2021. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence* 5, 5 (2021), 726–742.

A DETAILED SYSTEM DESCRIPTIONS

A.1 Loopy Belief Propagation Algorithm in Synergi

A.1.1 Background. The use of LBP in prior work [11] was limited to a scalar conversion weighting of the probability (`0.58`) when messages are exchanged between connected nodes in the graph. In other words, when the user assigns a category to a paper, the papers connected to that via citations would receive messages to increase their marginal probabilities of also being assigned the same category, regardless of the specific citation context. Furthermore, while this simple message weighting is a suitable configuration for interaction scenarios where the user provides iterative supervision over graph nodes (*i.e.,* user assigns a category $c \in C$ for each node n; each node state $s(n) \in \{c, \neg c, \text{not-seen}\}$), which can be used to correct subsequently propagating errors due to insensitivity to diverse citation relations, it is not suitable for our problem setting where no iterative supervision from the user can be supplied during the initial outline generation phase.

In contrast, in our problem setting the user input consists only of *the initial set of seed references* as possible exemplars on the citation graph, along with the *citation context described in natural language in which they were referred to*, without iterative supervision.

A.1.2 Running LBP with context-specific message scaling. In order to prioritize papers that globally optimizes relevance and importance to the user input, we developed a multiplicative message weighting scheme which we assign to each factor in the factor graph to change the marginal probability after each local message passing between the two papers v_i and v_j:

$$\frac{\left(\sum_{s\in S,k\in K} \text{sim}\left(\text{emb}(a_{i,j,k}), \text{emb}(c_s)\right)\right)}{|S\times K|} \times \frac{1}{1+e^{-|\text{ref}(v_i)\cap\text{ref}(v_j)|}}$$

where $\{\forall s \in S : c_s\}$ is the set of seed clips, $\{\forall k \in K : a_k\}$ are the annotation texts stored on each edge between paper variable v_i and v_j (*i.e.,* note that $k \geq 1$ because the candidate paper's title text is always available even when no citation context text was found), $\text{sim}(\cdot,\cdot)$ represents the cosine similarity function that takes two embedding vectors as its input, $\text{emb}(\cdot)$ represents a text embedding using the Open AI's `text-davinci-003` model, and $\text{ref}(\cdot)$ represents a function that takes a paper v_i as its input to return the IDs of its referenced papers.

Intuitively, the first component of the multiplication corresponds to the average semantic similarity of possible pairings between the citation contexts in seed clips provided by the user and the citation contexts of the two papers. This is relevant because we are concerned with prioritizing papers with *similarity specific to the query aspect*, rather than the entire paper's topical or thematic similarity to another paper.

The second term of the multiplication corresponds to the degree of overlapping references between the two papers. Intuitively, the higher the number of overlapping references between the two papers, the more likely they would be building on similar threads of research, which can be a useful signal. Similar mechanism of triadic closure has been shown to be capable of surfacing missing friends [2, 42], relevant paper recommendations [19], and author recommendations [22]. However, the effect of a small increase of the count of the overlapping references early on (*e.g.,* consider the effect from a step change $0 \mapsto 1$, in terms of the number of overlapping references between two papers; because there are many more papers that do not share any references, this step change may contain more discriminative information for classification than any other subsequent increases) may exhibit a steeper effect than the same difference at a higher base count of overlapping references. As such, we model the diminishing returns of this signal using the sigmoid function. Finally, the LBP is run until conversion[12].

A.2 From Binary Tree to a 3-level Hierarchy

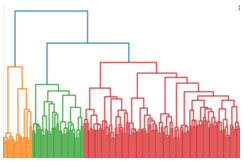

Figure 12: Example hierarchy from agglomerative clustering.

[12]We did not encounter a non-converging case in the user studies.

The resulting binary tree from the agglomerative clustering step in the algorithm (Section 4.2.2) may contain within it the high-level hierarchy that resembles the structure that emerges from bottom-up coding of clips via this clustering process. However, in practice each thread in a literature review outline may have more than just two children citation contexts supporting it. For example, in the example binary tree outputted in Fig. 12, the tri-colored branches may correspond well to three distinctive research areas and thus need to be grouped into three semantic categories. Therefore, we condense and re-structure the binary tree in a way that hides the unnecessary complexity arising from the particular clustering method, while preserving the high-level semantic groupings converted, into an 3-level N-ary tree by cutting it at 3 different heights and pruning the branches that form elongated chains.

A.3 Merging Similar Threads

After piloting the synthesized labels of threads (Section 4.2.3), we realized that the conversion of the full binary tree from agglomerative clustering into a 3-level hierarchy may have resulted in sub-groups that have similar citation contexts, that may be better described as a single larger high-level group. Therefore, we introduced a post-processing step that greedily merges parent threads that are highly similar in content from one another, thus reducing redundant sub-groups. We achieved this by using the pairwise cosine similarity of `0.92` as threshold, which was determined from pilot testing.

A.4 Chat-GPT4 Prompt for Label Synthesis

The input prompt to Chat-GPT4 consisted of a system message and a user message (Fig. 13). The outputs were generated using the OpenAI Playground interface[13] in the chat mode using the GPT-4 model. The temperature was set to `0`. The content of the user message was infilled with up to 25 citation context text snippets in each cluster.

B DETAILS OF THE STUDY

B.1 Tutorials

Before participants start with each of the two main task with different conditions, they were given a tutorial of the assigned systems via screen sharing. The interviewer demonstrated a step-by-step installation process and the main features of each system using a prepared script that took around 10 minutes in each condition. In the baseline condition, participants were instructed to clip citances using in-text highlighter directly in the PDF, and switch between the editor and PDF viewer to organize saved clips into an outline. Participants could search for the PDFs of relevant papers on the Web using any popular search engines and continuously collect relevant clips from them. Participants in the treatment condition were instructed to start by reviewing the Synergi-generated threads and recommended clips to construct an outline.

B.2 Chat-GPT4 Prompt for Literature Review

For the prompt in Fig. 14, the temperature was set to 1 for repeated random sampling. The content of the user message was infilled

[13]https://platform.openai.com/playground

```
[System Message]
You are an agent that summarizes scientific articles.
- Follow the user's requirements carefully & to the letter.
```

```
[User Message]
What is the topic commonly described in the following text snippets?
Summarize the topic succinctly (i.e., 6 words or less).
Reply with "Common topic: " followed by your response.
---
{input documents}
---
```

Figure 13: The prompt used to synthesize labels for each cluster using cluster members (`{input documents}`).

using the content of each clip used in timed tasks, augmented by the titles of the references included in the clip.

C DETAILED USER INTERACTION LOGS

A time-graph of user actions in each condition is shown in Fig. 15.

D FULL SURVEY RESULTS

Descriptions of survey items and participants' responses grouped by condition are presented in Table 1. Two-sided Wilcoxon's signed rank tests were performed to compute the p-values between conditions. See Section 6.3 for discussions of the results.

```
[System Message]
You are an assistant to a scientist who's conducting a literature review.
- Follow the user's requirements carefully & to the letter.
```

```
[User Message]
Complete the following survey paper:

Title: Using Annotations for Sensemaking about Code - A Survey

### Code comments are not commonly used for keeping track of facts learned or open questions

Code comments are commonly utilized for keeping track of open tasks [START_REF]The emergent
    structure of development tasks.[END_REF][START_REF]Work Item Tagging: Communicating
    Concerns in Collaborative Software Development.[END_REF] and can be used as navigational
    aids [START_REF]How Software Developers Use Tagging to Support Reminding and Refinding.[
    END_REF][START_REF]Work Item Tagging: Communicating Concerns in Collaborative Software
    Development.[END_REF], but are not commonly used for keeping track of the other previously
     mentioned information needs developers have such as facts learned or open questions. This
     may be partially because the cost of externalizing this information, especially when the
    information may be incorrect, is too high [START_REF]Resumption strategies for interrupted
     programming tasks.[END_REF], and these code comments must then be cleaned up [START_REF]
    TODO or to bug.[END_REF].

###
```

Figure 14: The prompt used to generate outlines for expert review (showing content for one of the two papers used in timed tasks of the experiment). (Top) The system message component of the prompt. (Bottom) The user message component of the prompt. The temperature was set to 1. The prompt for the first paper in the timed task was similarly constructed, using the clipped citation context with demarcated (*e.g.,* enclosed within each `[START_REF]...[END_REF]` pair) reference titles.

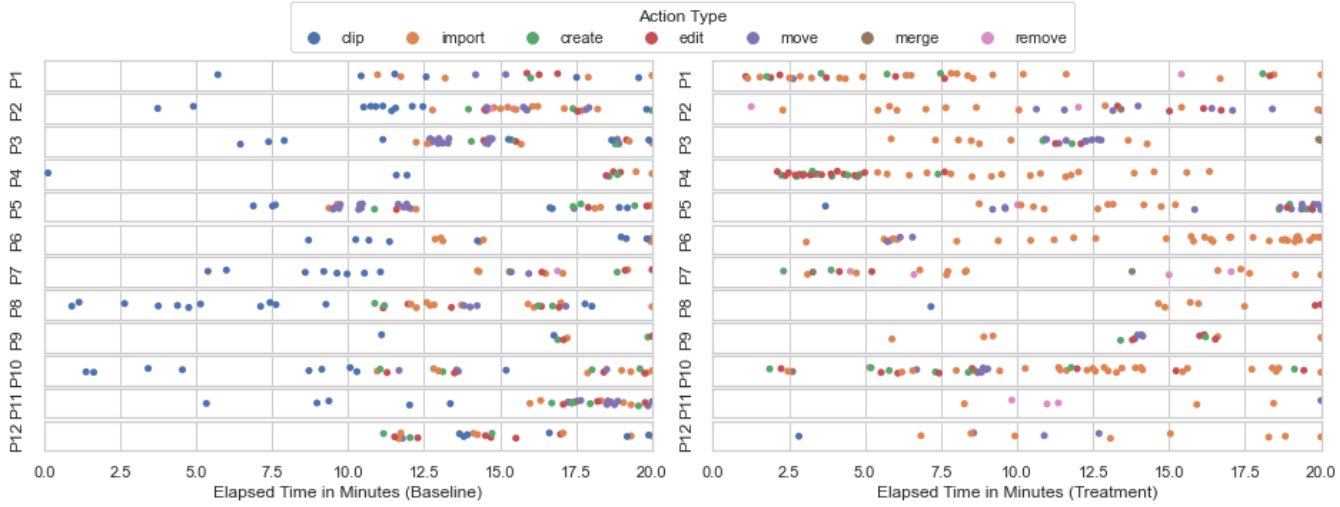

Figure 15: User interaction logs on each system showing the timestamps of seven types of actions.

	Description	Baseline	Synergi	p-val.
1. NASA-TLX	Sum of the participants' responses to the five NASA-TLX's [13] Likert-scale questionnaire items below. The original 21-point scale was mapped to a 7-point scale, similarly with [40].	22.3 (SD=6.00)	17.9 (SD=4.19)	.08
1a. Mental	"How mentally demanding was the task?"	4.8 (SD=1.36)	4.3 (SD=1.42)	.34
1b. Physical	"How physically demanding was the task?"	4.6 (SD=1.62)	3.8 (SD=1.47)	.32
1c. Temporal	"How hurried or rushed was the pace of the task?"	5.0 (SD=1.21)	3.5 (SD=1.31)	.003**
1d. Effort	"How hard did you have to work to accomplish your level of performance?"	4.4 (SD=1.44)	4.3 (SD=0.98)	.93
1e. Frustration	"How insecure, discouraged, irritated, stressed, and annoyed were you?"	3.5 (SD=2.11)	2.0 (SD=1.21)	.08
2. TAM	Sum of the participants' responses to the 4 questionnaire items below adopted from [49] measuring the technological compatibility with participants' existing scholarly discovery workflows and the easiness of learning.	19.1 (SD=4.48)	21.0 (SD=5.00)	.06
2a. Compatibility	"*Using the system is compatible with most aspects of how I search for scholars and their papers.*" (The response Likert scales for this question and below are 1: *Strongly disagree*, 7: *Strongly agree*)	4.1 (SD=1.51)	4.8 (SD=1.70)	.33
2b. Fit	"*The system fits well with the way I like to search for scholars and their papers.*"	4.7 (SD=1.83)	4.6 (SD=1.73)	.89
2c. Easy-to-Learn	"*I think learning to use the system is easy.*"	5.8 (SD=1.05)	6.2 (SD=1.02)	.48
2d. Adoption	"*Given that I had access to the system, I predict that I would use it.*"	4.5 (SD=188)	5.4 (SD=1.73)	.15
3. Discovery	"*The system helped me discover relevant threads of research in the literature.*"	3.3 (SD=2.14)	6.3 (SD=0.75)	.009**
4. Sensemaking	"*The system helped me make sense of relevant threads of research in the literature.*"	4.3 (SD=2.00)	5.3 (SD=1.66)	.09
5. Outlining	"*The system helped me outline a review of the literature.*"	5.1 (SD=2.02)	6.1 (SD=0.67)	.09
6. Curiosity	"*The system made me curious about different threads of research in the literature.*"	3.9 (SD=1.73)	6.0 (SD=0.74)	.01*
7. Confidence	"*The system increased my confidence in reviewing the literature.*"	4.0 (SD=1.71)	5.8 (SD=0.94)	.01*
8. Fear of Missing Out	"*The system reduced my fear of missing out on important research.*"	3.2 (SD=1.64)	5.2 (SD=1.22)	.01*
9. Organizing Clips	"*The system helped me organize the clips I found.*"	5.7 (SD=1.15)	5.5 (SD=1.73)	.79
10. Organizing References	"*The system helped me organize the references I found.*"	5.2 (SD=1.59)	5.8 (SD=1.66)	.34

Table 1: Descriptions of full questionnaire items and responses grouped by condition. p−values are from two-sided paired samples Wilcoxon's signed rank tests.