

QLARIFY: Bridging Scholarly Abstracts and Papers with Recursively Expandable Summaries

Raymond Fok*
 rayfok@cs.washington.edu
 University of Washington
 Seattle, WA, USA

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

Tal August
 tala@allenai.org
 Allen Institute for AI
 Seattle, WA, USA

Amy X. Zhang
 axz@cs.uw.edu
 University of Washington
 Seattle, WA, USA

Daniel S. Weld
 danw@allenai.org
 Allen Institute for AI &
 University of Washington
 Seattle, WA, USA

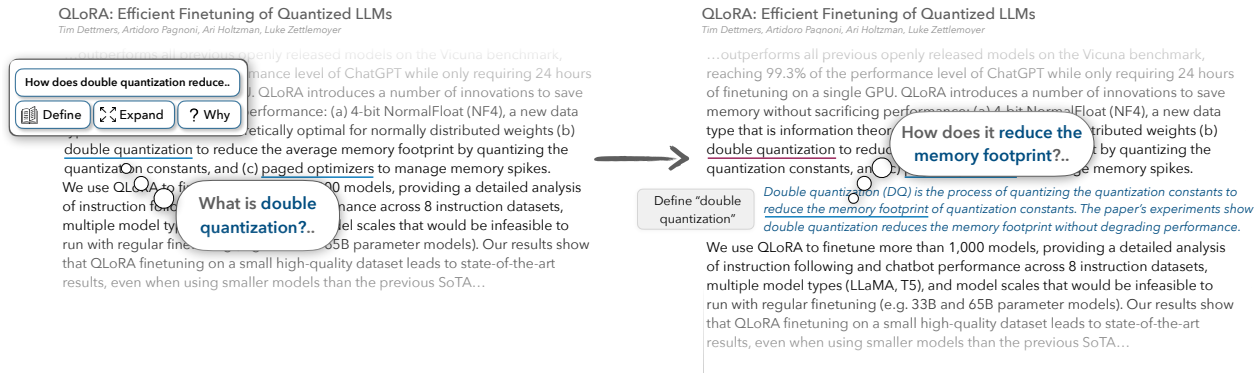


Figure 1: Paper abstracts are static author-generated summaries that cannot address the personalized information needs of every scholar. In this work, we present QLARIFY, an intelligent augmented reading interface that leverages large language models to support scholars expanding abstracts on-demand with just a few clicks, using information retrieved from full papers.

ABSTRACT

As scientific literature has grown exponentially, researchers often rely on paper triaging strategies such as browsing abstracts before deciding to delve into a paper’s full text. However, when an abstract is insufficient, researchers are required to navigate an informational chasm between 150-word abstracts and 10,000-word papers. To bridge that gap, we introduce the idea of *recursively expandable summaries* and present QLARIFY, an interactive system that allows scholars to recursively expand an abstract by progressively incorporating additional information from a paper’s full text. Starting from an abstract, users can brush over summary text to specify targeted information needs or select AI-suggested entities in the text. Responses are then generated on-demand by an LLM and appear in the form of a fluid, threaded expansion of the existing text. Each generated summary can be efficiently verified through attribution to a relevant source-passage in the paper. Through an interview study (n=9) and a field deployment (n=275) at a research conference, we use QLARIFY as a technology probe to elaborate upon the expandable summaries design space, highlight how scholars benefit from QLARIFY’s expandable abstracts, and identify future opportunities to support low-effort and just-in-time exploration of

scientific documents – and other information spaces – through LLM-powered interactions.

CCS CONCEPTS

- **Human-centered computing** → **Interactive systems and tools.**

KEYWORDS

interactive documents, expandable summaries, scientific papers, mixed-initiative systems

1 INTRODUCTION

Keeping current with the literature is an important part of many scholarly workflows. Literature review helps scholars survey what has already been done, identify open challenges, and find inspiration [36]. Unfortunately, the process of reviewing the literature has grown ever more challenging due to the exponential growth of scientific publication [6, 7]. In response, computational interventions have sought to aid scholars in the discovery and consumption of research literature. Systems supporting literature discovery such

*Work completed during an internship at Semantic Scholar, Allen Institute for AI.

as Google Scholar¹ and Semantic Scholar² allow scholars to search for papers, navigate citation graphs, and receive recommended papers based on their interests. In contrast, systems supporting literature consumption assist in making sense of individual papers, for instance by improving comprehension [2, 26], supporting skimming [19, 41], and synthesizing insights across papers [31, 34].

At the seams of literature discovery and consumption there exists another challenge which has received less attention — the preliminary, breadth-first exploration of papers. Scholars often need to browse collections of potentially related papers, such as a recurring email digest of recently published papers or proceedings of a conference. The exploratory triaging process allows scholars to assess papers for their relevance, identify future reading material, or satisfy an informational curiosity, without incurring the cognitive costs of reading each paper. A typical approach might involve reasoning over each paper’s abstract to determine its relevance and legitimacy [28]. Literature discovery tools can provide scholars with possible papers to read, and literature consumption tools can assist scholars in reading selected papers; however, neither set of tools assists in a preliminary exploration of papers through abstracts.

While abstracts can be appealing for triage since they succinctly encapsulate essential aspects of a paper, they come with significant limitations. As static summaries that condense upwards of 10,000 words into one or two paragraphs, abstracts rarely address all of a scholar’s information needs and require clarification to fully comprehend. For instance, an abstract could mention the size of a study, but give no further information a scholar may want about recruitment, participant demographics, or analysis techniques. This challenge of locating additional context is further complicated when scholars need to triage numerous papers. A scholar interested in understanding how prior work built on a specific dataset would need to manually open the PDFs for each search result, and search each paper for mentions of the dataset name to find passages that contained detailed descriptions about how the dataset was used.

We investigate how paper triage may require navigating the knowledge gaps between an abstract and a paper in a formative study with seven scholars. Our findings revealed four common information needs for which scholars may seek additional context while reading an abstract: *Definition*, *Instantiation*, *Clarification*, and *Motivation*. These needs were often expressed as questions about the abstract that could be answered with information from the paper’s full text.

In this work, we take inspiration from these observations and propose a novel interaction technique, *recursive summary expansion*, that supports users in interactively asking questions to expand abstracts with information retrieved from the paper’s full text. We reify this interaction within QLARIFY, an augmented reading interface for scientific paper abstracts powered by current Large Language Models (LLMs). In contrast to LLM-enabled chat-based interactions with papers (e.g., ChatPDF [12]) which require users to devise questions from scratch, QLARIFY envisions how LLMs can instead enable an effortless, threaded interaction paradigm where questions naturally arise from interacting with an abstract.

Specifically, with QLARIFY, users can query any part of an abstract and view in-situ answers drawn from the full paper that incrementally expand the abstract. QLARIFY bootstraps the question-asking process by underlining AI-suggested candidate entities within an abstract, indicating areas that are potentially informative to expand (Figure 1). In addition to selecting these AI-suggested expandable entities, users can highlight arbitrary text in the abstract to expand. Users can also recursively expand any expansion, probing further into details of the paper via the threaded expansion. QLARIFY reduces the cost of asking an appropriate question to a single click by providing two types of question scaffolding: (1) an AI-suggested question which aims to infer a user’s information-seeking intent, and (2) three static questions — *Define*, *Expand*, and *Why* — which mirror the common information needs scholars may desire in an expansion. QLARIFY enables verification of any expansion by providing attribution to relevant passages from the paper and one-click access to the passage within the paper.

We use QLARIFY to investigate how scholars would use an expandable abstract interaction to support their exploratory paper triaging workflows. Though an interview study (n=8) and online deployment (n=275) with scholars, we validate the design and efficacy of QLARIFY’s recursive expansions. We found scholars often used the AI-suggested expandable entities to facilitate a guided active reading of abstracts during triage, creating threaded expansions that thoroughly satisfied their informational curiosities. Scholars were also largely impressed with the ability for the LLM-suggested questions to infer their information-seeking intents and the LLM-generated expansions to answer the questions they asked. We conclude by discussing further opportunities for a recursive expansion interaction to bridge abstracts and papers and implications of our findings on future AI-infused scholarly support tools.

In summary, we contribute the following:

- A novel document-centered interaction technique, *recursive summary expansion*, that allows users to progressively expand abstracts on-demand with information from the paper’s full text, and a discussion of its design space.
- QLARIFY, an augmented reading interface for recursively expanding scientific paper abstracts, leveraging LLMs to reduce the costs of formulating and answering questions, as well as providing easy verification of answers.
- Insights from qualitative interviews (n=9) and an online deployment (n=275) with scholars, revealing how an expansion interaction can enable rapid yet thorough exploratory information-seeking processes over abstracts.

2 RELATED WORK

2.1 Addressing Document-Centered Information Needs with Summarization

The aim of summarization is to condense long documents into short and concise texts, encapsulating the most important information required for comprehension. With the advent of neural architectures [38], significant efforts have been dedicated to improving the capabilities of automatic text summarization systems. Some work has focused especially on summarizing domain-specific and long-form documents, such as scientific text [8, 27, 72]. Consuming a traditional summary requires little to no user effort, but the

¹<https://scholar.google.com/>

²<https://www.semanticscholar.org/>

static medium of a summary makes it impossible to capture the personalized and nuanced information needs of every individual. Some research has explored ways to incorporate humans in the loop to generate more personalized summaries [22, 62, 76], but once generated, these summaries lack the ability for iterative refinement, for instance to reflect updated information needs.

Prior work has explored hierarchical approaches to summarization that enable a reader to interactively specify the degree of depth they wish to explore. These approaches require the authoring of summaries at differing levels of depth (e.g., for summarizing large-scale online discussions [74], books [71], or web documents [4, 52]), forming a static summary tree artifact that enables structured navigation between the summaries and long documents. These summaries are typically constructed in a bottom-up fashion, working recursively starting from the full text, so as to break down the work into more manageable chunks. In contrast, we take a top-down approach; rather than condensing information to *generate summaries*, we instead leverage automated techniques to *expand summaries* – incrementally and recursively – to reveal relevant information from a long document on-demand. By constructing expandable summaries in the same direction that readers explore (i.e., by drilling down), the summary trees generated by our system are personalized to each reader’s exploration, as opposed to having to conform to a single rigid structure.

This concept of expanding text to incorporate more context and resolve ambiguities is also related to *decontextualization*, a recent NLP task exploring automated approaches for rewriting extractive excerpts divorced from a longer document to be interpretable out of context while preserving meaning [14, 53]. The expansion interaction we explore can be seen as a form of interactive decontextualization, sensitive to users’ personal information needs.

2.2 Querying Documents with Question Answering Systems

People often express their information needs within documents as natural language questions [29, 37, 67], and ask follow-up questions when an initial answer is not sufficient [47]. Question answering is a long-standing problem in NLP, including considerable efforts for answering questions over scientific documents [16, 60, 61]. Recent advances in LLMs have seen the proliferation of prototypes for conversational question answering over long documents (e.g., ChatDoc [11], ChatPDF [12]). The expansion interaction we propose in this work is related to these systems, but differs in two main ways: first, we aim to integrate answers in-situ by expanding the summary at the point where a question was asked to maintain the flow of reading, and second, we introduce mixed-initiative interactions that reduce the cost of forming and asking questions. To enable these question-driven expansions, we implement a retriever-reader architecture with a dense representation-based retriever and a generative, LLM-powered reader.

Attributed question answering [5], where AI models are tasked with returning evidence in addition to their answer, has gained recent interest due to the potential for hallucinations in LLM-generated answers [30, 46]. Some work has sought to more closely integrate attribution with generation (e.g., with post-editing [21]), while systems such as GopherCite [48], WebGPT [51], and LaMDA [68]

place the burden of fact verification on the user by providing URLs and excerpts as supporting evidence. We take a similar user-centric approach to attribution, providing simple affordances for users to drill down into excerpts and then the full paper from an LLM-generated response.

Prior conversational interfaces with documents often make two assumptions: that users know what to ask, and that the most natural interaction is for users to manually type their intended questions. Yet this is not always the case [2]. Instead, we propose that carefully crafted interactions and language understanding techniques can effectively infer the intent of users, reducing the cost of asking questions to a single click. In this work, we explore the potential for intelligent support to not only answer users’ questions, but also suggest questions prompting relevant entities to expand. Recent work has begun to explore the effectiveness of LLMs in this task of question generation, in particular the generation of follow-up and clarification questions [37, 39, 47]. As prompting LLMs can be challenging [49, 73], QLARIFY obviates the need for users to craft their own prompts by providing a set of common questions to select from. These questions are grounded in a taxonomy of document-centered information-seeking questions people ask while reading a document [37], and refined for the context of scientific documents through a formative study.

2.3 Supporting Exploration of Scholarly Literature

Scholars need to triage more papers in less time now more than ever [40], facing constant information overload from the growing number of publications [15, 20], shift from paper to digital publishing [42, 66], and distribution of ongoing work via online preprint archives (e.g., arXiv³). Scholars conducting exploratory research over a collection of papers often exhibit opportunistic and dynamic information needs [63], and managing papers across historical collections and continuous publication streams can be challenging [50, 65]. To support scholars in triaging, organizing, and reviewing papers, a plethora of systems have been introduced within both academic research and industry contexts.

One line of work aims to support scholars in exploring a collection of papers. For instance, many scholars nowadays use systems such as paper search engines (e.g., Google Scholar [3] and Semantic Scholar [1]), graph-based visual tools (e.g., ConnectedPapers [56]), and information extraction tools (e.g., Elicit [18]) to assist in discovering relevant work or scaffolding a literature review. Prior work has also developed research prototypes that facilitate visual exploration of the research landscape [13, 25, 45, 57, 65, 70], scaffold paper exploration through augmentations of related work sections [54] and citations [10, 58], leverage personalized cues for paper recommendations [32] and literature discovery [33], and synthesize research threads across papers [31, 34].

Another set of interactive systems aims to support scholars in reading and understanding individual papers. For instance, PaperPlain helps lay readers navigate medical literature with AI-generated section summaries and suggested questions [2], and ScholarPhi helps scholars understand paper-specific jargon with definition and equation augmentations [26]. ScholarPhi allows

³<https://arxiv.org/>

Table 1: Participants in our formative study and some questions they asked while reading abstracts of scientific papers.

	Title (Research Area)	Representative Questions
P1	PhD student (Explainable AI)	What does “TAM” mean? What is a “path analysis”? What does “visual question answering task” mean? What’s an example of a “trustworthiness cue”?
P2	PhD student (Computational Biology)	What is the “two-stage algorithm”? What are the 12 challenging reasoning tasks? What’s an example of this task? What is “the zero hypothesis”? What does “outlying failure cases” mean?
P3	PhD student (Human-AI Interaction)	What does “comparable to SoTA phrase based systems” mean? What is a “sequence transduction model”? What is the “path-x challenge”? What is the “path-256 task”? What does “retrieval set” mean?
P4	PhD student (NLP)	What loss function did they use? What’s the model architecture? What’s the metrics they use? Why do they call it “human-interpretable”?
P5	Post-doctoral scholar (HCI)	How do they define “trust” in human-AI teamwork; is this reliance? Is “human-agent teamwork” a defined sub-field of prior work, or is that just jargon the authors use? What are examples of “spatial crowdsourcing”? What are examples for the “two realistic task assignment settings”?
P6	PhD student (HCI)	What do the authors mean by “gigification of knowledge work”? What were the design recommendations? What are the key findings in a simplified sense? What papers are they building upon? What is their study design?
P7	PhD student (HCI)	What does “perceived valence” mean? What does “participatory foresight” mean? What do they mean by “early testing of AI-based features”? What does a report look like when they say “practitioners with reports”?

scholars to click on specific terms and view definitions in-situ within a floating card. We adopt a similar interaction design in QLARIFY, allowing users to click on suggested entities underlined within an abstract to expand with minimal effort (i.e., a single click). Other systems draw on visual aids to improve comprehension, for instance by highlighting conceptual relationships within papers using bubble-tree map visualizations [75], embedding animated figures into papers [24], and linking video summaries from authors’ talk videos with relevant passages in a paper [35]. Complementing these tools that support paper comprehension, some systems facilitate rapidly reading or skimming papers. For instance, Spotlights anchors visually salient objects as transparent overlays on a paper to facilitate high-speed skimming [41], and Scim uses faceted highlights to direct readers’ visual attention through a paper [19]. Our work presents an interaction technique situated at the intersection of supporting paper exploration and comprehension. Specifically, we seek to narrow the informational gap between a paper’s abstract and full text, addressing scholars’ personalized information needs as they arise during the triaging of abstracts.

3 FORMATIVE STUDY

To understand the types of information needs that emerge when exploring scientific paper abstracts, we first conducted a formative study observing scholars reading abstracts in a familiar domain. We recruited seven participants from an academic institution via group messaging channels and snowball sampling (Table 1). All participants actively conducted research across different areas of computer science, and reported familiarity with the research process, including triaging, reading, and organizing scientific papers. After introducing the study and obtaining consent, participants were asked to read 3–5 abstracts of their choice. Participants selected abstracts from various sources: many used results from a paper search engine seeded with a relevant prompt for their area

of research, one used a abstracts for papers they were currently reviewing, and one used papers they had previously saved for a later reading session. While reading each abstract, participants were asked to share aloud any thoughts, questions, or confusions they had about the information within the abstract or about the paper in general. All studies lasted approximately 45 minutes.

3.1 Findings

Our observations of scholars revealed four common types of information needs expressed while reading an abstract:

Definition. Participants wanted to define jargon such as unfamiliar terms, abstract language, or acronyms they encountered in the abstract. These definitions could be either contextual (i.e., grounded in information or explicitly defined in the paper) or context-free (i.e., pertaining to information not defined in the paper).

Instantiation. Participants sought examples to instantiate and provide concrete context for under-specified language. For instance, given the sentence, “We find our approach outperforms three baselines on a common question answering benchmark,” it is unclear which baselines were compared against or which question answering benchmark was used. When an abstract described an unfamiliar concept, such as a newly-introduced task or dataset, participants wanted to view an instance of the concept to help visualize its structure and compare it against familiar concepts such as existing tasks or datasets.

Clarification. Participants sought additional context to help explain technical or unfamiliar language in an abstract. Since abstracts are concise, self-contained summaries for a long paper, authors are compelled to withhold particular details and use dense language to convey information. As a result, scholars reading abstracts often have information needs expressed through implicit clarification questions, personalized by their own expertise and reading goals.

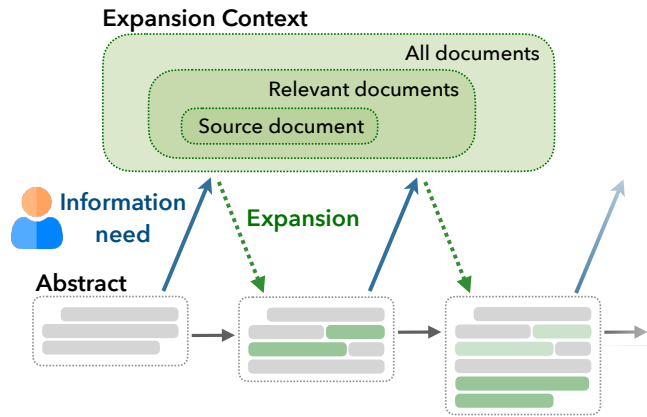


Figure 2: We describe a novel document-centered interaction technique, *abstract expansion*, for progressively expanding abstracts with clarifying information from a broader *expansion context* in response to users’ dynamic information needs.

Motivation. Participants occasionally expressed a desire to probe the authors’ motivations and justifications for various aspects of the described work. For instance, some participants questioned why the authors chose their specific approach (e.g., model architecture, loss function, or task), why particular language was used in describing their system (e.g., “human-interpretable”), or why the problem they tackled is important and different from prior work. Addressing this need may help scholars to satisfy curiosities, expand their understanding, or ascertain a paper’s validity.

3.2 Abstract Expansion and Design Space

Motivated by the information-seeking behaviors scholars exhibited when reading an abstract, we propose a novel document-centered interaction technique, *abstract expansion*, that augments abstracts with additional relevant information in response to dynamic user queries (Figure 2). This expandable text paradigm is inspired by StretchText (or transclusion) [4, 52], an early vision in Project Xanadu and hypertext design from the 1980s for structuring text on the web that allows users to choose the level of detail they want to see. When a specific area or keyword is selected, the originally concise text “stretches” to reveal additional details. While the original vision for StretchText requires carefully-authored, structured text and has not gained wide adoption, we revisit and build on this vision by using LLMs to dynamically generate on-demand summaries that support personalized and interactive expandable text.

For this, we focused on designing expandable summaries based on scholarly abstracts and papers. Abstracts are concise, static, author-crafted text summaries for a long scientific document; as such, no single abstract can concisely capture the interests of every reader or address dynamic information needs that arise while reading [76]. *Abstract expansion* ameliorates the static limitations of summaries by allowing scholars to interactively expand abstracts with additional clarifying information retrieved from an *expansion context*. For instance, the *expansion context* for most clarifying queries over an abstract is often the corresponding paper’s full text,

Dimension	Alternatives				
Information needs type	Agnostic	Grounded	Latent		
Information needs source	User-suggested	AI-suggested	Mixed-initiative		
Expansion context	Same doc	Related docs	Open-domain		
Expansion length	Short phrase	One sentence	Several sentences		
Expansion placement	Fluid	Inline	Appended	Popup	Sidebar
Expansion delineation	Bold	Italicize	Colorize	Indent	Quote
Attribution method	Embedded			Separate	
Attribution granularity	Phrase	Sentence	Entire expansion		
Attribution length	Phrase	Sentence	Paragraph	Page	

Figure 3: A design space for interactive systems that implement the *abstract expansion* interaction paradigm. Alternatives we explored through the iterative design of QLARIFY are highlighted in gray, and those included in the final system are outlined in red.

but expansions could also be drawn from information in a broader domain, such as related papers in the paper’s citation network or general information in an online resource (e.g., Wikipedia).

To help articulate how interactive systems could realize such a paradigm, we describe several dimensions and alternatives of a design space for an abstract expansion interaction (Figure 3).

3.2.1 Information needs. One aspect of the design space concerns what and how information is selected for expansion. First, what *type* of information needs should be expanded? The four types of needs identified in the formative study are similar in that they all represent an information need *grounded* in language from the abstract. These may emerge as scholars read an abstract, and are anchored to specific words in the abstract. For instance, in the following excerpt, “We evaluated against three baseline approaches on a popular question-answering benchmark,” a grounded information need might express, “What were the three baseline approaches?” or “What was the popular question-answering benchmark used?”

Information needs may also not be grounded in an abstract’s text. Scholars could read abstracts with a specific set of questions in mind, with needs *agnostic* to any particular abstract and instead relevant to their overall goals of exploration. For instance, an NLP researcher browsing a group of similarly structured papers may want to generally expand each abstract to learn more about the specific methods, experiments, or findings.

Finally, information needs could be *latent*, unknown to a scholar until explicitly surfaced. Information salient to a scholar’s goals could be located within the paper and yet be absent in the abstract. Such information may be valuable to expand but is challenging to directly query for, as it represents an unknown need.

Another dimension to consider is the *source*, or initiating agent, of the information needs to be expanded. For instance, expansions could be created in response to a user’s information-seeking question, to an AI’s suggested expandable entity, or perhaps a combination of both (i.e., mixed-initiative).

3.2.2 Expansion. Another aspect of the design space considers the content and visualization of information within an expansion. First, what is the desired expansion *context*? For instance, information used to form an expansion could be retrieved from the full text of the paper for an abstract, from other relevant papers, or from a broader corpora of information (e.g., Wikipedia).

Then, what expansion *length* is appropriate? An expansion should comprehensively address scholars' questions, yet be judicious in length to reduce cognitive burden. It could contain just a concise sentence fragment as an answer, some reasonable number of sentences, or a longer, more contextualized answer with details that could motivate further exploration.

Furthermore, what is the appropriate *placement* for the expansion in relation to the original text in the abstract? Expansions could be presented in an adjacent pane (similar to many chat-based applications with documents), in a popup card (similar to citation cards in augmented paper reading interfaces⁴ or page previews in Wikipedia⁵), appended at the end of a summary, or placed in-situ near an appropriate text anchor in the abstract. Selecting the optimal placement requires trading off the navigational effort between an expansion and the abstract, and the potential for visual distraction or clutter.

Finally, for placements that interweave expansions and the abstract, how should the two sources of text be visually *delineated*? This consideration is particularly important since the provenance of the original abstract is known and trusted, while expansions present additional text which could deviate from the underlying paper in unexpected ways. For instance, the expansion text could be indicated with standard visual cues, e.g., bold, italics, color, or through positional displacement, helping users to visually identify and switch between the two texts.

3.2.3 Attribution. A third aspect of the design space considers how information provenance may be conveyed for the generated expansions. For question-answering, provenance is often achieved through attribution, i.e., retrieving evidence from the expansion context to support a generated answer. One consideration is the *method* of conveying attribution. For instance, evidence could be embedded directly in an expansion through visual cues such as highlights or quotation marks to demarcate abstractive and extractive text. Alternatively, a system could provide attribution as separate excerpts, surfaced verbatim from the expansion context.

For longer or more complex expansions, multiple pieces of evidence may be necessary to support each claim in the generated text. In these cases, systems should consider the most appropriate *granularity* of attribution to determine the utility of providing evidence for each phrase, sentence, or entire expansion. Finally, considering the appropriate *length* of attributed evidence can assure sufficient evidence is provided to support verification of the generated expansion without introducing excessive cognitive burden.

4 THE QLARIFY SYSTEM

Next, we describe the design and implementation of QLARIFY, an augmented abstract reading experience that realizes the abstract expansion paradigm. QLARIFY's interactive features enable scholars

to recursively expand abstracts on-demand, progressively incorporating information from the full paper relevant to their triaging goals and dynamic information-seeking curiosities.

4.1 User Interface

The design of QLARIFY was motivated by the four types of information needs observed in our formative study and further refined through an iterative design process. In this process, alternatives in the design space were considered and evaluated by scholars both internal and external to the research team.

4.1.1 Eliciting Information Needs as Clarifying Questions. With QLARIFY, users begin by reading an abstract as they typically would. As information needs arise during reading, users can highlight any span of text within the abstract to request additional information. In response to a user's highlight, QLARIFY displays a *question palette* centered above the highlighted text (Figure 4B) that enables users to easily specify their information needs as clarifying questions anchored to a specific context in the abstract (i.e., the highlighted text). The question palette contains four buttons: three static buttons with the questions *Define*, *Expand*, and *Why*, and one dynamic button with an AI-suggested question.

The AI-suggested question aims to predict a user's intent, offering the most probable clarification question a user may want to ask given the text they highlighted. The three static questions are the same in every question palette regardless of the selected text, and were selected to reflect the common grounded information needs revealed in our formative study. Specifically, the *Define* question aims to address *Definition* and *Instantiation* needs, the *Expand* question aims to address *Clarification* and *Instantiation* needs, and the *Why* question aims to address *Motivation* needs. The *Expand* question is centralized, as we believe it could serve as a "catch-all" option to incrementally retrieve more details in most circumstances.

While increasing the number of static questions could provide more flexibility, based on feedback to initial prototypes of QLARIFY we determined that providing more questions could clutter the interface, occlude more of the abstract, and cause decision paralysis in selecting an appropriate question. For similar reasons, only the top-1 AI-suggested question is shown in the question palette.

To complement the manual highlighting of text users want to expand, QLARIFY also pre-selects several *expandable entities*. These entities capture spans of text in the abstract that the system believes could be further expanded from the expansion context; for instance, they could include under-specified language (e.g., "some", "several", "various") or jargon (e.g., acronyms). Expandable entities are visually indicated with a blue underline (Figure 4A), and users can click on an entity to reveal the question palette. Altogether, QLARIFY aims to reduce the cost of asking grounded information-seeking questions through these two lightweight interactions.

4.1.2 Expanding Abstracts with Clarifying Information. When users select a question from the question palette, QLARIFY creates an *expansion* by fluidly expanding the abstract with in-situ information retrieved from a larger expansion context (Figure 4C). Each expansion is an abstractive, LLM-generated response to a user's question, containing up to three sentences. In instances where a question cannot be answered, no expansion is created, and a toast alert is shown

⁴<https://www.semanticscholar.org/product/semantic-reader>

⁵https://www.mediawiki.org/wiki/Page_Previews

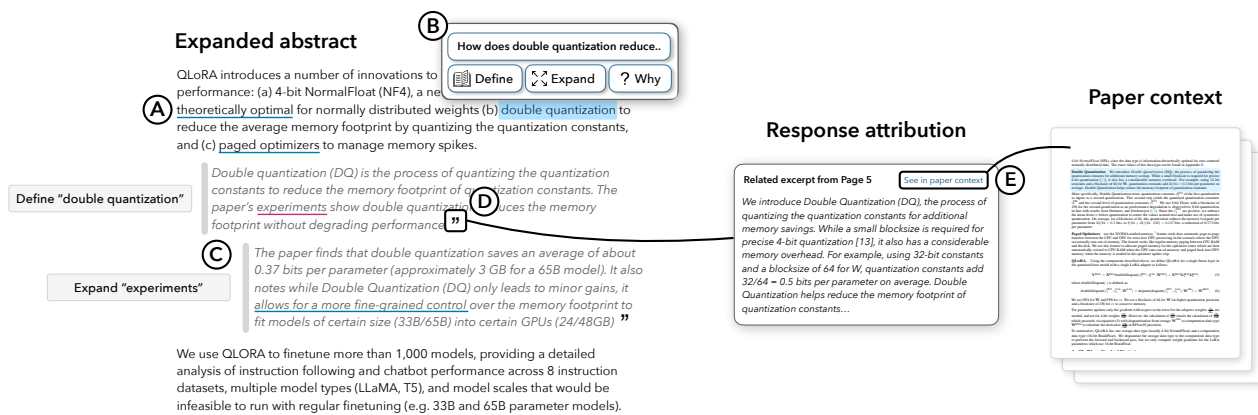


Figure 4: Expandable scientific paper abstracts in QLARIFY, instantiating a general *summary expansion* interaction technique. Users express information needs by highlighting text in the abstract or simply selecting an AI-suggested expandable entity (A), revealing a question palette (B). Clicking on a question in the palette scaffolds a question-asking prompt for an LLM, which provides a response visualized as a fluid, threaded expansion within the abstract (C). Users can drill-down to see evidence for a response in a paper excerpt (D) and within the full paper context itself (E). The excerpts above are from *QLORA: Efficient Finetuning of Quantized LLMs* [17].

in the bottom right of the screen instead to indicate an expansion could not be created.

Key to the expansion’s design is ensuring visual delineation between the original text of the abstract and that of the generated expansion. Expansions are presented as indented blocks, appended below the sentence containing the selected expandable entity. A vertical bar and tag containing the question used to generate the expansion are shown to the left of each expansion to help users easily identify the visual boundary and purpose of each expansion. The text of an expansion is also colored blue when initially created, gradually transitioning into a light gray after a few seconds. This produces a smooth animation that visually cues attention to a new expansion, before fading into a color (gray) that is similar but not overly distracting from the color of the original abstract text (black). To indicate parts of the abstract previously expanded, text used to seed the expansion – either from a user highlight or an expandable entity – is underlined in purple.

For each generated expansion, QLARIFY suggests additional entities that could be expanded. Similar to interactions with the original abstract text, users can select an AI-suggested entity or highlight any text in the new expansion to recursively expand further. Expansions created on other expansions form a threaded abstract reading experience, allowing users to easily dive deeper into aspects of interest from the abstract by gradually retrieving details from the full paper. As each expansion is limited to three sentences in length, users are less likely to be overwhelmed by information in any single expansion. They can continue expanding until their information needs are satisfied, at which point they can easily pop back up into the original abstract or expansions at any level. If a particular expansion is no longer needed, users can click on the question tag at the left of the expansion to collapse it into its parent expansion.

4.1.3 Drilling Into a Paper with Attributed Responses. Since each expansion is generated automatically by an LLM, there is a potential

risk of generating content that is unfaithful to the original paper or otherwise factually incorrect, a problem referred to as hallucination [30, 46]. To help mitigate these risks, QLARIFY provides *attribution*, or extractive supporting evidence, for each expansion. Users can click on a quote button at the end of each expansion to show a card with the most relevant paragraph from the full paper (Figure 4D). Within the card, users can further drill-down to open the paper in an integrated document viewer overlay, with the attributed paragraph navigated to and highlighted in the context of the paper (Figure 4E). Through these two levels of interactive attribution, QLARIFY allows users to explore the surrounding paper context and verify the accuracy of an expansion.

4.2 System Architecture

To create an expandable abstract, QLARIFY implements three natural language services (illustrated in Figure 5): (1) A *document preprocessing* service extracts and encodes information for a paper’s full text; (2) A *question answering* service generates attributed answers to users’ document-centered questions; and (3) An *expandable entity extraction* service identifies expansion candidates within an abstract or generated expansion. We provide an overview of our implementation of these services, which may serve as a starting reference for others exploring similar expandable summary interactions.

4.2.1 Document Preprocessing. Papers ingested by QLARIFY are first preprocessed to reduce latency when interactively generating expansions at query-time. Each PDF is parsed into its constituent tokens and bounding boxes, and sentences and paragraphs are constructed from the full text. Then, chunks are created with a chunk size of three sentences and a two sentence overlap. Each chunk is converted to a dense vector representation with encoder

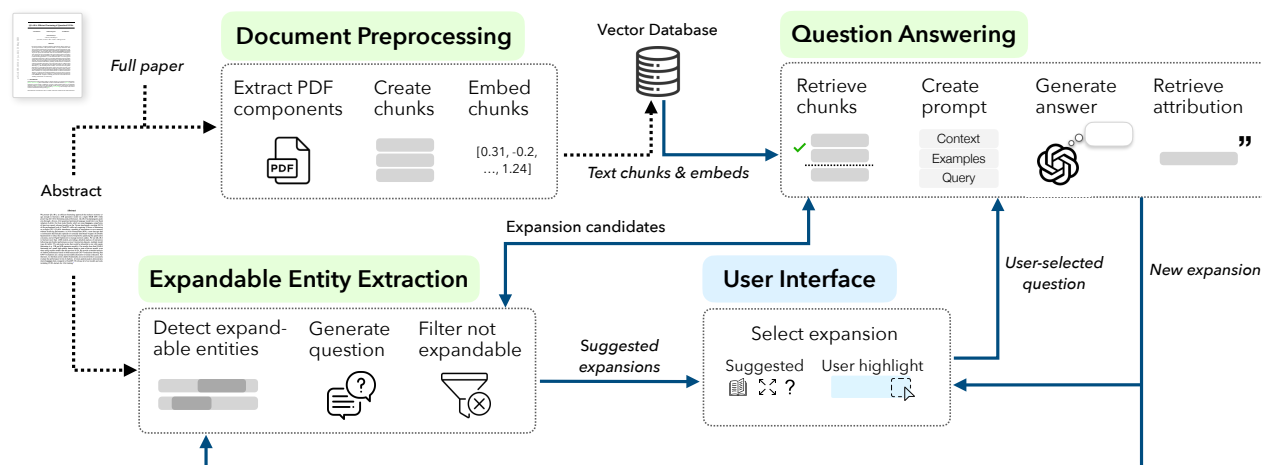


Figure 5: Overview of QLARIFY’s system architecture. Each paper is first preprocessed (*Document Preprocessing*) and initial expandable entities are extracted from the abstract (*Expandable Entity Extraction*). When a user selects a question for an expandable entity, QLARIFY uses a retriever-reader architecture with LLMs to generate a response with attribution (*Question Answering*). QLARIFY then suggests expandable entities for the new expansion, and users can iteratively expand upon expansions, creating a threaded abstract.

model `all-mpnet-base-v2` from the SentenceTransformers framework [59], and stored into a vector database. Embeddings of paragraphs are also created and stored in a separate index, which are used for retrieving attribution.

4.2.2 Attributed Question Answering. We use a common retriever-reader architecture with LLMs for question answering. When users select a question from the question palette, QLARIFY first embeds the question with the same encoder used during preprocessing. It then retrieves the 12 most relevant paper chunks (with relevance determined by cosine similarity between chunk and question embeddings) to form a context. An LLM prompt is then formed by concatenating a general description of the question answering task, the context, a few examples of question and answer pairs, and the question. The prompt further instructs the LLM to be concise, use language from the provided paper context when appropriate, generate answers containing no more than three sentences, and return no answer if the question cannot be answered given the context. These specific prompt tuning adjustments were made over several iterations of inspecting QLARIFY’s expansions. Finally, we use `gpt-3.5-turbo` to generate an answer for the question using this few-shot prompt. The current QLARIFY prototype answers questions using information from the full text of the source paper only; we leave consideration of other possible expansion contexts (e.g., other related papers) for future exploration.

QLARIFY further provides attribution for each of its expansions to enable users to verify the accuracy of the generated answer and ease into the full paper. To generate attributions, QLARIFY retrieves the most relevant paragraph to the generated text (by cosine similarity). We explored other attribution schemes in earlier iterations of QLARIFY. For instance, we tried retrieving chunks for each individual sentence, but found chunks were less preferred than paragraphs since they sometimes lacked sufficient context. We

also tried providing attribution for each sentence where expansions consisted of multiple sentences. However, we found the need to read and reconcile multiple attribution challenging sources introduced confusion and made verification more challenging.

4.2.3 Expandable Entity Extraction. To complement users in manually specifying their own expansions, QLARIFY proactively suggests parts of an abstract or expansion that could benefit from additional context. To identify these regions within an abstract, QLARIFY uses `gpt-4` with a few-shot prompting strategy. The model is instructed to identify short text spans (i.e., entities) which may be expanded to provide clarification for vague, dense, or jargon-rich language. The prompt also specifies that information required to expand each entity should not be already available in the abstract or expansion. For each entity, QLARIFY performs a dry-run expansion (using the same *Expand* question in the question palette), and removes entities for which no answer is found.

For each remaining entity, `gpt-4` is directed with zero-shot prompt to generate a single question that users might ask in expanding that entity. For instance, given the sentence, “We propose a new framework to address the ACTA task,” QLARIFY could identify the entities “a new framework” and “ACTA,” and generate the questions “What are the main characteristics of the proposed framework?” and “What is the ACTA task?,” respectively. All of these *expandable entities* are underlined in reading interface and the suggested question is shown in the question palette. The same question generation prompt is executed on-the-fly to generate the suggested question within the question palette when users create an expansion by highlighting any text.

4.3 Implementation Details

QLARIFY was implemented as a standalone web application using TypeScript, CSS, and the React framework [64] for the user interface.

The PDF reader for viewing expansion attribution in context was adapted from an open-source PDF reader library [43].⁶ Backend services and LLM-powered functions were implemented in Python and the Flask framework [55]. GROBID [23] was used to parse PDFs paper into a structured JSON format, and the MMDA open-source library⁷ was used to construct Document objects from the output from which full text paragraphs and sentences could be retrieved. The gpt-3.5-turbo (with a 4,097 token context window)⁸ and gpt-4 (with a 8,192 token context window) LLMs were accessed via OpenAI’s APIs,⁹ and responses were generated with a temperature of 0 and a maximum length of 256 tokens. The specific LLM prompts we used for each service is provided in Appendix C.

5 EVALUATION

We used QLARIFY as a technology probe to understand how scholars would benefit from an expandable abstract interaction in exploring scientific papers. We first conducted an interview study to elicit nuanced qualitative insights into scholars’ perceived benefits and limitations of QLARIFY, and then conducted a deployment study to characterize real-world usage behaviors with expandable abstracts.

5.1 Interview Study

5.1.1 Procedure. We conducted an interview study to better understand the costs and benefits of using QLARIFY to triage scientific papers. To ensure engagement with the study, we curated a personalized set of abstracts for each participant aligning with their research expertise and interests. In a screening survey, we asked participants to list 3 to 5 “seed” papers representative of their research interests but that they did not author. We then used the Semantic Scholar Recommendations API¹⁰ to obtain 25 additional recommended papers for each participant based on their seed papers. We combined the seed and recommended papers, and preprocessed all papers for which a valid PDF file could be found. The abstracts for these papers were then rendered in a list within QLARIFY.

During the study, participants first completed a tutorial that introduced them to QLARIFY’s features (~5 minutes). They were then asked to browse the list of abstracts as if they represented a collection of papers recently recommended to them, e.g., from a colleague or a paper search engine. Participants spent 25 minutes using the interactive expansion interaction to explore abstracts. During their exploration, they were asked to think aloud, sharing observations, questions, or frustrations as they emerged. Afterwards, we engaged participants in a semi-structured interview to elaborate on the perceived advantages and limitations of expandable abstracts as envisioned in QLARIFY. We recorded and transcribed all verbalizations throughout the exploration and subsequent interview (additional details provided in Appendix A).

5.1.2 Participants. We recruited 9 participants (6 male, 3 female; Age: $M = 27.8$, $SD = 9.3$) for the study via university mailing lists and Slack channels. Eight participants were doctoral students within computer and information science, and 1 participant was a research

scientist. Each study lasted approximately 45 minutes, and participants were compensated with \$25 USD for their time. The study was approved by a university institutional review board.

5.1.3 Results. An analysis of interview transcripts and interaction logs uncovered various ways in which QLARIFY supported the exploration of abstracts, such as using interactive expansions to retrieve additional information on-demand from full papers, threaded exploration to dive deeper into aspects of interest, and LLM-generated expandable entities and questions to guide attention. In the following results, we refer to participants with the pseudonyms P1–9.

Abstract expansions allowed an on-demand recursive extraction of high-quality information from papers. Participants were actively engaged with QLARIFY during the study. On average, each participant explored 4.0 papers ($SD = 1.1$, $Mdn = 4.0$) and created a total of 20.8 expansions ($SD = 7.8$, $Mdn = 18.0$). Based on the think-aloud, participants liked how QLARIFY allowed them to surface details from the paper using simple interactions with the abstracts over manually searching for them over the full papers. For example, one participant remarked, “I was impressed by the things that I was able to pull from the paper and the amount of additional details I can get just by reading this abstract” (P3). Additionally, participants pointed to how abstracts typically have a familiar structure that served as a scaffolding and jumping-off-points, and allowed them to use QLARIFY to pull-in additional detailed information from different parts of the full paper when needed:

“I enjoyed the fact that the abstracts were broken down for me. I think that one of the key things about being a PhD student is being able to quickly break down information without having to spend too much time reading the entirety of the paper. And so the abstract allowed for me to do that without having to even read that. Usually the rule of thumb is abstract, intro and conclusion. And with this, I feel I get a bit of the intro, conclusion, results, discussion, analysis, all that within the abstract breakdown.” – P7

Beyond the interaction design, many participants were genuinely surprised at how well the QLARIFY’s generated expansions answered the questions they asked (P1–3, P5, P6, P8). One participant appreciated how the expansions “didn’t just summarize, but seemed to extract meaning from the paper” (P6), reflecting the LLM’s capacity to join fragmented but relevant chunks across a paper to form a coherent, concise, and complete answer. Some participants began their exploration by browsing the abstract of a paper they were intimately familiar with, using their initial interactions to gauge the reliability and accuracy of the generated expansions (P5, P8, P9). Others instead mentioned explicitly trusting the model’s responses and felt confident they could dive into the paper to verify any point if needed (P4, P7, P8).

Moreover, participants found the LLM-suggested question in the question palette often aligned well with their information-seeking intents, and reduced the costs of forming a question to expand the abstract. P8 described the suggested question as, “It seems to almost read my mind when I click on something or highlight something.” P1 noted how “the question is excellent because it captures my intention,” and P5 said, “Every time I think of what the question is, that’s pretty

⁶<https://github.com/allenai/pdf-component-library>

⁷<https://github.com/allenai/mmda>

⁸At the time of submission.

⁹<https://platform.openai.com/docs/api-reference/chat>

¹⁰<https://api.semanticscholar.org/api-docs/recommendations>

much what the question it already thought of is.” We noticed this sentiment reflected in usage behaviors as well; within the question palette, participants selected the LLM-generated question most often (40.1% of clicks, *Define*: 23.5%, *Expand*: 22.5%, *Why*: 13.9%).

Participants also frequently utilized the recursive expansion feature of QLARIFY — 58% of expansions created by participants were threaded (i.e., formed by asking questions about text in another expansion rather than from the abstract). Based on their think-aloud, participants mentioned how the threaded expansions empowered them to dig deeper and ask follow-up questions if an initial expansion did not fully satisfy their information needs (P1, P3, P4, P6, P8). For example, P4 said, *“I really did love the way you could keep going and go branching into a tree.”* By automatically detecting and underlining successive expandable entities immediately after an expansion is created, participants felt QLARIFY could reduce the costs of continued engagement with the abstract. P1 described how these underlines scaffolded his sensemaking process:

“Another thing is in this interface I can keep diving more because that highlighting feature is not provided only on the first level of the abstract, it’s also provided in the next level. When they generate a response in the second level, they also provide these underlining features, which if I have any more questions or concerns or any thoughts I could keep using these interactions to help me make sense of the abstract.” – P1

In summary, the threaded expansions not only supported ongoing engagement but also enhanced the efficiency of sensemaking within the abstract.

AI-suggested expandable entities helped guide exploration of abstracts but could be overwhelming. The majority of the time, participants clicked on a pre-selected underlined entity to create expansions (77.5%) as opposed to manually selecting a custom text span (22.5%). While we initially designed this feature to lower the interaction costs, the think-aloud suggested that the pre-selected entities also served as information cues that can facilitate discovery. Many participants (P1–4, P6, P8) commented how the underlines for expandable entities were helpful visual cues to *“keywords that may be relevant”* and that *“tell me what to focus on”* (P1).

Conversely, if the underlines did not precisely capture participants’ needs, they also appreciated the ability to fall back to highlighting custom text to create an expansion (P1, P2, P8). P6 summarized how the interactive features of QLARIFY in concert could address all of her envisioned information needs over an abstract:

“I think the underlines were already really good. And yeah, that highlighting something as a backup to the underlines already gets you to probably 99% coverage for the things you would want to ask. And then for the last one percent, I could click on the quote and go into the paper.” – P6

On the other hand, visually augmented reading interfaces such as QLARIFY and its underlined entities can introduce distractions for some scholars. Since expanding the abstracts by interacting with the underlined entities required less effort than highlighting

text, P4 noticed how she *“tended to default to whatever was already underlined,”* and it became *“hard to remember that I can just like pick anything out unless I was really curious about it.”* This behavior is not necessarily undesirable, but suggests careful consideration should be given to how augmented interface elements may inadvertently guide or constrain user interactions. P3 further suggested how the underlined entities could open up a rabbit-hole of exploration, derailing the reading of an abstract:

“I can see how it can be distracting for me to be able to finish the entire abstract because I can get hung up on one small detail and be able to dig really deep into that, and just take five times as much time it takes for me to otherwise just read the abstract.” – P3

Similarly, P7 noted how the seemingly limitless freedoms afforded by an abstract expansion interaction could be double-edged and inhibit a sense of completion:

“The endlessness of the underlines, as a completionist, my mindset, I want to click them all. And so I liked the fact that there were no bounds, but I could also feel overwhelmed knowing there are no bounds. So I feel like I could miss something the AI could uncover for me if I just kept clicking all the underlines.” – P7

These observations suggest that while the underlines may help guide an in-depth investigation of details in the paper, it can also potentially hinder a process of triage. As such, balancing interactivity and efficiency is critical for an expandable abstract interaction in ensuring users can engage deeply when needed but also efficiently navigate the content within the abstract and generated expansions.

5.2 Deployment Study

5.2.1 Procedure. To investigate how scholars would interact with expandable abstracts in the wild, we deployed QLARIFY during the 49th International Conference on Very Large Data Bases (VLDB 2023).¹¹ During the duration of the conference, members of the research team invited conference attendees and other scholars to try out QLARIFY via social media announcements through email, Twitter, Slack, and LinkedIn. We created a landing page within QLARIFY to allow users to easily browse the conference proceedings (248 papers), with a paginated, scrollable list of paper metadata and a search bar for filtering papers. Clicking on a title navigated users to a separate page containing an expandable abstract for that paper.

5.2.2 Findings. During the week of the conference and two subsequent weeks, a total of 275 unique users interacted with 50 unique papers using QLARIFY. Based on the interaction logs, each user created 3.3 expansions on average (SD = 4.6, Mdn = 2.0). We found users expanded abstracts using the pre-selected expandable entities rather than manually selecting custom text spans (80.4% vs 19.6% of interactions), corroborating similar observations of usage from the interview study.

We further found that users more often selected one of the three static questions (i.e., *Expand*, *Define*, and *Why*; 88.2% of interactions) than the more specific LLM-generated questions (e.g., for the entity *“Fries scheduler with consistency guarantees”*, the suggested question was *“How does the Fries scheduler ensure consistency in*

¹¹<https://vldb.org/2023/?papers-research>

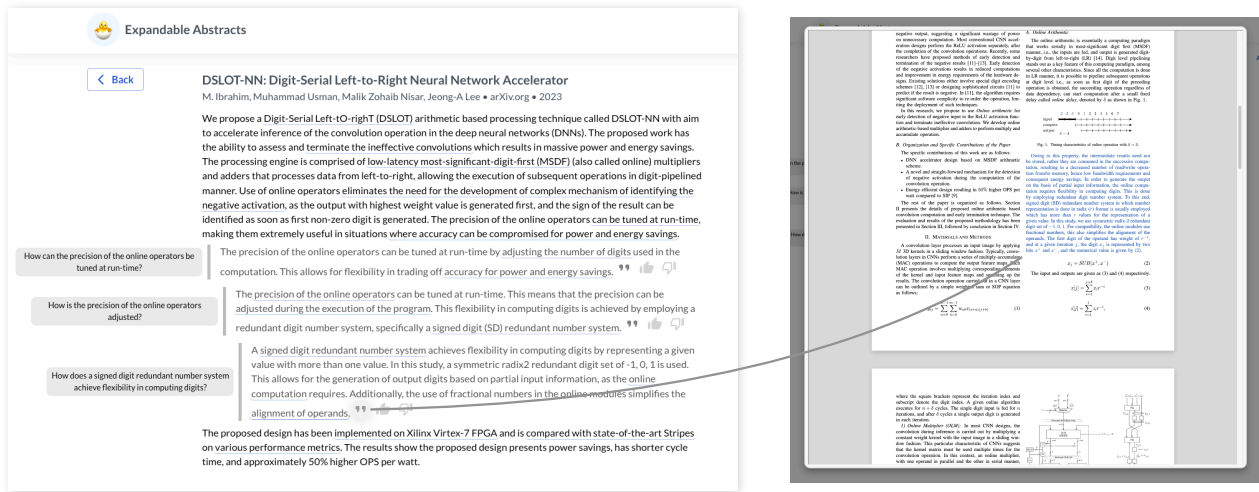


Figure 6: QLARIFY’s user interface, showing an expansion thread (left) created by P8 from the interview study while exploring an abstract. Users could easily dive into the paper with evidence for an expansion highlighted in context (right); during their exploration, P8 opened the paper for the third expansion.

runtime reconfigurations?”; 11.8% of interactions). Within the the static questions, users created 41.6% of expansions with *Expand*, 31.5% with *Define*, and 15.1% with *Why*, a distribution that closely corroborates the frequency of question types we observed in our formative study. These results are in contrast to our interview study, where participants were more likely to select LLM-generated questions (40.1% of expansions). This difference might be due to the higher relevance of abstracts in our interview study compared to our deployment (i.e., paper abstracts were selected based on personalized recommendations for each participant). These results from our deployment study suggest that even when users’ information needs were grounded in specific entities, they either do not always have specific questions in mind or that the suggested questions did not reflect their specific intents. On the other hand, when the suggested questions were found useful, they could lower interaction costs by allowing users to expand with a single click.

While the majority of the time users created a single level of expansion from the original abstract, a significant portion (27.7%) of interactions were threaded, meaning users recursively asked follow on questions by selecting additional entities in the expanded text. Some users recursively created up to 5 nested expansions. These results demonstrate the users’ needs for recursive expansion of abstracts and QLARIFY’s ability to support this. In comparison to our interview study, fewer users in our deployment study created threaded expansions (58.0% vs. 27.7%, respectively), perhaps due to differences in user engagement or relevance of abstracts.

We also observed users actively engaging with the attributed evidence paragraphs and viewing the paper itself. In 14.8% of expansions, users viewed the attributed evidence paragraph for a generated expansion. About 60% of the time users were satisfied with the extracted evidence, while 40% of the time they further opened the PDF to view the highlighted evidence in the context of the paper. These behaviors suggest users were either interested

in understanding the supporting information or sought to use the evidence as an efficient entry point into reading the paper.

Altogether, our deployment study suggests that an expandable abstract interaction is an intuitive and powerful way to ask clarification questions for abstract-grounded information needs, allowing users to retrieve answers and evidence on-demand and in-situ. Complementing the qualitative insights from our interview study, findings from this deployment study provide evidence within a natural context of user interaction, exemplifying real world usage behaviors of QLARIFY at a conference where people are actively triaging papers for themselves.

5.3 Evaluation of LLM-Generated Expansions

To inspect the quality of expansions generated by the LLMs, we analyzed a subset of the data collected in the deployment study. Members of the research team annotated 120 randomly sampled expansions for which an answer was found (30 for each of the *Define*, *Expand*, *Why*, and AI-suggested questions).

Of the 120 expansions, 105 (87.5%) were entirely accurate (i.e., all statements were grounded in verifiable information from the paper) corroborating perceptions of expansion quality by scholars in our interview study. We did not explicitly assess relevance, an important but challenging property to operationalize in user-facing systems. However, through our annotation of the sampled expansions and observations of participants in our interview study, we found information within the LLM-generated expansions tended to not only be accurate but also relevant to the asked question. Errors found within the analyzed expansions included:

- **Inaccurate details:** Seven (5.8%) expansions included detail inaccuracies, often involving numerical or mathematical content. These included false navigational references (e.g., attributing statements to an incorrect section in the paper), numerical values in experimental results (e.g., fabricated

numbers in “the additional mean overhead time of 0.47s is only 12.8% of the average episode duration of 3.67s”), and acronyms (e.g., describing FMs as “language guided models” rather than “foundation models”).

- **Missing content:** Eight (6.7%) expansions contained phrases such as, “the paper does not provide explicit details for. . .” This error tended to occur when a portion of the relevant information was provided in a table rather than in the body text of the paper. Rather than a limitation of LLMs, this perhaps reflects how QLARIFY preprocesses papers into a flat representation without delimitation for structural or visually salient content such as tables. In other cases, the context provided to the LLM may have lacked sufficient information to answer the question, suggesting the need to further investigate robust chunk retrieval techniques.

Our error analysis reveals the subtlety of hallucinations within LLM-generated text, such as how plausible-seeming yet unfounded details are embedded into an otherwise accurate expansion. Indeed, it is worth noting that no scholars in our interview study surfaced any errors, and we identified errors only through extensive checking with the original paper. These hallucinations can thus be challenging to detect — especially within exploratory processes such as triage — and potentially lead to harmful misinterpretations and an erosion of trust in the reliability of the generated expansions.

6 DISCUSSION

In this paper, we proposed a new document-centric interaction technique that leverages the generative capabilities of LLMs to recursively expand summaries. We presented a system, QLARIFY, that instantiates this technique to allow scholars to incrementally expand abstracts in response to personalized information needs and incorporate details from full papers. Findings from our interview and deployment studies reveal how QLARIFY enables scholars to retrieve information from a paper on-demand with one-click questions. Here, we discuss how recursive expansions can effectively bridge the information gap between abstracts and papers and the implications such AI-powered support tools may have for the scholarly research process.

6.1 Bridging Abstracts and Papers with Recursive Expansions

In this work, we demonstrated that enriching abstracts with expansion interactions helped scholars to navigate some of the informational gaps between an abstract and a paper. The initial positive response to QLARIFY in our studies opens up future exciting dimensions of the design space to explore.

As mentioned in Section 3.2 there are additional interface interventions and computational approaches for expanding abstracts for information needs not grounded in the abstract’s text. For instance, expanding abstracts with agnostic information needs, i.e., questions that scholars have divorced of any specific abstract (e.g., “What are their contributions”) could draw on prior work for the extraction of information for key mechanisms, e.g., contributions, method, findings (à la [9, 19]). Alternatively, an expansion interaction could allow users to ask free-form questions, which some participants in our interview study suggested could also be a useful fall-back in

QLARIFY if pre-selected entities or highlighting custom text did not capture their intended questions.

Handling latent information needs, or those unknown to a reader until explicitly surfaced, could involve identifying and surfacing information from a paper relevant to a scholar but that is not grounded in the abstract. The challenge here lies in expanding the abstract in a way that is tailored to each scholar’s exploratory goals, and possibly presented as a suggested key question index (e.g., [2]).

To support scholars in switching between the triage of abstracts and reading of papers, we envision opportunities to further improve abstract expansions and facilitate a more seamless transition between the two texts. One approach could involve surfacing additional relevant content from the paper. For instance, we found participants in our interview study used the ability to drill down into a paper not only for verification of the generated expansions, but also to quickly open the paper and browse visual content, e.g., a system diagram or a table of experimental results. Expansions could help streamline the retrieval of such non-textual content scholars desire from papers.

Expansions could also provide more entry points from an abstract into a paper. QLARIFY provides one per expansion in allowing scholars to jump into the paper’s context by drilling down into attributed evidence. Systems could instead seed paper reading by leveraging a scholar’s history of expansions. For instance, augmented reading interfaces could help guide a scholar’s attention to other related passages throughout the paper (e.g., [19]) or visualize regions of the paper related to other expansions created in an abstract. Finally, systems could consider how to persist a scholar’s newly gained knowledge from reading the paper when returning back to the abstract and triaging process.

6.2 Considerations for AI-Infused Scholarly Support Tools

Our results also highlight implications for designing and deploying future AI-infused scholarly support tools.

6.2.1 Provide control of AI augmentations. Our studies surfaced the possible benefit of allowing users to control the amount of augmentation in QLARIFY. Some participants mentioned that the in-situ expansions could be distracting to a user’s reading flow, especially with verbose or inaccurate expansions. Though most participants found the underlined entities as helpful indications of expandable information, some participants felt such strong visual indicators could derail reading the entire abstract. To meet the specific desires of individual scholars for AI assistance, we recommend future intelligent reading interfaces provide controls for users to effortlessly customize visually salient or AI-enabled features, such as the placement of expansions (e.g., fluid, in a sidebar, or as a popup) or density of pre-selected expandable entities.

6.2.2 Consider unintended consequences of long-term deployment. The development of scholarly support tools should be done while considering the implications of deploying such systems. While LLM-generated questions within QLARIFY can help scaffold an active reading process, they may also restrict a scholar’s agency over their exploratory triaging process. Furthermore, AI-augmented interactions such as abstract expansions may disincentivize scholars

from reading full papers, and instead encourage more superficial exploration through interactions with abstracts only. Participants in the interview study commented on how these risks could be most damaging to new scholars, as the “path of least resistance” provided by these AI-augmented scholarly support tools could hamper the learning that would otherwise accumulate over years of triaging and consuming the literature.

6.3 Future Work

While QLARIFY was designed and evaluated with papers largely within computer science, the underlying interaction paradigm can transcend disciplines. For instance, prior work has shown that reading medical literature can be challenging and overwhelming due to barriers such as dense and unfamiliar terminology, not knowing what to read, and the inability to find answers to specific questions [2]. One avenue for future work could explore how expandable abstracts with similar question asking and answering assistance could make medical papers more accessible by providing just-in-time responses to questions directly within an abstract, without requiring lay readers to interact with the full paper. A similar idea might allow policymakers to understand the implications of scientific developments, a critical need [69].

Furthermore, could a system learn about a user’s interests based on what they choose to expand in abstracts they interact with? Such a system could then automatically expand personalized entities of interest as users encounter new abstracts to further lower interaction costs and encourage exploration, or even regenerate abstracts to be personalized to a user’s interests. Another exciting direction is to explore support for expansion contexts that include multiple documents using the proposed abstract expansion interaction. For example, how should we design future versions of QLARIFY that can allow users to expand on related work sections and explore information across many cited papers, and how should such systems synthesize the information retrieved across these multiple documents? Finally, a wealth of opportunities exists to examine the merits of recursively expandable summaries for documents in other domains (e.g., legal documents, medical notes, or discussions in online forums).

6.4 Limitations

Through our studies, we sought to provide an initial characterization of how researchers can use and benefit from recursively expandable abstracts and identify opportunities for further iterative design. Additional evaluations could help empirically quantify the advantages of expandable abstracts, for instance in comparison to researchers’ current techniques for exploring paper abstracts and other LLM-enabled applications, e.g., conventional chatbot-style interactions with a paper. Our iterative design process and studies with QLARIFY suggest design principles for an expansion interaction that we believe can generalize beyond the scientific papers included in our evaluation.

Expanding abstracts also naturally requires access to a paper’s full text and PDF, for instance in question answering to populate an expansion, in providing extractive attribution, and in visualizing attributed paper evidence within the PDF. During the duration of our studies, we had the privilege of institutional access to full text

PDFs for our studies. However, we note a significant portion of the scientific literature remains inaccessible behind paywalls enacted by academic publishers. While numerous legal and institutional challenges remain, open access initiatives (e.g., the Open Access movement¹², arXiv, S2ORC [44]) have made notable strides in changing the landscape of scholarly publishing to encourage more accessible dissemination of scientific knowledge.

7 CONCLUSION

In this work, we introduce QLARIFY, a novel reading interface for scientific paper abstracts that supports scholars in recursively expanding abstracts on-demand with clarifying information from the full text of papers. To reduce the costs of expanding, QLARIFY provides intelligent assistance leveraging LLMs to identify informative entities within abstracts to expand, suggest intent-inferring questions that scholars can ask in a single click, and generate concise, attributed responses. Through an interview study with scholars, we found these interventions enabled scholars to more rapidly and thoroughly address information-seeking curiosities during paper triage. A subsequent deployment study revealed how scholars would use expandable abstracts for paper triage in a real-world environment. We believe QLARIFY and the abstract expansion paradigm contribute a valuable step toward understanding how mixed-initiative systems can effectively support the low-effort and just-in-time exploration of scientific documents.

ACKNOWLEDGMENTS

This work is supported by NSF Grant OIA-2033558, ONR grant N00014-21-1-2707, and the Allen Institute for Artificial Intelligence (AI2). The authors thank Luca Soldaini, Kyle Lo, Ben Newman, and Cassidy Trier for the insightful discussions and feedback. Finally, we thank the participants in all of our studies, without whom this work would not have been possible.

REFERENCES

- [1] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the Literature Graph in Semantic Scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. Association for Computational Linguistics, New Orleans - Louisiana, 84–91. <https://doi.org/10.18653/v1/N18-3011>
- [2] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. PaperPlain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Transactions on Computer-Human Interaction* 30, 5 (Sept. 2023), 74:1–74:38. <https://doi.org/10.1145/3589955>
- [3] Jöran Beel and Bela Gipp. 2009. Google Scholar’s Ranking Algorithm: An Introductory Overview. In *Proceedings of the 12th International Conference of the International Society for Scientometrics and Informetrics*, Vol. 1. Rio de Janeiro, Brazil, 230–241.
- [4] Mark Bernstein. 2009. On hypertext narrative. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*. ACM, Torino Italy, 5–14. <https://doi.org/10.1145/1557914.1557920>
- [5] Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roei Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed Question Answering: Evaluation

¹²cf. <https://www.doaj.org/>, <https://www.budapestopenaccessinitiative.org/read/>

- and Modeling for Attributed Large Language Models. <http://arxiv.org/abs/2212.08037> arXiv:2212.08037 [cs].
- [6] Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications* 8, 1 (Oct. 2021), 1–15. <https://doi.org/10.1057/s41599-021-00903-w> Number: 1 Publisher: Palgrave.
 - [7] 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. <https://doi.org/10.1002/asi.23329> arXiv:<https://arxiv.org/abs/1505.03001> eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23329>.
 - [8] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme Summarization of Scientific Documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 4766–4777. <https://doi.org/10.18653/v1/2020.findings-emnlp.428>
 - [9] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. SOLVENT: A Mixed Initiative System for Finding Analogies between Research Papers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–21. <https://doi.org/10.1145/3274300>
 - [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 (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3544548.3580847>
 - [11] ChatDOC. 2023. ChatDOC - Chat with your documents. Retrieved September 26, 2023 from <https://chatdoc.com/>
 - [12] ChatPDF. 2023. ChatPDF - Chat with any PDF. Retrieved September 26, 2023 from <https://www.chatpdf.com/>
 - [13] 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*. ACM, San Diego California USA, 739–742. <https://doi.org/10.1145/2020408.2020524>
 - [14] Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making Sentences Stand-Alone. *Transactions of the Association for Computational Linguistics* 9 (April 2021), 447–461. https://doi.org/10.1162/tacl-a_00377
 - [15] Johan S. G. Chu and James A. Evans. 2021. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences* 118, 41 (Oct. 2021), e2021636118. <https://doi.org/10.1073/pnas.2021636118> Publisher: Proceedings of the National Academy of Sciences.
 - [16] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 4599–4610. <https://doi.org/10.18653/v1/2021.naacl-main.365>
 - [17] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. <http://arxiv.org/abs/2305.14314> arXiv:2305.14314 [cs] version: 1.
 - [18] Elicit. 2023. Elicit - The AI Research Assistant. Retrieved September 26, 2023 from <https://elicit.com/>
 - [19] Raymond Fok, Hita Kambhampettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti Hearst, Andrew Head, and Daniel S Weld. 2023. Scim: Intelligent Skimming Support for Scientific Papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. ACM, Sydney NSW Australia, 476–490. <https://doi.org/10.1145/3581641.3584034>
 - [20] National Science Foundation. 2021. Publications Output: U.S. Trends and International Comparisons. <https://nces.nsf.gov/pubs/nsb20214>
 - [21] Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and Revising What Language Models Say. Using Language Models. <https://doi.org/10.48550/arXiv.2210.08726> arXiv:2210.08726 [cs].
 - [22] Samira Ghodrathnama, Mehrdad Zakershahrah, and Fariborz Sobhanmanesh. 2021. Adaptive Summaries: A Personalized Concept-based Summarization Approach by Learning from Users' Feedback. <http://arxiv.org/abs/2012.13387> arXiv:2012.13387 [cs].
 - [23] GROBID. 2008–2023. GROBID. <https://github.com/kermitt2/grobid>
 - [24] Tovi Grossman, Fanny Chevalier, and Rubaiat Habib Kazi. 2015. Your Paper is Dead! Bringing Life to Research Articles with Animated Figures. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. Association for Computing Machinery, New York, NY, USA, 461–475. <https://doi.org/10.1145/2702613.2732501>
 - [25] Jianguan He, Qing Ping, Wen Lou, and Chaomei Chen. 2019. PaperPoles: Facilitating adaptive visual exploration of scientific publications by citation links. *Journal of the Association for Information Science and Technology* 70, 8 (2019), 843–857. <https://doi.org/10.1002/asi.24171> eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24171>.
 - [26] 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 (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–18. <https://doi.org/10.1145/3411764.3445648>
 - [27] Nouf Ibrahim Altmami and Mohamed El Bachir Menai. 2022. Automatic summarization of scientific articles: A survey. *Journal of King Saud University - Computer and Information Sciences* 34, 4 (April 2022), 1011–1028. <https://doi.org/10.1016/j.jksuci.2020.04.020>
 - [28] Emi Ishita, Yasuko Hagiwara, Yukiko Watanabe, and Yoichi Tomiura. 2018. Which Parts of Search Results do Researchers Check when Selecting Academic Documents?. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL '18)*. Association for Computing Machinery, New York, NY, USA, 345–346. <https://doi.org/10.1145/3197026.3203867>
 - [29] Farnaz Jahanbakhsh, Elnaz Nouri, Robert Sim, Ryen W. White, and Adam Fourney. 2022. Understanding Questions that Arise When Working with Business Documents. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 341:1–341:24. <https://doi.org/10.1145/3555761>
 - [30] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (Dec. 2023), 1–38. <https://doi.org/10.1145/3571730>
 - [31] 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 (UIST '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3526113.3545660>
 - [32] 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 *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–23. <https://doi.org/10.1145/3491102.3517470>
 - [33] Hyeonsu B Kang, Nouran Soliman, Matt Latzke, Joseph Chee Chang, and Jonathan Bragg. 2023. ComLitte: Literature Discovery with Personal Elected Author Committees. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–20. <https://doi.org/10.1145/3544548.3581371>
 - [34] Hyeonsu B. Kang, Sherry Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. 2023. Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking. <https://doi.org/10.1145/3586183.3606759>
 - [35] Tae Soo Kim, Matt Latzke, Jonathan Bragg, Amy X. Zhang, and Joseph Chee Chang. 2023. Papeos: Augmenting Research Papers with Talk Videos. <https://doi.org/10.1145/3586183.3606770> arXiv:2308.15224 [cs].
 - [36] Jeffrey W. Knopf. 2006. Doing a Literature Review. *PS: Political Science and Politics* 39, 1 (2006), 127–132. <https://www.jstor.org/stable/20451692> Publisher: [American Political Science Association, Cambridge University Press].
 - [37] Wei-Jen Ko, Te-yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li. 2020. Inquisitive Question Generation for High Level Text Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6544–6555. <https://doi.org/10.18653/v1/2020.emnlp-main.530>
 - [38] Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An Empirical Survey on Long Document Summarization: Datasets, Models, and Metrics. *Comput. Surveys* 55, 8 (Dec. 2022), 154:1–154:35. <https://doi.org/10.1145/3545176>
 - [39] Vaibhav Kumar and Alan W Black. 2020. ClarQ: A large-scale and diverse dataset for Clarification Question Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7296–7301. <https://doi.org/10.18653/v1/2020.acl-main.651>
 - [40] Esther Landhuis. 2016. Scientific literature: Information overload. *Nature* 535, 7612 (July 2016), 457–458. <https://doi.org/10.1038/nj7612-457a> Number: 7612 Publisher: Nature Publishing Group.
 - [41] Byungjoo Lee, Olli Savisaari, and Antti Oulasvirta. 2016. Spotlights: Attention-Optimized Highlights for Skim Reading. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5203–5214. <https://doi.org/10.1145/2858036.2858299>
 - [42] Ziming Liu. 2005. Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *Journal of Documentation* 61, 6 (Jan. 2005), 700–712. <https://doi.org/10.1108/00220410510632040> Publisher: Emerald Group Publishing Limited.

- [43] 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 Langgan, Matt Lutzke, 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. <http://arxiv.org/abs/2303.14334> arXiv:2303.14334 [cs].
- [44] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4969–4983. <https://doi.org/10.18653/v1/2020.acl-main.447>
- [45] Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2021. Paper Forager: Supporting the Rapid Exploration of Research Document Collections. In *Proceedings of Graphics Interface (GI 2021)*. Canadian Information Processing Society, Virtual Event, 237–245.
- [46] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1906–1919. <https://doi.org/10.18653/v1/2020.acl-main.173>
- [47] Yan Meng, Liangming Pan, Yixin Cao, and Min-Yen Kan. 2023. FOLLOWUPQG: Towards Information-Seeking Follow-up Question Generation. <http://arxiv.org/abs/2309.05007> arXiv:2309.05007 [cs].
- [48] Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. <http://arxiv.org/abs/2203.11147> arXiv:2203.11147 [cs].
- [49] Aditi Mishra, Utkarsh Soni, Anjana Arunkumar, Jinbin Huang, Bum Chul Kwon, and Chris Bryan. 2023. PromptAid: Prompt Exploration, Perturbation, Testing and Iteration using Visual Analytics for Large Language Models. <http://arxiv.org/abs/2304.01964> arXiv:2304.01964 [cs].
- [50] Sheshera Mysore, Mahmood Jasim, Haoru Song, Sarah Akbar, Andre Kenneth Chase Randall, and Narges Mahyar. 2023. How Data Scientists Review the Scholarly Literature. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval (CHIIR '23)*. Association for Computing Machinery, New York, NY, USA, 137–152. <https://doi.org/10.1145/3576840.3578309>
- [51] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. WebGPT: Browser-assisted question-answering with human feedback. <https://doi.org/10.48550/arXiv.2112.09332> arXiv:2112.09332 [cs].
- [52] Theodor H Nelson. 1983. *Literary machines*. (1983).
- [53] Benjamin Newman, Luca Soldaini, Raymond Fok, Arman Cohan, and Kyle Lo. 2023. A Controllable QA-based Framework for Decontextualization. <http://arxiv.org/abs/2305.14772> arXiv:2305.14772 [cs].
- [54] 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. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–20. <https://doi.org/10.1145/3544548.3580841>
- [55] Pallets. 2023. Welcome to Flask – Flask Documentation (2.3.x). Retrieved August 27, 2023 from <https://flask.palletsprojects.com/en/2.3.x/>
- [56] Connected Papers. 2023. Connected Papers | Find and explore academic papers. Retrieved September 26, 2023 from <https://www.connectedpapers.com/>
- [57] 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 (CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 2264–2271. <https://doi.org/10.1145/2851581.2892334>
- [58] Napol Rachatasumrit, Jonathan Bragg, Amy X. Zhang, and Daniel S Weld. 2022. CiteRead: Integrating Localized Citation Contexts into Scientific Paper Reading. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 707–719. <https://doi.org/10.1145/3490099.3511162>
- [59] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. <https://doi.org/10.18653/v1/D19-1410>
- [60] Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *Comput. Surveys* 55, 10 (Feb. 2023), 197:1–197:45. <https://doi.org/10.1145/3560260>
- [61] Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, Ryan A. Rossi, and Franck Dernoncourt. 2023. PDFTriage: Question Answering over Long, Structured Documents. <http://arxiv.org/abs/2309.08872> arXiv:2309.08872 [cs].
- [62] Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Ido Dagan, and Yael Amsterdamer. 2022. Interactive Query-Assisted Summarization via Deep Reinforcement Learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 2551–2568. <https://doi.org/10.18653/v1/2022.naacl-main.184>
- [63] Ayah Soufan, Ian Ruthven, and Leif Azzopardi. 2022. Searching the Literature: An Analysis of an Exploratory Search Task. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*. ACM, Regensburg Germany, 146–157. <https://doi.org/10.1145/3498366.3505818>
- [64] Meta Open Source. 2023. React. Retrieved August 27, 2023 from <https://react.dev/>
- [65] 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*. ACM, Salerno Italy, 1–5. <https://doi.org/10.1145/3399715.3399830>
- [66] Carol Tenopir, Donald King, Sheri Edwards, and Lei Wu. 2009. Electronic Journals and Changes in Scholarly Article Seeking and Reading Patterns. *Carol Tenopir* 61 (Jan. 2009), 5–32. <https://doi.org/10.1045/november2008-tenopir>
- [67] Maartje ter Hoeve, Robert Sim, Elnaz Nouri, Adam Fournay, Maarten de Rijke, and Ryen W. White. 2020. Conversations with Documents: An Exploration of Document-Centered Assistance. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval (CHIIR '20)*. Association for Computing Machinery, New York, NY, USA, 43–52. <https://doi.org/10.1145/3343413.3377971>
- [68] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kuleshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YuGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. <https://doi.org/10.48550/arXiv.2201.08239> [cs].
- [69] Chris Tyler, K L Akerlof, Alessandro Allegra, Zachary Arnold, Henriette Canino, Marius A Doornenbal, Josh A Goldstein, David Budtz Pedersen, and William J Sutherland. 2023. AI tools as science policy advisers? The potential and the pitfalls. *Nature* 622 (2023), 27–30. <https://api.semanticscholar.org/CorpusID:263104425>
- [70] Yun Wang, Dongyu Liu, Huamin Qu, Qiong Luo, and Xiaojuan Ma. 2016. A Guided Tour of Literature Review: Facilitating Academic Paper Reading with Narrative Visualization. In *Proceedings of the 9th International Symposium on Visual Information Communication and Interaction (VINCI '16)*. Association for Computing Machinery, New York, NY, USA, 17–24. <https://doi.org/10.1145/2968220.2968242>
- [71] Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively Summarizing Books with Human Feedback. <http://arxiv.org/abs/2109.10862> arXiv:2109.10862 [cs].
- [72] Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R. Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. 2019. ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks. <https://doi.org/10.48550/arXiv.1909.01716> arXiv:1909.01716 [cs].
- [73] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–21. <https://doi.org/10.1145/3544548.3581388>
- [74] Amy X. Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 2082–2096. <https://doi.org/10.1145/2998181.2998235>
- [75] Xiaoyu Zhang, Senthil Chandrasegaran, and Kwan-Liu Ma. 2021. ConceptScope: Organizing and Visualizing Knowledge in Documents based on Domain Ontology. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–13. <https://doi.org/10.1145/3411764.3445396>

- [76] Xiaoyu Zhang, Jianping Li, Po-Wei Chi, Senthil Chandrasegaran, and Kwan-Liu Ma. 2023. ConceptEVA: Concept-Based Interactive Exploration and Customization of Document Summaries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–16. <https://doi.org/10.1145/3544548.3581260>

A SEMI-STRUCTURED INTERVIEW QUESTIONS

We used the following questions to elicit qualitative insights from participants in the interview study. These questions were used to initially guide the discussion, and probing questions were used to further elaborate on responses.

- (1) Summarize your overall impressions of this interactive abstract interaction. What did you like or dislike?
- (2) What other features or improvements would you want in a future system with interactive abstracts?
- (3) Did you prefer asking questions by highlighting text or by clicking one of the underlined entities, and why?
- (4) How do you feel about the choice of three static questions provided in the question palette? Are there other questions you would have wanted?

- (5) Did you use either the attributed evidence paragraph or the ability see the evidence in the context of the paper? If so, how did you use it and was it useful?
- (6) How did you feel about the quality of the generated responses?
- (7) How did you feel about the quality of the suggested question in the question palette?
- (8) Do you think this interactive abstract would be useful in your own research workflows, and if so, how?

B EXAMPLE OUTPUTS FOR QLARIFY'S ABSTRACT EXPANSION PIPELINE

Figure 7 shows the augmented reading interface in QLARIFY for [34], with expandable entities underlined and no expansions created. Table 3 lists LLM-generated questions and expansions for each detected expandable entity, and Table 2 lists example expansions for the static questions.

C LLM PROMPTS

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

Hyeonsu B Kang, Sherry Wu, Joseph Chee Chang, A. Kittur • arXiv.org • 2023

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.

Figure 7: QLARIFY’s user interface with expandable entities underlined and no expansions created (abstract shown is from [34]).

Table 2: Examples of LLM-generated expansions for the three static questions in QLARIFY (from [34]; abstract in Figure 7).

Question	Entity	Expansion
<i>Define</i>	“mixed-initiative workflows”	Mixed-initiative workflows refer to a design approach that combines both human input and automated assistance in the process of conducting literature reviews and synthesizing knowledge from scientific papers. It involves the integration of user input, such as relevant seed threads, with computational algorithms and interaction features to expand and structure the threads. This approach allows scholars to start with a pre-generated structure and customize it according to their needs, facilitating efficient sensemaking, broadening perspectives, and increasing curiosity.
<i>Expand</i>	“broaden their perspectives”	Participants in the study engaged with Synergi-generated threads of research to “broaden their perspectives.” This means that the threads provided them with a wider range of information and insights, allowing them to consider different viewpoints and explore various aspects of the literature. By broadening their perspectives, participants were able to gain a more comprehensive understanding of the topic they were reviewing.
<i>Why</i>	“building research threads scattered across multiple papers remains a challenge”	Building research threads scattered across multiple papers remains a challenge because it requires synthesizing information from multiple sources and organizing them in a coherent manner. Prior approaches have supported largely manual bottom-up synthesis workflows, which can be time-consuming and inefficient. Synergi addresses this challenge by automatically retrieving important papers from a 2-hop neighborhood on the citation graph and summarizing them in a hierarchical manner, allowing users to quickly discover relevant threads and understand them through synthesis by other scholars.

Table 3: Examples of QLARIFY’s LLM-generated expandable entities, suggested questions, and expansions. All examples are actual system output for the paper *Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking* [34] (abstract in Figure 7).

Entity	Question	Expansion
“top-down synthesis (and LLMs)”	What is meant by “top-down synthesis” and “bottom-up synthesis”?	Top-down synthesis refers to the approach of using large language models (LLMs) to generate summaries or syntheses of research papers. It is characterized by limited personalization and iteration on the output. On the other hand, bottom-up synthesis involves practices such as citation chasing and footnote chasing to discover important papers related to a research problem. It is more time-consuming but allows for a more comprehensive synthesis.
“mixed-initiative workflows”	What is the definition of a “mixed-initiative workflow”?	A mixed-initiative workflow is a type of workflow that combines both human input and automated processes. It involves user input of relevant seed threads, which are then expanded and structured using citation graphs and LLMs. Scholars can start with a pre-generated structure and customize it as they wish, allowing for personalization and iteration.
“relevant seed threads”	What is meant by “relevant seed threads”?	“Relevant seed threads” refers to the structured research threads that are generated based on the selected clips and their associated citation contexts and seed references. These threads are considered relevant because they align with the user’s interests and provide valuable information about specific research areas in the literature.
“citation graphs and LLMs”	How does Synergi use citation graphs and LLMs to expand and structure research threads?	Synergi uses citation graphs to retrieve important papers from a 2-hop neighborhood and summarize them in a hierarchical manner. It generates a 3-level hierarchy of research threads by capturing the core commonalities among the papers. It also utilizes LLMs like GPT4 to condense scholarly text and provide synthesis.
“iterate and customize”	What is the process for scholars to iterate and customize the threads-and-subthreads structure?	Scholars can review the system-generated threads, curate the ones that interest them, and build upon them iteratively. They can easily curate useful threads, references, and contexts into an editor using drag-and-drop. They can also prioritize references by using the group-by-reference view. They can expand and collapse the structured thread recommendations to access the relevant citation contexts and source materials.
“In our evaluation”	How was the evaluation of Synergi conducted?	The evaluation of Synergi involved 12 participants and domain experts. They compared Synergi to a baseline system, Threddy, and GPT4. The evaluation measured the quality of the outlines, efficiency in constructing them, and participants’ perception of Synergi-generated threads and experience. The evaluation used experts’ judgment, survey responses, qualitative data from interviews, and participants’ interaction logs.
“broaden their perspectives”	How does Synergi help scholars broaden their perspectives?	Synergi helps scholars broaden their perspectives by providing them with system-generated threads of research that they can quickly scan to understand how different sub-group structures map to the broader literature. This understanding allows scholars to orient their attention towards specific areas that align with their interests.
“future design implications”	What are the future design implications for thread-based, mixed-initiative scholarly synthesis support tools?	The future design implications for thread-based, mixed-initiative scholarly synthesis support tools include incorporating LLMs as components in computational pipelines, exploring the design space of interaction designs that benefit users in discovering, prioritizing, extracting, organizing, and synthesizing knowledge, and conducting additional ablation studies to understand the contributions of each component in the pipeline.

Table 4: Prompts used in QLARIFY. { } refers to a placeholder.

Task	Prompt
Expandable Entity Extraction	<p>You are a helpful research assistant that asks questions about abstracts of scientific papers.</p> <p>List all questions that a curious reader might have after reading this abstract. These questions must not be answerable given the abstract, but may be answerable given the full paper. These questions could help clarify vague terms, define jargon, request for more detail, or ask for justification. Each question should be short and not contain multiple sub-questions. Provide a phrase (three words or less) verbatim from the abstract that motivates each question.</p> <p>Title: {Title} Abstract: {Abstract}</p> <p>{Examples}</p> <p>Questions:</p>
Question Generation	<p>You are a helpful research assistant that predicts what question a reader might have.</p> <p>A reader has highlighted a span of text in the abstract. What is the most likely question they could ask about the span? The question must not be answerable given the abstract, but may be answerable given the full paper. The question may help clarify vague terms, define jargon, request for more detail, or ask for justification. The question should be short and not contain multiple sub-questions. Try framing the question as: How? Why? What? Such as?</p> <p>Abstract: {Abstract} Target span: "{Entity}", in the sentence "{Sentence}" Question:</p>
Question Answering	<p>You are a helpful research assistant that answers questions about scientific papers.</p> <p>Answer the question based on the following excerpts from the full text of the paper. Incorporate quotes verbatim from the excerpts when relevant. If the question cannot be answered from the provided context, reply "No answer." Your answer should be {Response Length}.</p> <p>{Examples}</p> <p>Context: {Context} Question: {Question} Answer:</p>