박 사 학 위 논 문
Ph.D. Dissertation

AI 시스템과 인간 인지과정의 정렬을 통한 지식 이해 지원

Aligning AI Systems with Human Cognitive Processes for Knowledge Understanding

2025

이 윤 주 (李 倫 周 Lee, Yoonjoo)

한 국 과 학 기 술 원

Korea Advanced Institute of Science and Technology

박 사 학 위 논 문

AI 시스템과 인간 인지과정의 정렬을 통한 지식 이해 지원

2025

이 윤 주

한 국 과 학 기 술 원

전산학부

AI 시스템과 인간 인지과정의 정렬을 통한 지식 이해 지원

이 윤 주

위 논문은 한국과학기술원 박사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2025년 6월 2일

심사위원장	김 주 호	(인)
심 사 위 원	오 혜 연	(인)
심 사 위 원	서 민 준	(인)
심 사 위 원	Eytan Adar	(인)
심 사 위 원	Sherry Tongshuang Wu	(인) Tongshuang Wu.

Aligning AI Systems with Human Cognitive Processes for Knowledge Understanding

Yoonjoo Lee

Advisor: Juho Kim

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science

Daejeon, Korea
June 26, 2025

Approved by

Juho Kim
Associate Professor of Computer Science

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

DCS 이윤주. AI 시스템과 인간 인지과정의 정렬을 통한 지식 이해 지원. 전산학부 . 2025년. 111+vi 쪽. 지도교수: 김주호. (영문 논문)
Yoonjoo Lee. Aligning AI Systems with Human Cognitive Processes for Knowledge Understanding. School of Computing . 2025. 111+vi pages. Advisor: Juho Kim. (Text in English)

Abstract

Despite the growing role of Artificial Intelligence (AI) in knowledge processing, current models fail to effectively tailor their outputs to individual users' needs for processing information. Rather than AI systems adapting to users, users often need to adjust their behavior—engaging in iterative prompting and piecing together fragmented responses across multiple interactions to make sense of AI-generated knowledge. This challenge is particularly evident in domains where users engage with complex information, such as education, decision-making, and scientific knowledge consumption, leading to inefficient interactions and missed insights.

To address this, I develop AI systems that adapt their outputs to align with users' cognitive processes by representing and modeling how humans understand information. My work takes a dual approach—enabling AI systems to better understand human cognition while helping humans interpret and evaluate AI's capabilities—ultimately enhancing AI's role in augmenting human knowledge. In this dissertation, I will discuss (1) methods for building AI systems that personalize knowledge delivery based on users' tasks and contexts, such as helping children grasp scientific concepts or assisting researchers in finding relevant papers, and (2) evaluation methods to assess whether AI models can generate outputs aligned with human cognitive processes. My work shows that when AI systems adapt to each user by aligning with their cognitive processes, they enhance engagement and deepen comprehension. By bridging the gap between AI-generated knowledge and human understanding, we can make AI more useful and interpretable in real-world settings.

Keywords AI Alignment, Human-centered AI, Adaptive AI Systems, Cognitive Alignment, Human-AI Interaction, Personalized Knowledge Delivery, AI Evaluation

Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Background

Artificial Intelligence (AI) systems are becoming increasingly embedded in how people seek, engage with, and understand knowledge. From students using tutoring assistants to professionals navigating technical documentation and researchers digesting scientific papers, AI has become a prominent partner in knowledge-intensive tasks. In these contexts, understanding is not limited to passively consuming presented information—it involves identifying what needs to be learned, acquiring new knowledge, and integrating it into one's existing knowledge schema.

Despite recent advancements, current AI systems often fail to support this nuanced and cognitively demanding process. For example, a developer who comes across a new technique in machine learning might input a relevant paper into an AI assistant and ask what the method is and why it matters. While the system may respond with a technically rich summary covering key concepts like "reward model" or "relative log probability", the user may not recall what a reward model is or how PPO works and they need to clarify each point through multiple follow-up questions and piecing together fragmented responses across multiple interactions. Alternatively, the user might attempt to refine their query using strategies like asking for simpler explanations, requesting analogies, or specifying their own professional background (e.g., "Explain like I'm a 5-year-old" or "I'm a software engineer—how does this relate to me?"). Although they may eventually arrive at a satisfactory explanation through such iterative interactions, this process shifts the burden of adaptation onto the user. Understanding is already a cognitively demanding task; Actually, understanding itself is already a cognitively demanding task—it requires actively processing new information and integrating it with prior knowledge. But when AI systems require users to adapt to their limitations, the effort increases. Users need to figure out what makes it challenging to understand the information and identify how to bridge those gaps (e.g., What strategies should they use?). They also should learn how to communicate effectively with the AI systems to acquire the information they need, through often opaque and trial-and-error interactions. This additional burden can lead to confusion, fatigue, and disengagement, especially for users who are unfamiliar with the domain, facing multi-step reasoning tasks, or lacking the prompting expertise to steer the system effectively.

1.2 Aligning AI Systems with Human Cognitive Process

To address the presented challenges in understanding knowledge with AI systems, this thesis draws inspiration from the theory of Human-Computer Symbiosis [141]. Rather than viewing AI systems as passive tools where humans adapt to the systems, to truly augment human cognition, AI systems need to align with how people naturally think, reason, and learn. To achieve such a partnership, this thesis proposes ways to **align AI systems with human cognitive processes**.

This focus on cognitive alignment offers a distinct contrast to dominant paradigms of AI alignment, which often emphasize aligning systems with human intentions, values, or preferences. Such approaches tend to assume well-defined goals and evaluate success based on outcome correctness or agreement with users' preferences. In contrast, this thesis centers on alignment with the process of understanding itself. By shifting the focus from static and general alignment targets to dynamic and cognitive alignment, this thesis reimagines how AI systems can serve as collaborative partners in human learning and reasoning.

1.3 Dual Research Threads

The thesis is organized into two complementary threads: (1) building AI systems that align their outputs with individual users' cognitive processes, and (2) developing methods to interpret and evaluate how well AI systems align with human cognitive processes.

1.3.1 Building AI Systems

Users vary in what they know and how they understand, depending on their goals, tasks, and cognitive states. To support diverse users, this thesis introduces a method to design interactive AI systems that personalize knowledge delivery by aligning their outputs with individual users' cognitive processes. This is achieved by eliciting knowledge states through interaction and generating tailored outputs using structured knowledge representations informed by cognitive models relevant to the target user group. This approach is demonstrated across two contrasting use cases: (1) children asking "why" and "how" questions, and (2) researchers triaging newly recommended papers.

Depending on the differing needs and cognitive capacities of different users, this thesis proposes that interaction design needs to be different. In contexts such as children's learning, explicit interactions—such as asking clarification or diagnostic questions—can help surface knowledge gaps and sustain engagement. In contrast, such strategies may be disruptive or burdensome in cognitively demanding professional tasks. In these cases, adaptive support can be achieved through implicit interaction signals, such as prior engagement history, allowing the system to personalize its responses without requiring additional input.

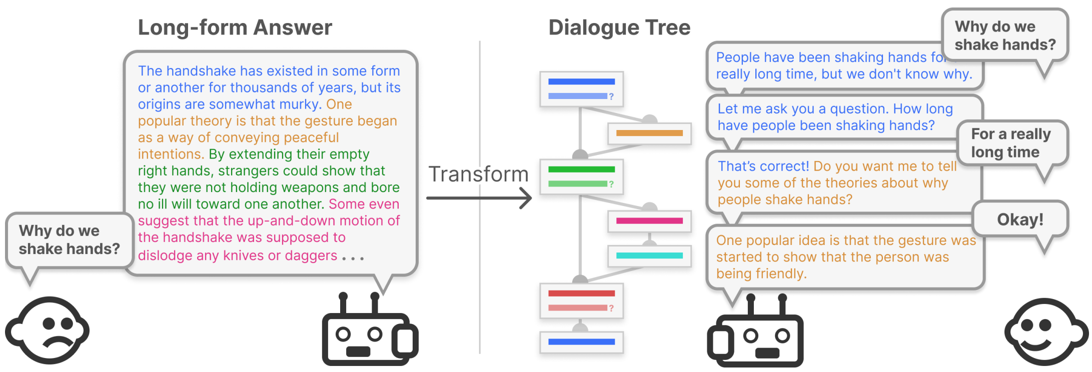

Figure 1.1: An AI-powered pipeline generates dialogue trees that deliver step-by-step explanations, prompting children to engage and assess their understanding.

Transforming Long-form Answers to Interactive Dialogues for Children Existing AI-powered conversational agents (CAs) rely on data not designed with children in mind, making their responses overly complex, insufficiently interactive, and poorly adapted to children's understanding levels. To address these limitations, the DAPIE system [128] was introduced to deliver adaptive, step-by-step explanatory dialogues in response to children's "why" and "how" questions. Informed by extensive literature reviews and consultations with experts in child development, education, and cognitive psychology, the system was designed to transform static, long-form answers into interactive dialogues. This transformation is achieved through a two-step pipeline:

1. **Base Dialogue Tree Construction**: Automatically build a dialogue tree with essential information and optional branches for auxiliary details, determining the high-level flow of the conversation.

2. **Interactive Dialogue Augmentation**: Enrich the dialogue tree by generating targeted questions to elicit children's prior knowledge and interests through simulated conversations between the agent and the child.

A user study with 16 children and their parents further demonstrated that DAPIE improved children's understanding and engagement compared to simpler CAs that deliver information sentence-by-sentence.

Providing Personalized Explanations about New Papers for Researchers With the recent proliferation of published materials, researchers face an even bigger challenge of keeping up with the literature. Researchers sometimes have challenges in making sense of how new papers are relevant to their research context and overlook papers that could be relevant. PaperWeaver [126] was developed to explain new research papers by contextualizing them in relation to a researcher's prior interests, implicitly inferred from previously read or saved papers. Instead of depending solely on global textual similarity or direct inference of comparable aspects by language models, the system constructs aspect-based comparisons that are cognitively meaningful to researchers (e.g., "Paper A and Paper B address a similar *problem* but solve it using different *methods*"). These structured comparison guides an LM to generate personalized explanations that uncover meaningful relationships between new and previously seen papers. Researchers can interact with the system to select papers or comparison structures (e.g., problem-focused or method-focused) to explore and tailor explanations to their needs. In a user study, PaperWeaver enhanced researchers' learning and discovery of connections across papers without increasing cognitive workload compared to a baseline. While personalization can narrow users' focus on what is familiar, this approach balances relevance with exploration, helping users discover previously overlooked but relevant papers.

1.3.2 Evaluating AI Alignment with Human Cognition

In cognitively complex tasks, existing AI models often exhibit low performance and lack well-established evaluation methods to measure their performance. In such settings, users and developers alike face challenges in interpreting what the model is capable of, where it fails, and why. This thesis addresses this gap by proposing an evaluation approach grounded in human cognitive processes. By decomposing complex tasks into fine-grained reasoning steps, the proposed method enables a more structured and interpretable analysis of model behavior. This evaluation helps model developers identify missing capabilities, informs system designers in selecting and integrating models, and allows users to engage with intermediate reasoning steps to support their own understanding.

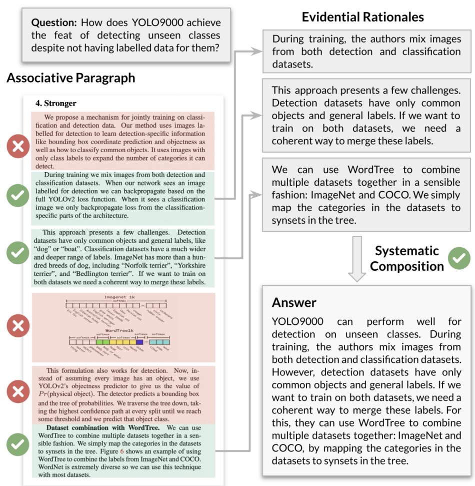

Figure 1.2: An example of QASA: A question posed by the reader/author while reading a paper. Paragraphs are classified for evidence, evidential rationales are written, and these are systematically composed into a comprehensive answer.

QASA: QA Benchmark for Full-stack Reasoning in Scientific Articles While existing question answering (QA) benchmarks have primarily focused on questions requiring shallow or surface-level reasoning, recent work has shown that reading scientific articles naturally elicits a broader range of cognitive demands [129]. Through a think-aloud study and a question analysis grounded in prior literature on human questioning behavior, it is demonstrated that such reading induces diverse question types, spanning from surface questions to deep questions. To address this gap, the QASA benchmark was introduced, consisting of 1,798 novel question-answer pairs that demand full-stack reasoning across scientific articles in AI and ML domains. The accompanying QASA approach decomposes the reasoning process into three stages: associative evidence selection, evidential rationale generation, and systematic answer composition. Empirical evaluations indicate that this full-stack approach significantly outperforms the state-of-the-art InstructGPT model in complex reasoning tasks.

Figure 1.3: Literature review table generation task: (1) synthesize input papers into a table with (2) columns and (3) values, where rows represent input papers.

ArxivDIGESTables: A Benchmark to Evaluate Language Models' Capability of Synthesizing Scientific Articles to Tables The core of the synthesizing process is identifying a schema, a set of important aspects that are useful for comparing and contrasting multiple papers, and identifying values for each paper under those aspects. The results of this process are often presented in the form of literature review tables, whose rows are a set of papers and whose columns are a set of aspects that the papers share. Therefore, to evaluate LLM's capacity to compare multiple papers, ArxivDIGESTables [171] is the evaluation framework that evaluates LM's capacity to compare multiple papers by decomposing this task into two sub-tasks: (1) schema generation, and (2) value generation. ArxivDIGESTables overcomes a lack of high-quality datasets to benchmark table generation by curating and releasing a dataset of 2,228 high-quality literature review tables extracted from ArXiv papers that synthesize a total of 7,542 research papers. Second, to support scalable evaluation of model generations against human-authored reference tables, ArxivDIGESTables introduces DECONTEXTEVAL, an automatic evaluation method that aligns elements of tables with the same underlying aspects despite differing surface forms. In the experiment that evaluated LMs' abilities to reconstruct reference tables, finding this task benefits from additional context to ground the generation (e.g. table captions, in-text references).

1.4 Contributions

This thesis makes two primary technical contributions:

1. Methods for building AI systems that adaptively align their outputs to individual users' cognitive processes, and

2. Evaluation methods that assess how well AI outputs align with the human cognitive process by decomposing tasks into interpretable reasoning steps.

These contributions are instantiated through a series of systems and benchmarks designed for diverse user groups and knowledge tasks. The adaptive systems elicit users' knowledge states and generate personalized interventions, while the evaluation frameworks decompose complex understanding tasks to reveal how AI performance aligns—or misaligns—with human reasoning. Together, these approaches demonstrate how aligning AI systems with human cognition can improve both the effectiveness and interpretability of AI-assisted knowledge consumption.

The contributions are enabled by uniquely combining and extending the following methodological foundations: **human-computer interaction**, to design user-centric human-AI interactions, **natural language processing**, to identify relationships between knowledge and generate cognitively meaningful textual interventions, **cognitive science and human learning theories** to inform system behavior aligned with human cognitive processes, and human-centered AI evaluation, to assess AI output quality from the perspective of alignment with human cognitive standpoints in realistic usage scenarios.

Thesis statement: AI systems that align with human cognitive processes can effectively support adaptive knowledge understanding across diverse users and tasks.

1.5 Thesis Overview

- **Chapter 2** reviews prior work across four areas foundational to this thesis: (1) AI alignment, (2) interactive human-AI systems for knowledge support, (3) knowledge-intensive natural language processing, and (4) cognitive science and human learning theories.
- **Chapter 3** presents *DAPIE*, a system that transforms long-form answers into interactive, adaptive dialogues for children by decomposing explanations into conversational units guided by learning science and child development principles.
- **Chapter 4** introduces *PaperWeaver*, a system that supports researchers in understanding new scientific papers by generating personalized explanations through structured comparisons with previously consumed work using a LLM.
- **Chapter 5** describes *QASA*, a benchmark designed to evaluate how well LMs handle advanced, multi-step questions about scientific articles by decomposing the full reasoning process from evidence selection to answer synthesis.
- **Chapter 6** presents ARXIVDIGESTABLES, a task and dataset for synthesizing multiple scientific papers into structured literature review tables, by curating real-world expert-created tables and designing an automatic evaluation method that compares human-authored tables with LM-authored tables.
- **Chapter 7** discusses the differences between the alignment method with cognitive process and the general existing alignment method, design guidelines for adaptive AI systems, and a cycle of adaptation and evaluation.

Chapter 2. Related Work

This thesis builds upon four key areas of prior research: (1) **AI alignment**, which examines how to align AI systems with human goals, values, and reasoning processes; (2) **human-AI systems** designed to support knowledge work across diverse user groups; (3) **knowledge-intensive NLP tasks**, which enable the generation of cognitively meaningful textual interventions; and (4) **cognitive science and human learning theories**, which inform the design of AI systems that align with human cognitive processes.

2.1 AI Alignment

In recent years, the fields of AI and HCI has increasingly recognized the importance of designing AI systems that center human needs and real-world contexts. Within this movement, AI alignment research has emerged to ensure that AI systems produce intended outcomes while avoiding harmful or unintended side effects [227]. While most existing work on AI alignment has explored how to align systems with human intentions [178, 6], values [218], instructions [13, 148], and preferences [15], this thesis emphasizes aligning AI systems with human cognitive processes. This shift in perspective lays the foundation for exploring how AI can more effectively support human knowledge acquisition and reasoning in complex, real-world tasks. To contextualize the notion of cognitive alignment proposed in this thesis, this section reviews prior work focusing on two key aspects: (1) the types of human values that have been the focus of alignment research, and (2) strategies for aligning AI systems to these values.

2.1.1 Human Values Aligned with AI Systems

Human values relevant to AI alignment have been categorized into multiple sources [213]. At the individual level, alignment may target personal interests [93], cognitive biases [112, 214], or moral judgments [93]. At the social level, values are shaped by group norms, cultural expectations, and fairness considerations [208]. A third category focuses on interaction-level values, such as usability, trust, and collaboration, which emerge during real-time human-AI interaction [66, 257]. While these efforts have contributed significantly to ensuring that AI systems behave in ways that reflect external human expectations, they often focus on outcome-level correctness or agreement. In contrast, this thesis advocates for cognitive alignment—an approach that prioritizes supporting users' understanding processes and adapting to their cognitive states.

2.1.2 Methods for AI Alignment

Various approaches have been proposed for AI Alignment, spanning offline learning, fine-tuning, and an interactive way, that target different stages and mechanisms of system development. Offline learning focuses on integrating broadly accepted human values—such as fairness, safety, and respect for autonomy—into the foundational training process of AI models [**?** 259]. These values are embedded using large-scale value-sensitive datasets or human feedback, often through supervised learning or reinforcement learning from human feedback (RLHF) [178]. This method has been used to align powerful language models like ChatGPT and GPT-4, trains a reward model using human feedback that subsequently guides policy optimization through reinforcement learning techniques. It aligns model outputs with externally expressed preferences or correctness. Offline learning typically occurs during pre-deployment stages, allowing the model to internalize general ethical guidelines and socially desirable behavior before any task-specific adaptation is applied.

In contrast, fine-tuning methods enable targeted alignment by adapting pretrained models to specific communities, domains, or user goals. These methods include group-based training [261], active learning [184], and parameter-efficient adaptation techniques such as adapter modules [224] or mixture-of-experts architectures. Fine-tuning is especially effective when value alignment needs to reflect local norms or evolving user preferences that may not be captured by generic pertaining.

Interactive alignment techniques engage users or stakeholders during system use, allowing them to iteratively influence model behavior through real-time feedback [158, 225], correction, or steering prompts [123, 108]. This interactive layer supports both user control and model adaptivity, and it is particularly useful in high-stakes or dynamic environments where static alignment may not be sufficient. Together, these strategies offer a multi-faceted approach to aligning AI systems with human values at various levels of generality, specificity, and responsiveness.

2.2 Human-AI Systems for Knowledge Support

2.2.1 Conversational Agent for Children

Recently, several studies investigated the experiences of children with conversational agents (CAs) in the form of smart speakers or voice assistants [57, 210, 153]. By observing interactions in natural settings, these studies demonstrated that children interact with CAs on diverse topics [210] and ask questions on various domains (e.g., science, culture, language) [153]. Further, studies showed that children view CAs as friendly, trustworthy, safe, and always available for them [57, 153, 256]—providing evidence on the potential of CAs as conversational partners for children. Despite their potential, research on CAs has revealed several challenges regarding child-CA interactions [107]. For example, CAs' speech recognition frequently misinterprets or fails to understand children's speech [107] which leads to breakdowns in conversation [210]. To resolve these breakdowns, Cheng et al. [35] observed that more capable adults can provide scaffolding strategies, and Xu et al. [256] designed conversational patterns to guide children's responses and prevent breakdowns. However, beyond such breakdowns, children can struggle to parse and understand CAs' responses as they can be long and complex [153, 68] and little research has investigated how to scaffold children's understanding of CAs' responses. DAPIE [128] works as conversational partners for children by introducing a novel approach that automatically creates interactive dialogues that answer children's questions on-the-fly, but with the adequate scaffolding that children necessitate.

2.2.2 Question Answering Systems for Learners

In HCI, researchers have developed QA systems that consider learners' understanding and engagement across various domains such as programming [246], mathematics [29], and factual knowledge [201]. While these systems demonstrate educational benefits, they often require significant manual effort to design diverse and pedagogically meaningful QA interactions. To reduce authoring burdens and enhance the diversity of generated content to cover different levels of prior knowledge of diverse learners, recent work has explored the use of LLMs to automate educational question creation [53, 209]. For instance, pretrained LMs have been used to generate educational assessment questions from existing content [241], and question generation models have been applied to resources such as textbooks [245].

However, many of these approaches focus narrowly on scaling question generation at the model level, without explicitly addressing the pedagogical needs of instructors or the varying understanding levels of learners. In contrast, Promptiverse [125] take a more human-centered approach by transforming annotated knowledge graphs into interactive QA prompts, thereby supporting instructors in generating diverse and relevant questions with less manual effort. ReadingQuizMaker [154] similarly aims to assist instructors by streamlining high-quality question generation. While these systems improve instructor support, they often overlook learner adaptability. VIVID [**?**], for example, enables instructors to collaboratively co-design dialogues with LLMs that are tailored to learners' comprehension levels—supporting interactive, adaptive learning experiences rather than static assessments.

2.2.3 Facilitating Sense-making on Literature for Researchers

Due to the increasing difficulties for scholars to keep up with the rapidly growing scholarly publications, significant research has been devoted to supporting scholars in better understanding the literature. Most prior work in scholar support tools has focused on two stages of this process: **broadening** scholars' reach in discovering relevant papers and helping scholars to **deeply** understand the literature as they read. PaperWeaver [127] focuses on a less explored area that bridges between the two stages: providing lightweight and contextualized sensemaking support when scholars are presented with a set of paper recommendation alerts.

Facilitating Broad Scholarly Exploration and Paper Discovery Significant research has focused on helping scholars explore and discover relevant papers in the literature. For example, SPECTER [43, 215] leveraged citations between documents to train dense vector representations to encode the content-similarity of research papers. The vectors can then be used to power different downstream applications such as search or recommender systems. Prior research has also leveraged alternative similarity signals such as co-citations [186], domains [99], and authorships [102, 187] to broaden the range of recommendations. Recent work has begun to explore facilitating research paper retrieval with a deeper understanding of its semantic content. For example, [31, 100, 87] focused on extracting a *problem-method* schema from paper, which is rooted in cognitive science theories of analogical processing and creativity [72, 71]. Diversifying retrieval based on one aspect (*e.g.,* mechanisms used to tackle a problem) while constraining the similarity on another (*e.g.,* the problems tackled in a paper), can effectively broaden recommendations across different domains and increase creativity in scientific ideation [100].

In addition to paper retrieval, some research has explored how to help users prioritize and understand large volumes of recommendations. For instance, visual overviews and lightweight relevance signals have been used to help scholars make sense of recommended paper sets [33, 222, 105]. These systems provide semantic or social signals to aid triage, such as surfacing personalized social signals based on co-authorship and citations [96] or publication venues and institutions [105]. However, most previous systems either rely on external structured signals that are easy to understand but are divorced from the content of the papers (*e.g.,* "you have cited the authors before" [96]) or latent semantic signals that are based on the content of the papers but difficult to understand (*e.g.,* clusters of papers based on embedding distances[1]). In contrast, recent systems like PaperWeaver [126] draws from literature on schematic processing [73] and leverages recent advancements in LLMs to extract problem-method-finding aspects across papers and generate easy-to-understand compare and contrasting statements that anchor the recommended papers to the user's familiar papers and context.

Facilitating Deeper Scholarly Sensemaking There is a recent thread of research focusing on how to help scholars deeply understand the literature. For example, Relatedly [180] allowed its users to search across related work sections extracted from many published papers and provided support for reading scattered paragraphs with automatically AI-generated subsection title. It highlighted how rich synthesis contained in related work sections can help scholars better understand the landscape of a research field while gaining a deep understanding of how different

[1] https://www.connectedpapers.com/

prior work compare and contrast with one another. CiteSee [32] and CiteRead [188] are two paper reading tools that extracted citing sentences, or *citances* as coined by [170], from other papers to provide in-situ sensemaking support when reading a new paper. Particularly, CiteRead presented incoming citances as margin notes in relevant regions of the paper a user was reading to help them better contextualize the current paper with relevant follow-on work. Another line of research focused on helping users clip citances and organize them into notable threads of research while reading papers to build up a better understanding of a field. For example, Threddy [98] supported dynamic clipping and organization while reading to help preserve users' context of reading when switching between different papers, and Synergi [97] extended this idea to automatically structure a thread-based hierarchy of relevant prior work, contextualized to user-selected citances in a paper. Building on these efforts to reuse related work sections for deeper scholarly sensemaking, PaperWeaver [126] leverages large language models to extract relationships from the related work sections of recommended papers and generates concise contrastive statements that help users anchor unfamiliar papers to familiar ones, making it easier to recognize meaningful connections across the literature.

2.3 Knowledge Intensive Natural Language Processing

Knowledge-intensive natural language processing (NLP) tasks involve reasoning over complex, often domain-specific information that extends beyond the boundaries of a single document. These tasks play a critical role in supporting user needs in real-world scenarios, such as answering open-ended or multi-step questions and synthesizing insights from multiple sources.

2.3.1 Information-Seeking Question Answering

QA has become an increasingly important capability for advancing NLP systems that can understand and respond to human queries. QA datasets serve as crucial benchmarks for developing and evaluating QA systems across different domains and complexity levels.

Long-form Question Answering focuses on generating free-form, explanatory answers to open-ended questions, rather than simply extracting short, fact-based answers from a passage. To support this task, researchers have developed a number of datasets, including SQuAD [193], Natural Questions [120], and ELI5 [62], These datasets vary in complexity and coverage. For example, ELI5 collects open-domain questions and paragraph-level answers from Reddit, supplemented with relevant sentences retrieved from web documents.

Recent research has also examined the internal structure of long-form answers [114] and techniques to improve their factual consistency [159]. Stelmakh et al. [220] argued that many factoid questions in ELI5 are ambiguous and proposed decomposing them into sub-questions. While prior work has focused on answering with informative long-form responses, DAPIE [128] retrieves explanatory passages and transforms them into interactive dialogues. It engages users through follow-up questions that probe prior knowledge, interests, and comprehension, enabling adaptive support aligned with each child's cognitive needs.

Conversational Question Answering extends QA into multi-turn interactions, where a questioner and an answerer engage in a dialogue (e.g., CoQA [194], QuAC [38], QReCC [4]). In CQA, questions beyond the first turn typically depend on conversation history, requiring systems to resolve references and maintain context across multiple exchanges. This mirrors how humans naturally gather information through interconnected questions and answers [194]. The evaluation of conversational QA involves three components: an evidence passage, a human questioner who cannot see the passage, and a model that has access to the passage and must answer questions based on both the passage and conversation history [138]. Although CQA models offer more interactivity than conventional single-turn QA, communication is one-way and lacks considerations on how the answerer can help or engage with the questioner.

Question Answering on Academic Research Papers Several datasets have been proposed for QA on academic research papers including emrQA [181], BioRead [182], and BioMRC [219]. They automatically construct their QA examples by extracting entities and relations as well as structure knowledge resources. Thus these datasets would unlikely reflect real-world scenarios where users have more advanced and open-ended questions [122]. QASPER [49] consists of 5K QA on NLP domain papers. However, most examples in QASPER represent shallow questions focused on completing concepts because the annotators produced the questions after reading only the title and abstract of a provided paper. Additionally, in QASPER, more than 70% of answerable questions consist of short-form answers, such as yes/no and small extractive span. In contrast, QASA [129] collected QAs by asking annotators to read further into main sections, demanding various types of questions based on our studied schema. As a result, the questions in QASA cannot be simply answered with extracting spans form selected evidence paragraph, which urges full-stack reasoning.

2.3.2 Multi-Document Understanding and Synthesis

Schema Generation for Literature Review Synthesizing schemas from research papers has been previously studied in contexts like identifying relations between papers [212, 126], organizing research threads [103], discovering papers for ideation [88, 101], or constructing intermediate scaffolds for better multi-document summarization [211]. These works often assume fixed or sparse schemas, focus on a sub-component of schema generation, or do not evaluate intermediate tables. More closely related to our work, SciDaSynth is an interactive interface for creating "data tables" from a set of papers [239], which infers aspects from users' questions about the papers. However, identifying and articulating good comparison aspects can be nontrivial for users, motivating our aim of automatically inducing salient aspects. Hashimoto et al. [83] explore automated aspect extraction for literature review tables and point out that more specific aspects are useful but hard to generate.

Scientific Table Generation from Multiple Documents Prior work has also released datasets of tables [12, 78]. Bai et al. [12] build a dataset of numeric result tables, while Gupta et al. [78] release 4.4k distantly supervised and 1.5k manually annotated tables with material compositions from papers. Unlike , these datasets do not necessarily link tables to input papers. Multi-document summarization datasets, like Multi-XScience [155] and MSˆ2 [55], are related to table generation but yield sparse tables or use fixed schemas. Finally, there are datasets for other table-related tasks such as table extraction from PDFs [70], table retrieval [67], column annotation [113], table-to-text generation [165], table transformation [34], and table generation [248]. While these datasets either do not focus on scientific tasks or comparing papers, ARXIVDIGESTABLES [**?**] focused on comparing papers in scientific domains.

Query-focused Multi-Document Summarization (qMDS) For qMDS, some datasets in various domains have been proposed, such as QMSum for meeting transcripts [262], Squality for science fiction [237], and AQuaMuSe for wikipedia [117]. The goal of these tasks is to find an answer over multiple documents, which is similar with ours. However, qMDS datasets such as AQuaMuSe and QMSum have the limitation of using noisy and insufficient contexts as multi-documents, since they used automatically-generated passages extracted by lexical matching. To address the issue of insufficiency of dedicated training data, the previous work [16] adopts transfer learning techniques. In comparison to qMDS, our task provides human-annotated evidences aligned with a particular paragraph and answer summaries composed of multi-evidences. Additionally, qMDS focuses on summarizing text without redundancy, while we aim to generate rich long-form answers including multiple rationales.

2.3.3 Adaptive and User-Centered NLP

Recent work in NLP [206] has investigated how to guide LLMs to personalize their generated text based on "user profiles" (*e.g.,* previous data or content that a user has written) which can help to explain, paraphrase, or summarize text into language that is familiar to the user. Building on this thread of research, PaperWeaver [126] investigates an approach to synthesizing papers collected by a scholar into folder descriptions. These descriptions are then leveraged as representations of both the user's interest and prior understanding when generating descriptions to contextually explain new paper recommendations. Regarding scientific literature understanding, ACCoRD[167] defines unfamiliar scientific concepts in terms of different reference concepts by taking advantage of diverse ways a concept is mentioned across the scientific literature. Notably, this work found that users prefer multiple descriptions to a single best description. At a high level, ACCoRD and PaperWeaver both aim to make unfamiliar ideas more accessible by creating bridges upon the user's prior knowledge. However, the approach explored in ACCoRD operates at the unit of concepts, while PaperWeaver aims to help scholars understand the relevance between different papers that may employ many interconnected concepts and other aspects (*e.g.,* problem, method, findings) of academic papers.

2.4 Cognitive Science and Human Learning Theory Foundations

2.4.1 Ausubel's Meaningful Learning Theory

Knowledge acquisition requires assimilating new information into existing knowledge [11]. The overarching idea in Ausubel's theory is that knowledge is hierarchically organized [10, 9, 8]. Based on this idea, he proposed that meaningful learning involves understanding the relationships between concepts and identifying new relations. When meaningful learning is done, knowledge is easily retained and applied, whereas rote learning lets learners just memorize all scattered knowledge [9, 174]. Meaningful learning is achieved when the instructional design considers these hierarchical relationships between prior knowledge and new knowledge. Ausubel also described three learning processes by which new knowledge is assimilated into the existing cognitive structure. The first is **superordinate learning**, where learning of a concept is facilitated by connecting it to many well-acquainted examples. For example, when learning deciduous trees, knowing about instances of deciduous trees, such as maples, oaks, and apple trees, would help understanding. The second is **subordinate learning** which occurs when learners subsume new information to the prior knowledge in a hierarchical manner. This type includes two subtypes of subsumptions which are **correlative subsumption** and **derivative subsumption**. **Correlative subsumption** occurs when learners have to alter or extend their previously learned concept to include the possibility of new information. For example, when learners encounter a tree that has red leaves but only know those with green leaves, then they need to extend the concept of trees to include the cases of red leaves. This process enriches the higher-level concept. **Derivative subsumption** is where new knowledge is an instance or an example of a previously learned concept so learners can leverage existing knowledge to learn the new one. For example, a learner who knows that a tree has a trunk would be able to use that knowledge when learning about a new tree, that the new tree would also have a trunk. The last is **combinatorial learning**, where learners relate previously acquired knowledge to learn new information that is neither more inclusive nor more specific than the previously acquired one. For example, to learn something about pollination in plants, a learner might relate it to the previously acquired knowledge of how fish eggs are fertilized.

Building on this theoretical foundation, this thesis emphasizes the importance of relational structure when integrating new knowledge into existing cognitive frameworks. Accordingly, the systems developed in this thesis elicit individuals' prior knowledge in order to design cognitive interventions that help users understand how new information relates to what they already know. By making these relationships explicit and tailored to each user's knowledge state, the interventions aim to promote more meaningful learning and deeper understanding.

2.4.2 Children's Question Answering Behavior

Children ask many questions to acquire information and develop knowledge about the world [40, 81]. Beyond fact-based questions, children also ask "why" and "how" questions that require explanations about causal relationships or mechanisms [40, 119]. According to developmental psychologists, questions and answers (QAs) help children construct complex causal knowledge [40], so QA-based conversations with more knowledgeable others (e.g., parents, instructors, or CAs) are important for children's development [236]. Studies observed that children prefer answers with satisfactory amounts of information [65, 45] and will ask follow-up questions if not satisfied [163, 64], indicating that question asking is not simply to seek attention. Due to children's well-known need for information, substantial research investigated how to effectively answer children's questions. However, these factors are rarely considered in the design of existing CAs. To address this gap, Lee et al. [128] proposed qualitatively analyzed literature related to answering children's question and consulted experts in children development to propose design guidelines for creating dialogues that effectively answer children's "why" and "how" questions.

2.4.3 Use of Knowledge Representation

Learning involves integrating new information into one's existing knowledge structures [166]. To support this process, structured knowledge representations—such as concept maps, flow diagrams, knowledge graphs, and tree diagrams—have long been employed in both cognitive science and education [2, 198]. These representations serve to externalize and organize knowledge in ways that facilitate comprehension, recall, and higher-order thinking [175, 176]. In the HCI community, several systems have been developed to support learners in constructing or interacting with such representations. For instance, ConceptScape [146] crowdsources concept maps to assist learners in navigating and understanding complex topics. Similarly, TexSketch [221] enables users to generate visual diagrams while reading, helping them synthesize ideas into coherent mental models.

Building on these insights, this thesis leverages structured knowledge representations as a foundation for AI-powered systems that support knowledge understanding. Unlike prior work that often relies on manually created representations, our systems automatically extract and organize information from source texts into task-appropriate formats. This allows the generated outputs—such as explanations, comparisons, or questions—to be pedagogically effective and cognitively aligned with the user's learning goals.

Chapter 3. DAPIE: Transforming Long-form Answers to Interactive Dialogues for Children

Chapters 3 and 5 present a method for building AI systems that interactively adapt to individual cognitive processes across different users and knowledge understanding contexts. This chapter demonstrate how we designed cognitively aligned AI systems to support children in understanding scientific concepts. This chapter has adapted, updated, and rewritten content from a paper at CHI 2023 [128]. All uses of "we", "our", and "us" in this chapter refer to coauthors of the aforementioned papers.

3.1 Motivation and Contributions

Asking "why" and "how" questions is an important theory-building mechanism for children, as answers to those questions can help the child gain a causal understanding of the world [45]. Children as young as three years old can formulate sophisticated questions to resolve gaps in knowledge or perceived inconsistencies [40] on a variety of phenomena: natural, biological, physical, cultural, and social [45]. As children can frequently perceive such gaps or inconsistencies, they require responders who could provide quality answers to their questions in time [231]. Thus, the availability of responders is important for young children. Moreover, knowing that a responder is available [231] can encourage children to ask more questions and improve learning. In terms of availability, conversational agents (CAs), like Alexa or Google Assistant, can offer great value in acting as responders to children's questions. Through CAs, children gain access to a huge amount of information available on the internet without being fluent in reading or writing [68]. Moreover, unlike adult responders, CAs are always available to answer children's questions [228] and have become increasingly common in home settings [173].

However, beyond challenges caused by CA's inaccurate speech-to-text translation, prior work has demonstrated CA's answers are frequently inadequate for children [153, 68]. For example, when asked *"Why do polar bears have white fur?"*, Google Assistant responds with the following: *"Polar bears have white fur so that they can camouflage into their environment. Their coat is so well camouflaged in Arctic environments that it can sometimes pass as a snow drift. Interestingly, the polar bear's coat has no white pigment; in fact, a polar bear's skin is black and its hairs are hollow."*

These types of long responses are challenging for children to understand because they often require the child to possess the prior knowledge needed and to interpret possibly complex reasoning chains [153]. Furthermore, existing CAs provide long responses at once without prompting, which leads to them not being able to identify what a child did not understand or to engage them in a conversation. These challenges stem from the fact that most CAs are powered by computational pipelines designed for adults. For example, information retrieval models, which are commonly used models in pipelines to make CAs, identify relevant passages from the internet and bring these raw long answers to be presented as a response [120, 104, 250]. On the other hand, generative long-form question answering (LFQA) models have also been designed to generate, long-form answers to given questions [62, 136, 137, 169]. However, these models typically provide the answers only focusing on answers' factuality or accuracy with no careful considerations for children. To answer children's "why" and "how" questions by leveraging CA's availability and their ability to connect children to vast amounts of information, we aim to transform long answers into interactive conversations that can enhance children's understanding and engagement.

To identify effective techniques for answering children's questions and for presenting explanations in a form that is comprehensible for children, we propose design guidelines for step-by-step interactive dialogues that scaffold children's understanding on their own "why" and "how" questions. The guidelines present common conversational turns and strategies that can be employed to engage children in conversations, diagnose their understanding, and provide adequate interventions to help them overcome difficulties. To construct these guidelines, we first conducted an iterative inductive analysis on challenges and lessons from prior literature in child development, and conducted consultancy sessions with four child education experts to refine and validate our guidelines.

Applying these guidelines, we propose DAPIE (**D**ialogic **A**nswering via **P**iecemeal **I**nteractive **E**xplanations), a novel system that answers children's questions through step-by-step interactive dialogues that adapt explanations contained in existing long answers. To power this system, we propose an AI-based pipeline that automatically transforms existing long answers into dialogue trees that CAs can perform. The pipeline consists of two main steps that adhere to our design guidelines: (1) decompose and structure the long-form answer into chains of sub-explanations (e.g., tree structure) that provide the information step-by-step; and (2) augment the tree with additional dialogue turns to diagnose the children's understanding and provide adaptive interventions. Through a technical evaluation of the modules in our pipeline, we found that our pipeline outperformed baseline techniques according to measures that correspond to our design guidelines.

To understand whether our interactive explanations improve children's understanding of the information and engagement, we conducted a within-subjects study with 16 participants aged five through seven. They experienced DAPIE and the baseline which provide a sentence at a time using the same source as DAPIE. Our study revealed that children when using our system got a significantly higher score in an immediate assessment and showed a significantly higher level of engagement than when using the baseline system. Children reported that DAPIE was a better teacher, and provided more comprehensible and enjoyable dialogue. We believe that DAPIE is a first step at extending CAs and smart speakers to interactively and adaptively answer children's questions to foster children's curiosity and enhance their understanding about the world.

The contributions of this work are as follows:

- Design guidelines for supporting explanations that answer children's "why" and "how" questions through step-by-step and interactive dialogues.
- DAPIE, a system that serves interactive dialogues through an AI-based pipeline that transforms existing long-form answers into dialogues that follow our guidelines.
- Findings from a user study demonstrating how these generated interactive dialogues can help children's understanding and promote engagement.

3.2 Guidelines for Designing Explanatory Dialogues

In this work, we propose a set of guidelines that outline how to construct explanatory dialogues to answer children's "why" and "how" questions. While research in child development, cognitive psychology, and education has investigated how children ask questions and understand explanations, little work has compiled and organized the findings into design lessons. To address this gap, we present guidelines that describe how to deliver explanations through dialogues that are catered to children's understanding and engagement.

3.2.1 Method

To identify relevant literature, we conducted a keyword-based search on Google Scholar and the ACM Digital Library using the terms: *"question-asking behavior of children"*, *"answering children's questions"*, and *"explanations for children"*. Through several cycles, we expanded our set of search terms by collecting keywords mentioned in sampled literature, and sampling more literature by combining the terms. Details in the Supplementary Materials.

Based on the collected papers, three of the authors conducted iterative coding through inductive analysis to organize the findings and lessons in the papers, and discover recommendations from the data. Any discrepancies in coding were negotiated until mutual agreement was achieved. Based on the analysis, we categorize these recommendations into design guidelines.

To verify and revise our guidelines, we then conducted a design consultancy with four experts in child education. The experts all had majored in child education (one M.S., three Ph.D.) and two also had more than five years of experience teaching children. During the consultancy, the experts were asked to evaluate our framework by revising dialogues that the authors made by applying the framework, and to apply the framework themselves by designing dialogues for given pairs of questions and long-form answers.

By qualitatively analyzing the experts' feedback, we found the following general guides which served to support and extend our guidelines. First, all experts mentioned that it is essential to *"decompose information into smaller steps"* when explaining verbally due to children's limited attention span and working memory. To decide on what information to omit when simplifying explanations, experts suggested considering importance (i.e., what the child needs to know) and acceptance (i.e., what the child can understand). After providing a small chunk of information, three experts recommended asking the child if they understood as each child might understand differently and might need further explanations based on their prior knowledge. They suggested that an explainer can then provide *"further information if the child understands, or provide adjusted explanations if they do not"*. When checking children's understanding, two experts mentioned that true/false or multiple choice questions are used often in practice. More detailed feedback is reflected in our guidelines below.

3.2.2 Guidelines

Based on our analysis of literature and design consultancy with experts, we present guidelines for effectively constructing interactive dialogues that answer "why" and "how" questions from children. The first section in our guideline describes how answers can be decomposed into a step-by-step explanation to scaffold children's reasoning. Beyond reasoning, children can struggle to understand answers as they may lack prior knowledge and can have difficulties in staying engaged. Therefore, the second section describes how, during an explanation, the explainer can interact with the child to promote engagement and check their understanding to provide suitable interventions.

Constructing a Chain of Explanatory Units To present explanations step-by-step, we suggest that explainers construct chains of explanatory units. Specifically, explainers should decompose the explanation into sub-units, identify relevant sub-units and their relationships, and present these units based on their identified relationships. By building relationships between concepts as they learn new concepts, explainers can help children achieve meaningful learning [11].

Decomposing Complex Explanations into Simpler Units When providing complex explanations, we suggest that explainers decompose the explanation into simpler units to help children understand complex concepts while avoiding significant cognitive load [59, 153]. By decomposing, explainers can lower complexity by unpacking the varying factors, entities, and relations contained in an explanation so that the child can process each one independently [80].

Identifying Relevant Units and Relations To aid and guide children's reasoning, explainers should identify and denote the factors in an explanation that are necessary to achieve an understanding. By highlighting relevant factors for a child, the child can identify them as well, focus on them, and reason about the relationships between them [115].

Connecting the Units With the relevant units and relationships identified, explanation should be presented in a cumulative and causal manner. Specifically, an explanation should indicate how one unit in the explanation leads to the subsequent unit, since this guides children to carry out sophisticated reasoning on the relationships between the presented information [203]. Moreover, prior work demonstrated that children show greater curiosity and learning when explanations elaborate on causal connections [106].

Designing Interactive and Understandable Dialogues Beyond constructing chains of explanations, we recommend that explainers expand on these chains to carry out interactive dialogues that support understanding and promote engagement. For this purpose, we suggest a systematic structure for guided dialogues between an explainer and a child, and recommend multiple strategies for the explainer to engage with the child and to adapt explanations to the child's level of understanding. Specifically, we adapted the three key components of effective dialogues with children (i.e., questions, feedback, and scaffolding), which were designed for contexts where adults ask questions to children [217], to our context where a child asks questions to an explainer. Several studies have applied these components to create CAs for preschool-aged children [254, 256].

We propose a dialogue structure where each turn of dialogue is composed of three sub-turns: **Feedback**, **Explanation**, and **Question**. The explainer first provides feedback by building upon the child's utterance. Then, the explainer provides an explanatory unit relevant to the child's answers. The explainer then ends the turn with a question that invites the child to engage in the dialogue. For example, when a child cannot understand an explanation about how bees make honey, the explainer can respond with *"That's alright. Nectar is a sugary liquid that flowers produce. Do you get that?"* where the sentences represent feedback, explanation, and question, respectively. Below, we describe each sub-turn, the various roles that each sub-turns can perform to support children, and strategies through which these roles can be performed.

Feedback involves the explainer verbally commenting on the child's response to the explainer's prompt. For example, the explainer corrects the child's answer (contingency) or praises their attempt to answer through contingency feedback (encouragement). Direct and specific feedback helps children clarify their confusion and increase their engagement [234, 22, 132].

Explanation delivers information after the explainer has provided feedback on the child's response. Explanations can perform two roles: *extension* and *adjustment*.

For *extension*, the explainer elaborates on the topic by providing a new explanatory unit like further details or new pieces of information to deepen the dialogue [110, 27, 46, 40, 163]. *Adjustment* is adapting the explanation to the child's developmental levels, and the cognitive and linguistic demands the child faces during the dialogue with the aim of facilitating child's understanding. For adjustment, explainer gives an explanatory unit that was already provided once, but adapts its content or language.

We propose several strategies for both extension and adjustment sub-turns (Table 3.1). Dialogues can be designed with these strategies to help children understand new information provided in an extension, or, in case the child was unable to understand, the strategies can be applied to create additional alternative explanations to use as adjustments. Local strategies (i.e., Simplifying, Providing examples, Summarizing, Providing analogies, Providing personifications, Representing or demonstrating) are applied individually on explanation sub-turns according to the intervention that the child requires. Global strategies (i.e., Textual simplification, Explicitly mentioning coherence, Global adjustment, Highlighting relevancy) apply to all explanation sub-turns in a dialogue, and act as general support that can benefit all children. Table 3.1 presents local and global strategies and guidelines for each strategy.

Question After providing information to a child, the explainer can ask the question the child to invite them to

Strategies		Guidelines
Local	**Simplifying**	**G1. Use language that matches child's level of comprehension [61].** **G1-1. Change scientific, technical, or formal terminology into simpler language.**
	Providing examples	**G2. Provide various examples that represent new or unfamiliar concepts [200].** **G3. Clearly explain the relationship between the original concept and examples to help generalization [116, 36].** **G4. Consider child's prior knowledge when choosing examples.** **G5. Provide examples with high similarity to the original concept [69].**
	Summarizing	G6. Clearly indicate the core principles of a concept. G7. First provide an immediate and summarized answer to a question before diving into the details [134].
	Providing analogies	G8. Consider the child's unique interest and experiences when choosing a comparison target [234, 75, 142]. G9. Explicitly guide the child to recognize the similarities between the source and target [235, 234]. G10. Choose a target that presents similar entities and relations to those in the source [75].
	Providing personifications	G11. Explain unfamiliar or complex entities and concepts by personifying them or granting them human attributes [89]. G12. Personification is more effective when the entity or concept shares similarities with humans [90, 75].
	Representing or demonstrating	G13. Use representations and demonstrations to illustrate or visualize concepts.
Global	**Textual simplification**	**G14. Apply simplification to all information [153] by considering the average child (local strategy G1-1 simplifies for a specific child).** **G15. (Lexical) Replace difficult terms with simpler ones [109].** **G16. (Syntactical) Simplify the syntactic construction of sentences [52].** **G17. (Length) Use intermediate-length sentence [65].**
	Explicitly mentioning coherence	**G21. Explicitly mention the coherency between turns [116] or use explicit linking language [133, 59] (e.g., "before that", "then").** **G22. For cause-and-effect relationships, explicitly mention how an event leads to the response (e.g., "When all the pieces touch, energy can travel from the battery to the light") [118].**
	Global adjustment	G18. Adapt explanations according to a child's level of prior knowledge [116, 134]. G19. Adapt explanations according to a child's personal experiences. G20. Propagate adjustments made in prior turns through the whole dialogue.
	Highlighting relevancy	G23. Redirect children's attention to the crucial content to help them engage in deeper processing [59, 116].

Table 3.1: Guidelines for local and global strategies that can be used in adjustment and extension sub-turns. Strategies in bold were used in devising our system, DAPIE, presented in Section 3.3.

Role	Description	Strategies
Guiding	Explainers can use *guiding questions* to scaffold children's understanding by helping them narrow down their focus [48] or by leading them to consider other information [162, 255, 64].	**Let children know what information is missing or what information they can ask about [48, 131, 64, 162, 255].**
		Guide children to understand detailed information from a prior turn [48].
Diagnosis	Explainers should *diagnose* children's understanding to provide interventions if they failed to understand [116]. Diagnosis is more effective and reliable if the child is prompted to apply information from the explanation.	**Providing all but one piece of information and asking children to fill-in-the-blank.**
		Ask children to give predictions.
		Ask children to self-explain.
		Ask children to decide whether a fact generalizes across items of varying similarity.
Eliciting	Explainers should *check or ask about children's prior knowledge* to adjust explanations with knowledge that is more familiar to the child. [27].	**Ask children about their knowledge.**
		Ask experience-based questions.

Table 3.2: Guidelines for the different roles that questions can take and specific strategies that explainers can apply to enable these roles. Strategies in bold were used in devising our system, DAPIE, presented in Section 3.3.

participate and engage in the dialogue by answering the question. Our guidelines suggest three roles for questions: *guiding*, *diagnosing understanding*, and *eliciting prior knowledge*. Table 3.2 presents description of the roles and strategies for asking questions.

3.3 DAPIE: Conversational Agent to Support Interactive Dialogues

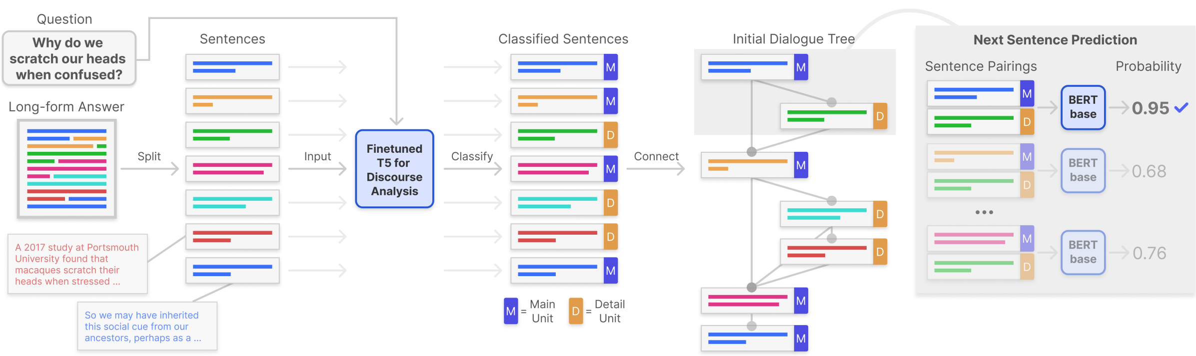

Figure 3.1: Overview of the first step (Section 3.3.1) in our computational pipeline for DAPIE. Starting from an existing long-form answer, it constructs a chain of explanatory units (i.e., an initial dialogue tree). First, it splits a long-form answer into sentences and judges their relevance to a given question, whether it is a main or detail unit, with a T5-based analyzer [251]. Then, the sentences are connected to each other according to their pair-wise relevance score (i.e., next sentence prediction by BERT model [54]) to form a tree structure.
Diagram showing a long-form answer to a question being split into sentences and then each sentence is passed as input to a finetuned T5 model for discourse analysis which classifies each sentence into main or detail units. Then, the classified sentences are connected into an initial dialogue tree by performing next sentence prediction using the BERT base model, where pairs of sentence predicted to have a high probability to be consecutive sentences are connected.

Based on our guidelines, we propose DAPIE, a novel system that automatically transforms existing long-form answers into interactive explanatory dialogues for children. Particularly, our computational pipeline applies our guidelines through state-of-the-art NLP techniques (e.g., large language model (LLM)-based few-shot generation) to structure and augment long-form answers into comprehensible and interactive dialogue trees. Though this, DAPIE can leverage and adapt existing long-form answers on the internet, which are inaccessible to children, to answer children's various 'why" and "how" questions. Our pipeline follows the two main processes from our guidelines: (1) *constructing* chains of explanatory units from the long-form answer; and (2) *designing* an interactive and understandable dialogue by augmenting the chains.

For examples of the final outputs generated by our pipeline, see Figure 3.4 for a dialogue tree.

3.3.1 Constructing Chains of Explanatory Units

In the first phase, our pipeline constructs chains of explanatory units by structuring the explanation in the long-form answer (Fig. 3.1). This phase involves (1) decomposing answers into units, (2) identifying relevant units, and (3) connecting the units into step-by-step explanatory chains.

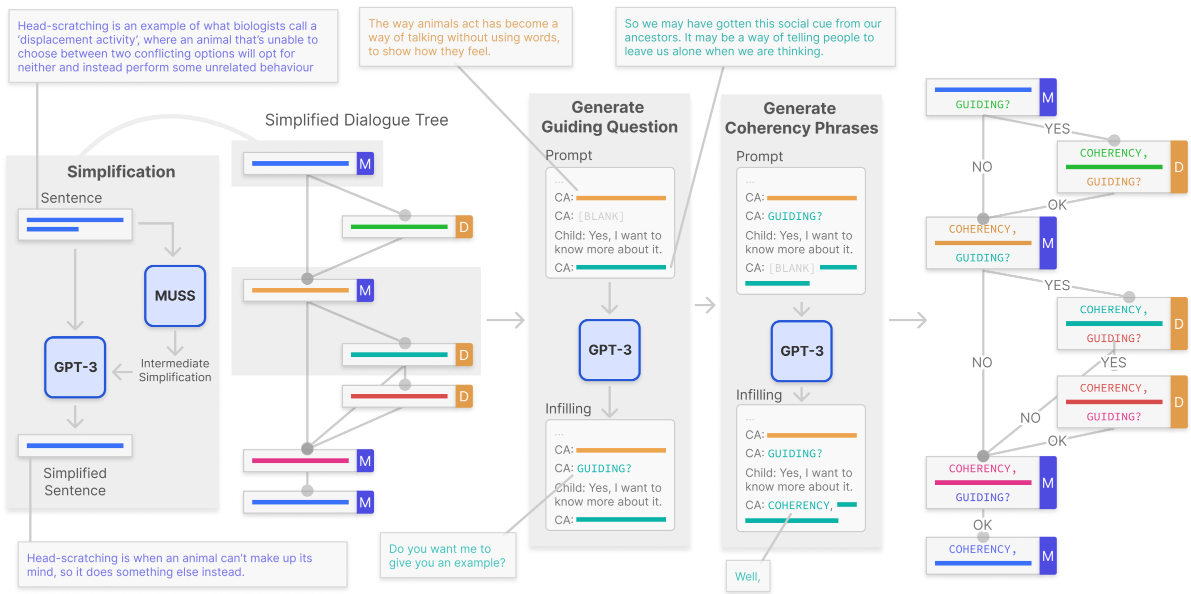

Figure 3.2: In the second step, the pipeline simplifies all the explanations in the dialogue tree, and then integrates guiding prompts and coherency phrases to bridge children's understanding across consecutive explanations. Diagram showing how each sentence block is simplified by simplifying the sentence with the MUSS model first, and then using the original sentence and the MUSS simplification as input to GPT-3 to simplify the sentence once more. The diagram then shows how the dialogue tree with simplified sentences is passed through a component which generates guiding questions between turns connected in the dialogue tree using GPT-3 infilling, and then how the result is passed through a component that generates coherency phrases using the same technique. The end of the diagram shows a dialogue tree where each node consists of a simplified sentence, a guiding question, and a coherency phrase.

Decompose Our pipeline decomposes a long-form answer by splitting it into its constituent sentences: each sentence is an explanatory unit. We assume that each sentence represents one explanatory unit as writers are frequently encouraged to encapsulate one point or thought per sentence. By qualitatively analyzing a sample of 10 QA pairs from the "BBC Science Focus Magazine", two of the authors verified that this assumption generally held for professionally written explanations.

Identify Our guidelines suggest that explainers should identify factors that are relevant to guide children's focus. As existing answers can include auxiliary information that is less relevant to a question [251], we employ the T5-based discourse analyzed by Xu et al. [251] to distinguish between relevant and auxiliary information. By employing this model, our pipeline first classifies the sentences in a long-form answer into their functional roles: "summary", "answer", "example", or "auxiliary information". Then, it assigns those classified as "summary" or "answer" as *main* units (i.e., directly relevant), and those classified to the other two roles as *detail* units (i.e., less relevant).

Connect With the units classified, the pipeline connects the *main* units in the order that they appear in the answer. This chain serves as the main thread of the dialogue tree. As *detail* units are less relevant, we incorporate these as optional extensions for when a child desires more information. As our guideline suggests that explanations should be cumulative, we connect each *detail* unit to the *main* unit that it is most likely to build on. For this, the pipeline uses BERT base model [54] to performs next sentence prediction (NSP) between all *detail* and *main* units. We used an additional model to connect units, instead of relying on their ordering in the answer, as our qualitative analysis of 10 QA samples showed that detail units were not consistently adjacent to their relevant main units. For details, see *"Connecting Explanatory Units"* in the Supplementary Materials.

With these steps, the pipeline produces an initial dialogue tree where each turn consists of an extension explanation sub-turn. Following this tree provides all of the most relevant information to the child's question, with optional branches that provide additional details.

3.3.2 Designing the Interactive and Understandable Dialogue

In the second phase (Fig. 3.2 and 3.3), the pipeline augments the initial dialogue tree by incorporating feedback and questions to interact with a child, and adjustment explanations to scaffold their understanding. We devised the augmentations based on our guidelines. For questions (Table 3.2), the pipeline incorporates (1) guiding questions to lead children to further information (i.e., other main or detail units), (2) diagnosis questions to check children's understanding of main units through fill-in-the-blank questions, and (3) questions that elicit prior knowledge to identify what causes difficulties in understanding. For adjustment explanations, we adopted simplification and provided examples as local strategies (Table 3.1). Analogies and personifications were not used as they are similar to exemplification, but only apply in narrower situations. Additionally, the pipeline applies global strategies (Table 3.1) to simplify and mention coherency in all units.

Our pipeline's goal is to maintain the core information in an explanation, but to incorporate additional turns that are coherent and follow our guidelines. With this goal, we employ an LLM as these models can produce text that coheres with the given context and follows given examples (i.e., few-shot learning). Specifically, we use GPT-3 [25] to extend the *dialog inpainting* technique [47] that generates simulated dialogues where an LLM fills in questions from a "reader" and an "author" answers with sentences from a document. We extend this technique to simulate dialogues where a CA interacts with a child through feedback-explanation-question sub-turns. To employ our extended technique, *turn inpainting*, we designed dialogue templates by imagining dialogues where a CA follows our guidelines to provide questions and adjustment explanations to a child. With the same 10 QA samples analyzed in Section 3.3.1, we performed prompt engineering to iterate on the templates until they produced satisfactory results. As LLMs can generate harmful words (e.g., swear words, vulgar words), our pipeline checks each generation output for such words and re-generates if found—our pipeline never had to re-generate during this work.

(A) Generate Guiding Question	**(B) Generate Coherency Phrase**
CA: [Explanation sub-turn in t_n] CA: [BLANK] Child: Yes, I want to know more about it. CA: [Explanation sub-turn in t_n + 1]	[Turns t_1 to t_{n-1}] CA: [Explanation sub-turn in t_n] CA: [Guiding question from t_n to t_n + 1] Child: Yes, I want to know more about it. CA: [BLANK] [Explanation sub-turn in t_n + 1]
(C) Generate Diagnosis Question	**(D) Generate Elicit Question**
CA: [Explanation sub-turn in t_n] CA: Let me ask you a question. [BLANK] Child: The answer is [Answer for t_n].	CA: [Explanation sub-turn in t_n] CA: [BLANK] Child: Hmm. I don't know. CA: It's okay. [Definition for term in t_n]
(E) Generate Term-based Example	**(F) Generate Clause-based Example**
CA: [Definition for term in t_n] Did you get it? Child: No, I couldn't understand it. CA: Don't worry. Let me give you examples. As you know well, [BLANK]. They are all [Term in t_n].	CA: [Explanation sub-turn in t_n] Did you get it? CA: No, I couldn't understand it. CA: Don't worry. Let me give you an example. [Clause in t_n] is like [BLANK]

Table 3.3: Prompt templates used as input for GPT-3 to produce the functionalities in our pipeline. The few-shot examples that are prepended to each template are available in the Supplementary Materials.

Simplify Due to children's developing language skills, it can be beneficial to simplify all of the explanation sub-turns in the dialogue tree (G14 in Table 3.1). For simplification, our pipeline first uses MUSS [160], a sentence simplification model with controllable attributes for the degree of lexical, syntactic, and length simplification (G15, G16, G17 in Table 3.1). While adequate for syntactic and length simplification, we observed that this model's lack of knowledge led to limited or incorrect lexical simplifications. Thus, our pipeline uses GPT-3, which contains vast language knowledge, to simplify sentences one more time by combining the original sentence, the MUSS simplification, and few-shot examples into an input prompt (T1 in Supplementary Materials).

Integrating Guiding Questions According to our guidelines, guiding questions can help children to engage further with a topic by previewing information to come ("Guiding" in Table 3.2). To generate these questions, our pipeline uses *turn inpainting* by constructing a template (Table 3.3A) with two consecutive explanation turns, t_n and t_{n+1}. With this template and few-shot examples as input, the model fills in a guiding question, in place of `[BLANK]`, that asks the child if they want to learn about the second turn. In the dialogue tree, the CA asks the question and moves to the next turn when the child accepts. If the next turn can be a main or detail unit, the CA asks the guiding question to the detail unit. With the guiding questions, the pipeline also uses *turn inpainting* to generate phrases that explicitly describe the coherency between turns and how an event leads to the response (global strategy G21, G22 in Table 3.1). For details, see *"Creating Coherency Phrases"* in Supplementary Materials.

Designing Diagnosis Questions As our guideline suggests, it is crucial that an explainer checks whether a child understood an explanation and, if they did not, to provide suitable adjustments ("Diagnosis" in Table 3.2). We chose to generate fill-in-the-blank questions as the other strategies require free-form responses that are difficult to verify with existing techniques. To generate these questions, the pipeline identifies two potential difficulties in the explanations to use as the "blank": unfamiliar terms and complex cause-effect relationships. According to surveyed literature, these two were common challenges in children's understanding [75, 61], and are core factors for answering "why" and "how" questions (i.e., prior knowledge and mechanistic reasoning). With GPT-3, our pipeline identifies these difficulties, and then extracts a word/phrase from the difficulty to use as the answer for the diagnosis question For details, see *"Identifying Difficulties and Correct Answers"* in Supplementary Materials.

With the difficulties and answers extracted, the pipeline generates a diagnosis question using GPT-3 with a template (Table 3.3C) where the CA asks a [BLANK] question and the child gives the answer extracted by the pipeline (prompt T6 in Supplementary Materials). Finally, to narrow down children's possible answers, we use GPT-3 to create alternative but wrong answers (prompt T7 in Supplementary Materials).

In the dialogue tree, diagnosis questions are asked after the main turns. For correct answers, the dialogue provides contingency feedback ("Feedback" in Sec. 3.2.2) like "That's correct!"and asks the guiding question. For incorrect answers, the dialogue provides feedback (i.e., "Hmm, I don't think so") and moves to adjustment turns according to the difficulty (i.e., term or cause-effect). We provide adjustment turns after the main turns as our consulted experts suggested that children should first be provided with information relevant to their question and then, if needed, provided with support. They explained that this retains the engagement of children who can understand, while guaranteeing fallback support for those who cannot.

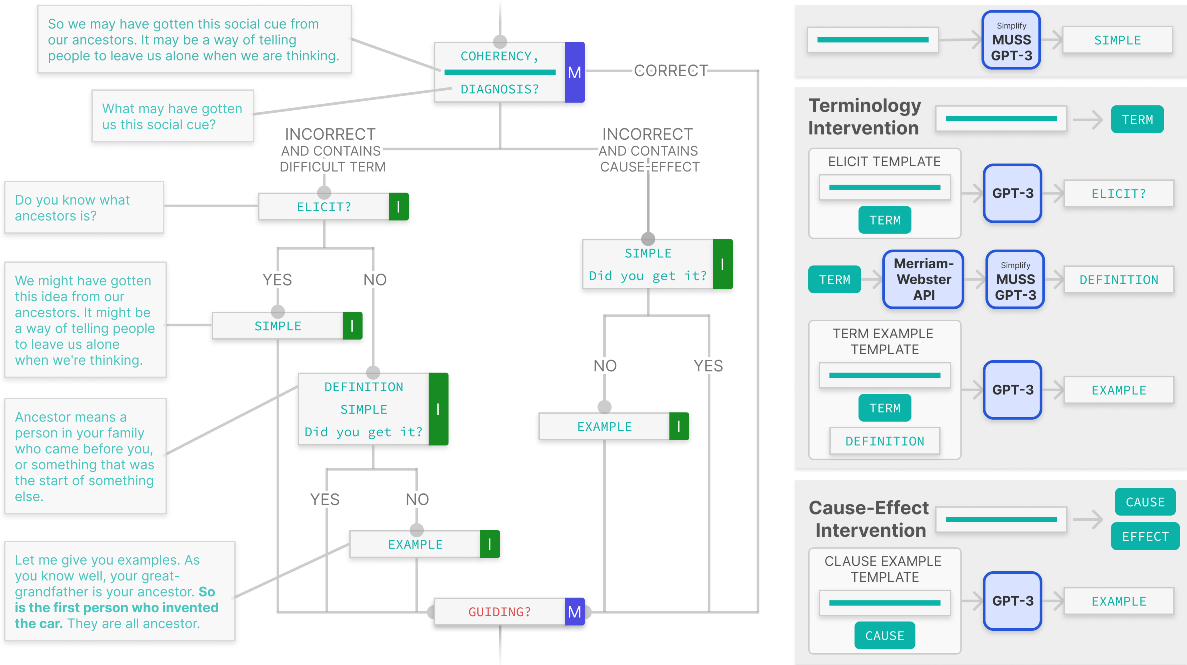

Figure 3.3: In the final step, the pipeline constructs adjustment sub-trees that scaffold children's understanding of (1) difficult terms, or (2) cause-effect relationships. The types and number of adjustments provided depend on children's answers to intermediate questions. After receiving the corresponding adjustments, the dialogue proceeds to the guiding questions for the next turn in the dialogue tree.
The diagram shows, on the left, a dialogue tree with augmented turns which branches out from the diagnosis question depending on whether the original explanation contained a difficult term or a cause-effect relationship. The diagrams shows how the dialogue tree provides different types of adjustment turns to the child based on this.

Adjustment Turns for Difficult Terms To verify whether the child failed to understand because they did not know the difficult term, the pipeline generates a question to elicit prior knowledge ("Eliciting prior knowledge" in Table 3.2). The pipeline constructs a template (Table 3.3D) where the CA asks a [BLANK] question, the child responds that they do not know, and so the CA provides a definition for the difficult term. To visualize how the adjustment turns are connected in the dialogue tree, refer to Figure 3.3.

Simplification Turn. If the child answers that they know the term, the pipeline offers a more simplified explanation (G1 in Table 3.1). For this, the explanation in the main turn is simplified again using the same simplification method as in Section 3.3.2.

Definition Turn. If the child does not know the term, the pipeline provides an extension explanation for the term. As LLMs can hallucinate [161] (i.e., generate false information), we retrieve definitions from a verified source, Merriam-Webster API[1], instead of generating them. After retrieving, our pipeline simplifies the definitions using our simplification method. The definition turn provides this definition and a simple diagnosis question asking the child if they understood or not. A simpler diagnosis is used to not exhaust children with frequent quizzing.

Term Exemplification Turn. If the child could not understand the definition, the CA should provide an additional adjustment to help them understand. Based on our guidelines, the pipeline generates various examples to illustrate the unfamiliar term by creating a template (Table 3.3E) where the CA provides `[BLANK]` examples of the unfamiliar term and explicitly indicates that they are all examples of the term (G2 and G3 in Table 3.1). Also, the prompt includes few-shot examples illustrating effective exemplification—i.e., familiar to children and have high similarity to the original concept (G4 and G5 in Table 3.1) We identified that, beyond questions and coherency phrases, *turn inpainting* could also produce context-relevant adjustment explanations (e.g., examples) based on a simulated dialogue.

Adjustment Turns for Cause-Effect Relationships Causal reasoning can be challenging for children due to limited prior knowledge on various cause-effect relationships [75]. To help children understand cause-effect relationships, the pipeline constructs two consecutive adjustment turns: a simplification turn, and a clause exemplification turn. The simplification turn is the same as that in the adjustment turns for difficult terms. To view how the adjustment turns are connected, refer to Figure 3.3.

Clause Exemplification Turn. When simplification is insufficient, the pipeline creates an example of another similar cause-effect relationship. We generated examples based on the causes as we observed that generating from the effects lead to broader and more unrelated examples (G5 in Table 3.1). The pipeline constructs a dialogue template (Table 3.3F) where the CA compares the cause in the explanation to another `[BLANK]` phenomena. While this dialogue template was designed to create examples, the pipeline occasionally creates analogies—possibly due to the presence of commonly used analogies in GPT-3's training data.

3.3.3 Interface

We developed DAPIE (Fig. 4.2), a CA that serves the interactive dialogues generated by our pipeline. DAPIE interacts with the child by first saying an utterance, and then reading options to which the child can respond. The CA, implemented as a web-based interface, was designed based on prior work and insights from the expert consultancy.

For each turn of a dialogue, DAPIE first says their utterance and also displays the text. The speaking rate was set lower than default to prevent overloading the child, and we show text as an additional modality to benefit children who are able to read. After the agent finishes its utterance, it reads out the options the child can respond with one-by-one while revealing them below the utterance. DAPIE pauses briefly after each time it speaks to let the child think and process what was just said.

After all the options have been read, the child can respond by clicking on one of the options. Although we could have designed DAPIE as a fully voice-based interface, automatic speech recognition is still limited and error-prone—specifically with children's speech [107]. Since our goal is to help children understand and engage with explanations, we designed our interface so that the child can accurately express their intent through clicks—protecting their understanding from being hindered by speech recognition errors. Furthermore, experts suggested limiting the child's response options, instead of allowing free-form responses, as children might already

[1] https://dictionaryapi.com/

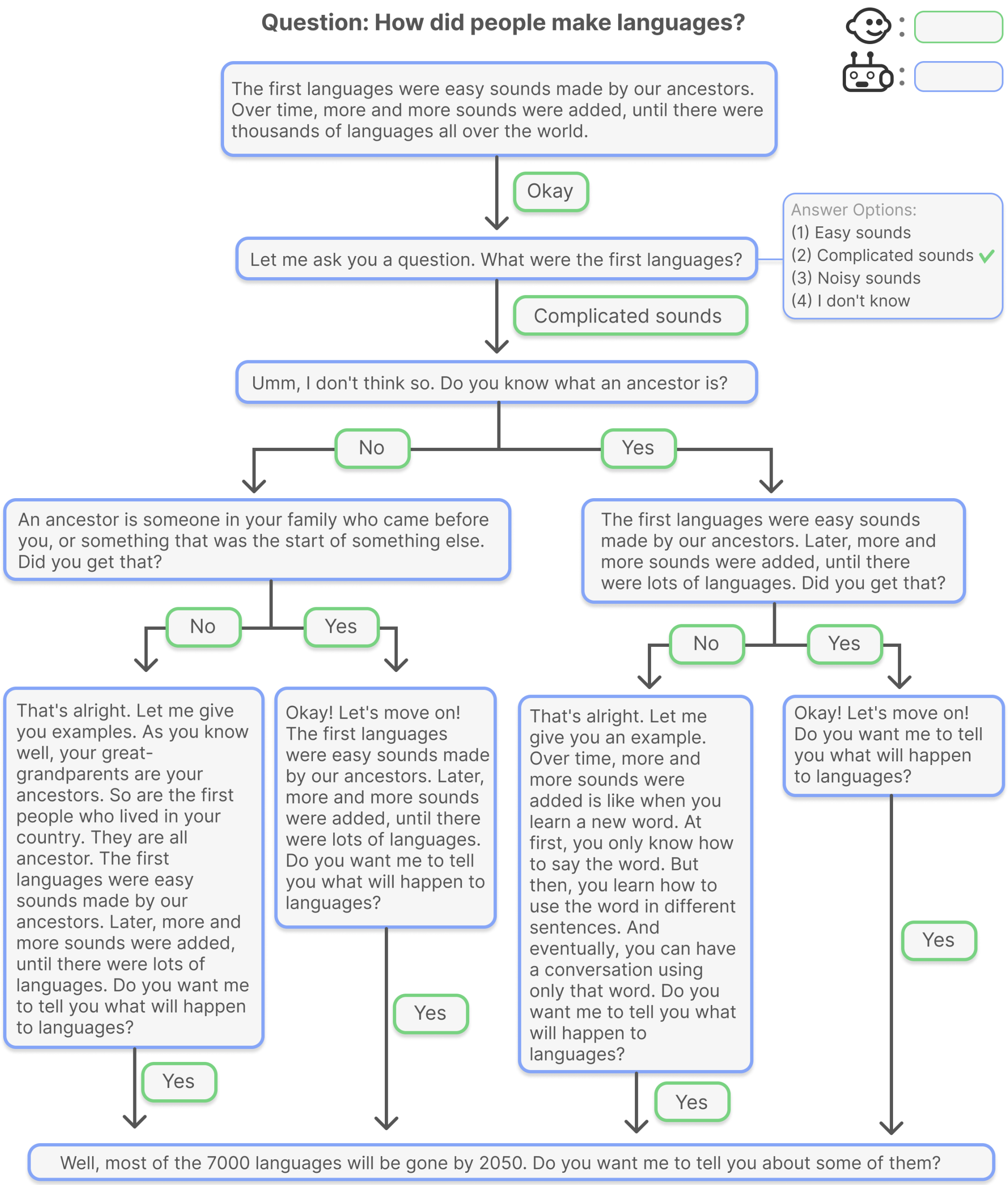

Figure 3.4: Subtree taken from the dialogue tree generated by our computational pipeline from the answer to the question *"How did people make languages?"* The dialogue tree shows how the CA can provide an explanation, diagnose the child's understanding, and the provide interventions (e.g., definition, example) to help them if they have difficulties understanding.
Subtree shows the various explanations that are provided by DAPIE to the child and the different types of responses that the child can provide. For example, one of the branches in the subtree show how the conversational agent explains that an ancestor is someone in your family that came before the child and asks the child if they got that, and the child has the option to respond "no" or "yes".

be cognitively burdened from understanding the explanations.

When the child clicks on an option, the text is selected and it is read to them again. This allows the children to re-listen to any option (or even the utterance) in case they may have forgotten or failed to hear what was first said. While an option is selected, the child can click on the option again to choose it and the dialog will then proceed to

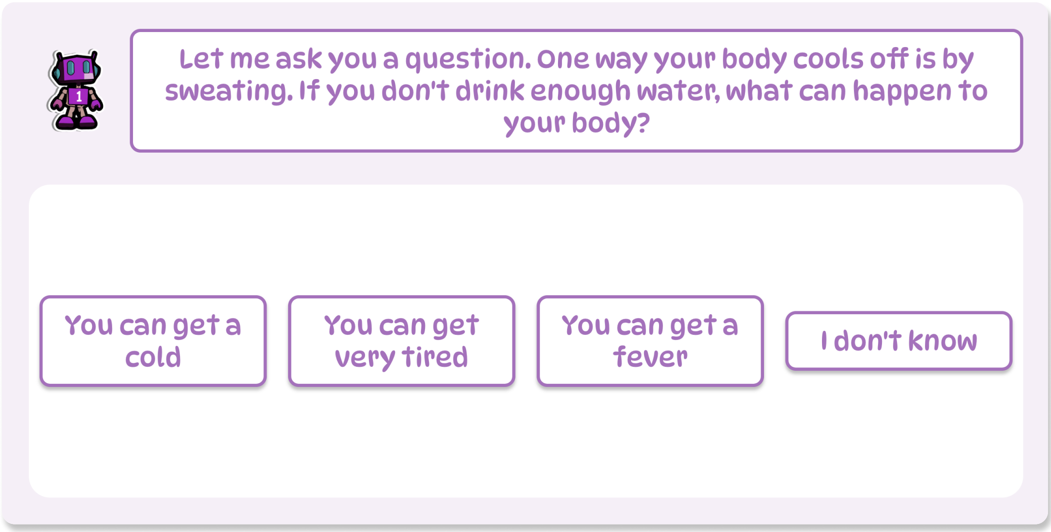

Figure 3.5: In the interface for DAPIE, the CA's explanation is shown on the panel above and spoken through text-to-speech. After speaking the explanation, the interface presents the child's response options in the panel below.
Interface of DAPIE shows the diagram of a robot on the top-left and next to it, on the right, the explanation provided by DAPIE. Below these items, the interface shows four different options that the child can respond with as buttons.

the next turn corresponding to that choice. This interaction repeats until the child reaches the end of the dialogue, after which they return to the main menu.

3.4 Technical Evaluation on LLM-generated Dialogue

We evaluated DAPIE through a pipeline evaluation and a user study (Section 3.5). The pipeline evaluation was conducted to verify the performance of each step in the pipeline, and the user study was to evaluate whether DAPIE's interactive dialogues, as a whole, help children's understanding and engagement. To validate our pipeline, we conducted human evaluations on the five main sub-modules used to generate the dialogues: simplification, exemplification, guiding question, diagnosis question, and elicit question. We focused on evaluating the text generation modules for questioning and adjusting (Section 3.3.2), since the modules for chain construction (Section 3.3.1) use off-the-shelf models. We provide a comprehensive analysis of each sub-module's performance compared to their baselines, and an in-depth post-analysis according to characteristics of the input question-answer pairs (e.g., source type, domain, and question type) to understand whether our pipeline is generalizable.

3.4.1 Method

We compared the quality of generated sentences from the five sub-modules (i.e., simplification, exemplification, guiding question, diagnosis question, and elicit question) to those from corresponding baseline models. For the evaluation metrics, we saw that existing metrics for evaluating generated text depend on adult-centric datasets so they cannot adequately evaluate text generated for children. Also, several of our modules perform novel tasks for which evaluation criteria do not exist. Therefore, by referring to our guidelines and commonly used measures in NLP, we defined different evaluation questions for each sub-module (Table **??**).

Test Data Collection: To evaluate whether our pipeline can be applied to various types of explanations, we collected test data (N=32) from multiple sources and domains. For sources, we chose an expert-generated source, BBC Science Focus Magazine, and a user-generated source, ELI5 dataset [62]. From each source, we collected data that corresponds to four domains: natural phenomena, biology, physics, and cultural and social conventions. Then, for each source and domain combination (total of 8), we collected two QA pairs for "why" questions and two for "how" questions. This totaled to 32 QA pairs (and more details are available in the Supplementary Materials).

Baselines: For the baselines, we selected state-of-the-art models for existing tasks and, for our new tasks, we used an LLM as it could perform the task from given instructions. For exemplification, we adopted GPT-3 with a zero-shot prompt (T11 in the Supplementary Materials) as a baseline since there is no appropriate specialized model for this task—Wang et al.'s [238] model is not open-sourced and is retrieval-based while our task is generative. For simplification, we adopted the state-of-the-art model MUSS [160]. For questioning, we trained a dialogue inpainting model following Dai et al. [47] using four datasets: QuAC [38], QReCC [4], DailyDialog [140], and Taskmaster [26]. Specifically, we finetuned the T5-large model [189] for three epochs with a learning rate of $3e^{-4}$.

Procedure: Inspired by the ACUTE-EVAL method [139] which is widely used for comparing generated dialogues, we showed a human evaluator two *responses*, one from our sub-module and one from the baseline. Both responses were generated from the same input *context*. For simplification and exemplification, we provided the original sentence and a context sentence as input context. For question generation, we provided a multi-turn dialogue including the `[BLANK]` that the model filled in. The human evaluators looked at the input context and the two generated *responses* side-by-side. The evaluator was also shown the measuring questions that corresponded to the sub-module and, for each question, were asked to make a choice between the two responses or to choose "tie". For each data point, we assigned three evaluators to collect three trials of such pairwise judgments, and used majority voting to designate whether our sub-module performed better, the baseline did, or whether it was a tie (e.g., the majority chose "tie"). We hired crowd workers as evaluators as the experts from our consultancy mentioned

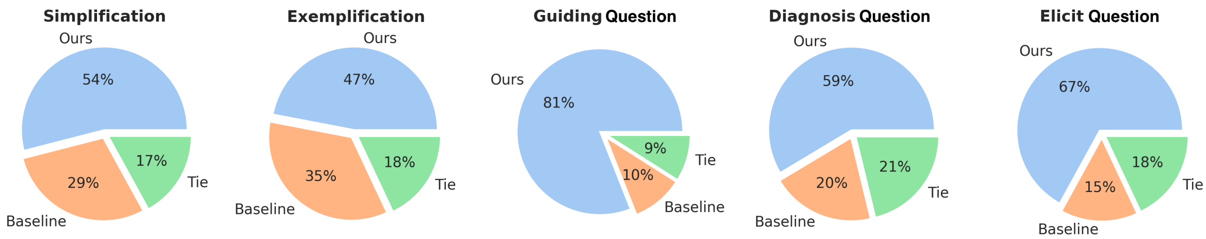

Figure 3.6: Overall human evaluation results for the five sub-modules. Outputs from our pipeline were assessed favorably compared to those from the baselines for all of the sub-modules.

Pie charts showing the proportion of evaluations for each sub-module. For simplification, 54% voted for our pipeline. For exemplification, 47% voted for our pipeline. For guiding question, 81% voted for our pipeline. For diagnosis question, 59% voted for our pipeline. For elicit question, 67% voted for our pipeline.

that even the general public can evaluate how helpful and easy it would be for a child to understand given content. We recruited crowd workers from Amazon Mechanical Turk who were in the US and had task approval rates higher than 98%. Each worker evaluated all five sub-modules (five pairs per sub-module) and answered one gold standard question. Our task took around 30 minutes and we paid workers $6 for their time. The final inter-annotator agreement was rated as fair (Fleiss's kappa=0.338).

3.4.2 Results

In short, the results generated by our pipeline were assessed favorably when compared to their corresponding baselines across all the criteria (Fig. 3.6). As seen in Table 3.5, the biggest difference between our sub-modules and

	Simplification			Exemplification				Guiding Question			
Measuring Questions	**QS1**	**QS2**	**QS3**	**QC1**	**QC2**	**QC3**	**QC4**	**QG1**	**QG2**	**QG3**	**QG4**
Ours	**52%**	**54%**	**57%**	**49%**	**46%**	**42%**	**49%**	**81%**	**80%**	**88%**	**76%**
Baseline	34%	32%	21%	40%	38%	32%	31%	9%	11%	6%	16%
Tie	12%	14%	22%	11%	16%	26%	20%	10%	9%	6%	8%

	Diagnosis Question				Elicit Question				
Measuring Questions	**QD1**	**QD2**	**QD3**	**QD4**	**QE1**	**QE2**	**QE3**	**QE4**	**QE5**
Ours	**55%**	**70%**	**38%**	**69%**	**64%**	**83%**	**72%**	**53%**	**64%**
Baseline	27%	21%	17%	17%	14%	9%	21%	17%	14%
Tie	18%	9%	45%	14%	22%	8%	6%	30%	22%

Table 3.5: Human evaluation results on five sub-modules shows, for each measurement question, the percentage of workers that preferred our pipeline's outputs, the baseline's outputs, or chose that it was a tie. For all five sub-modules, our pipeline outperformed the baselines.

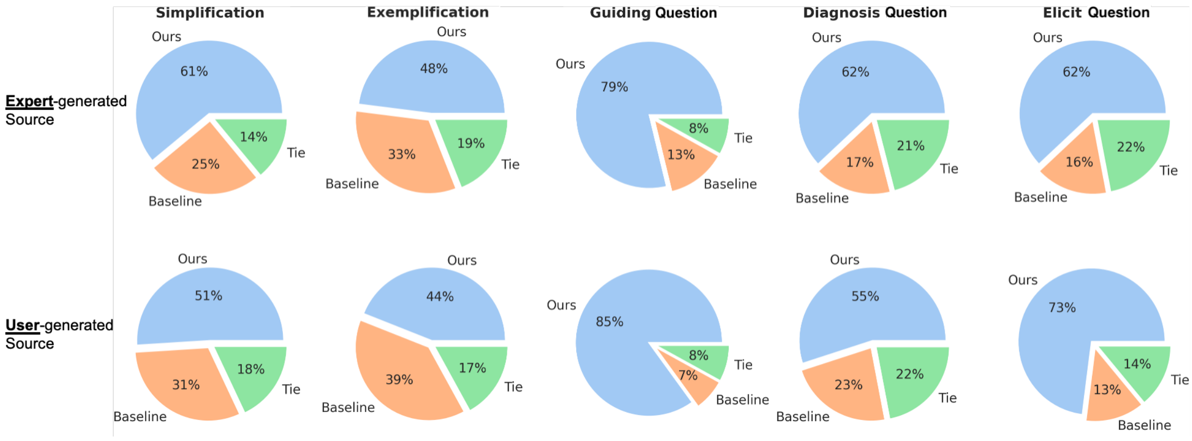

Figure 3.7: Human evaluation results on the five sub-modules according to the type of the source (i.e., expert-generated or user-generated). For all five sub-modules, our pipeline outperformed the baselines in the both type of sources.

Pie graphs showing the proportion of evaluations for each submodule according to whether the source was expert-generated or user-generated. In summary, our pipeline was voted more highly for all sub-modules across the two types of sources.

the baselines was for the **guiding questions** (QG, 81% vs 10%), while the smallest difference was for **example** generation (QC, 47% vs. 35%). Evaluators judged that our **simplification** module generated text with easier words (QS1, 52% vs. 34%) and simpler sentence structures (QS2, 54% vs. 32%) while preserving the core information in the original text. The differences between our **examples** and the baseline's were relatively small in terms of helpfulness (QC1, 49% vs. 40%) and familiarity (QC2, 46% vs. 38%) since both generally composed examples with easier words. Regarding the **guiding (QG), diagnosis (QD), and elicit (QE) questions**, there were apparent differences between *turn inpainting* and the baseline across all criteria. We presume that the substantial differences are derived from whether the models explicitly considered that the recipient of the question is a child or not. The baseline failed to generate adequate outputs, since the model had likely never seen dialogues with children at training time. Specifically, evaluators rated all the questions from our *turn inpainting approach* to be more understandable for a child (QG1, QD1, QE1) and closer to what a teacher would ask (QG2, QD2, QE2). We found that the difference for grammatical errors is significant for guiding questions (QG3) while this gap is reduced for diagnosis questions (QD3) since these are usually simpler, e.g., "Did you get that?" or "What's the problem?"

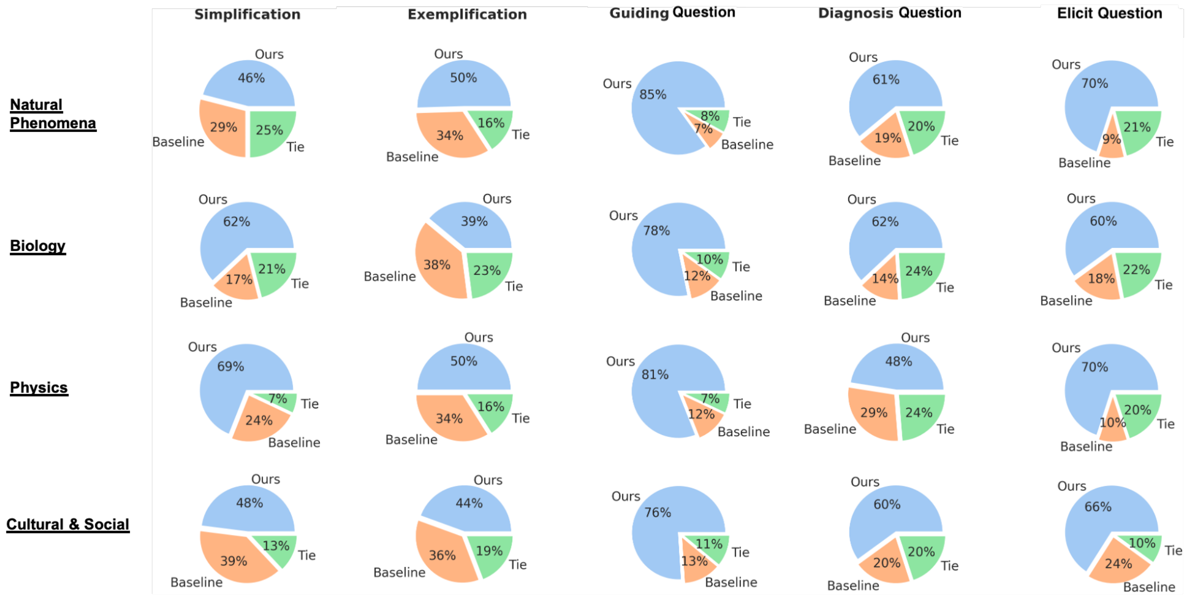

Figure 3.8: Human evaluation results on five sub-modules according to the domain. For all five sub-modules, ours outperformed the baselines in all the domains.
Pie graphs showing the proportion of evaluations for each submodule according to the domain of the question. In summary, our pipeline was voted more highly for all sub-modules across all domains except for exemplification in Biology where our pipeline was voted for by 39% of evaluators and the baseline by 38% of evaluators.

For the QA pairs from both expert-generated and user-generated sources, our sub-modules produced better generations than the baseline did (Fig. 3.7 and 3.8). In terms of questioning, the differences between our sub-modules and the baselines were irrespective of source and domain. For simplification and exemplification, the difference between the models was greater in the expert-generated sources, which are written more formally. When analyzing by domain, the differences in simplification and exemplification seem to be greater in the Physics domain, which contains the most scientific content. Furthermore, we observed that the relatively poor exemplification performance in the Biology domain (39% vs. 38%) was due to our sub-module significantly underperforming QC3 (27% vs. 42%). This implies that our exemplification outputs often included text duplicated from the given context, but the outputs still contained valid examples as they were assessed better for the other criteria (i.e., QC1, QC2, and QC4). Nevertheless, we conclude that our sub-modules overall perform better than the baselines regardless of source and domain, but show greater differences in formal or scientific text for simplification and exemplification.

Additionally, as LLMs can generate hallucinations (i.e., non-factual or nonsensical information) [161] or errors (e.g., grammatically incorrect or incoherent text), we qualitatively analysed 20 sample generations for each of our sub-modules or baselines. For simplification, our sub-module and baseline produced hallucinations for the same 10% of the inputs and, for exemplification, our sub-module produced fewer hallucinations (20%) than the baseline (30%). Overall, both our sub-modules and baselines tended to hallucinate minor superfluous or incorrect details for the same inputs, but our sub-modules were generally more robust to hallucinations. Also, while we recognize that exemplification had a relatively higher chance of hallucinations, the potential negative impact of these is the least significant as examples do not modify the original explanation and are the last resort support in our dialogues—only provided when children fail to understand after several adjustments. For question generation, all of our sub-modules produced few errors (<10%) while the baselines frequently produced questions that were incoherent or grammatically incorrect.

3.5 User Study

We conducted a controlled study to investigate whether interactive conversations from DAPIE improve children's understanding of concepts (RQ1) and increases engagement (RQ2) compared to a simpler CA that provides the same information sentence-by-sentence.

3.5.1 Participants and Apparatus

We recruited 16 participants (5 female, 11 male) aged five through seven through snowball sampling and by posting advertisements on online forums (e.g., Twitter, Reddit, and the online communities of several colleges). Before the study, we assessed children's English language proficiency using a computer-based assessment (i.e., Quick Interactive Language Screener [135]) to ensure that participants can understand the questions in the conversations, assessment, and usability survey. Table 3.6 summarizes the participants' demographic information.

For explanatory material, we selected question and answer pairs for four domains: Natural Phenomena, Biology, Physics, and Cultural and Social Science. We chose these domains as they are commonly asked by children [153] and are topically diverse helps us test generalizability. For each domain, we selected two QA pairs, one for each question type (i.e., "why" and "how"). Participants were assigned a total of four questions where they saw one question per domain and two questions per question type.

We compared DAPIE's interface to a baseline interface with the same UI. However, instead of providing interactive conversations, the baseline provided the information in the original answer by presenting one sentence at a time. After each sentence, the interface showed "Okay" as the only option the user could click to respond with. Unlike existing real-world voice-based CAs which provide lengthy explanations in a single turn, this baseline provides greater interactivity and allows the user to consume the information step-by-step. Thus, we believe this is fair baseline since it provides a higher level of interactivity than what is supported in existing voice-based CAs. Each session lasted about 60 minutes and participants were compensated with $50.

	Sample Ratio
Female	31.25%
Age	
5-year-old	12.50% (N = 2)
6-year-old	31.25% (N = 5)
7-year-old	56.25% (N = 9)
Predominant Home Language	
English	62.50% (N = 10)
Other (Korean)	37.50% (N = 6)
Race/Ethnicity	
Asian	50.00% (N = 8)
White	18.75% (N = 3)
Black	12.50% (N = 2)
Other	18.75% (N = 3)
Parents' Education	
Bachelor's degree or higher	81.25% (N = 13)
Other	18.75% (N = 3)
Usage of CA	
Daily or Weekly	25.00% (N = 4)
Monthly	25.00% (N = 4)
Rarely	50.00% (N = 8)
N	16

Table 3.6: Demographics of the participants in our study.

3.5.2 Study Procedure

The study was conducted remotely. Children participated in the study from their homes and communicated with the researcher via a video conferencing tool[2]. Children first followed a simple tutorial dialogue where they were introduced to DAPIE and the baseline, and learned to use the systems by answering a few simple questions like "Are you ready?" Then, the children went through the explanations for four questions. They interacted with each condition for half of the questions—the order was counterbalanced. Each question was from a different domain (i.e., Natural Science, Biology, Physics, and Social Science). As learning Natural Science can affect understanding of Biology and vice versa, we grouped the domains such that participants saw questions from Natural Science and Biology in one condition, and Physics and Social Science in the other—limiting learning effects across conditions. Additionally, participants saw one "why" question and one "how" question in each condition—the order was counterbalanced. After each question, we conducted an assessment that asked about specific knowledge in the explanation. After each condition, we conducted usability surveys and semi-structured interviews. After the children completed the study, we conducted semi-structured interviews with their parents regarding their child's experience with the CAs. The child's screen and their camera video were recorded. The procedure was pre-approved by the IRB of the authors' institution.

3.5.3 Measures

For measures, we evaluated the participants' understanding of the information in the dialogues, their engagement with the dialogue, and their perceived usability of the interface. We also qualitatively analyzed the interview data.

Immediate Assessment To assess children's understanding of concepts from the dialogue, we developed three questions for each dialogue. The questions assessed children's recall and understanding of facts introduced in the dialogue. These questions were different from the explanations embedded in the dialogue and did not overlap with the diagnosis questions provided in the dialogues. We designed these questions by consulting experts on children's learning and language development. All the assessment items are included in the Supplementary Materials.

For all of the questions, we first asked children open-ended questions and allowed them to freely formulate their answer. If they were unable to provide the correct answer, we provided them with two answer options to choose from. Children received a score of 2 if they answered correctly without options, a score of 1 if they required options to answer correctly, and a score of 0 if they could not answer correctly.

Engagement The evaluation of engagement was based on coders' assessments of each child's engagement in terms of three behaviors: eye gaze, verbal comments, and nonverbal comments. Eye gaze considered instances when participants stared at places other than the screen where the explanation was presented, which has been used as a negative indicator of engagement in children's book reading [95] and video watching [254]. On the other hand, verbal and nonverbal comments were considered as positive indicators. Verbal comments considered when a participant would verbally answer the CA's question, ask a question about on-topic information, or react to agent's response. These comments could be either to their parent or to the agent, although the agent could not understand these. Nonverbal comments included pointing at the screen, moving the cursor around, or clicking the interface to re-play the CA's explanation. Two coders observed the study sessions and, for each dialogue turn, recorded whether the turn included these engagement behaviors. Then, for each of the behaviors, we calculated

[2] https://zoom.us/

$(number\_of\_turns\_with\_the\_behavior)/(total\_number\_of\_turns) \times 100$. The IRR calculated by Intraclass Correlation for the two coders was 0.73, which is considered as substantial [195].

Usability For usability, we used a survey to elicit children's enjoyment on the usage experience, and their perceived trust towards DAPIE and the baseline. We adapted the four questions from the Giggle gauge [56] to measure enjoyment, and adapted two questions from Richards and Calvert's survey [197] for measuring children's perceived trust. For all items, children were first asked to indicate whether they agree with a statement (i.e., "yes" or "no") and then asked to clarify the magnitude to which they agree or disagree (i.e., "a bit" or "definitely"), leading to four possible ordinal response: "definitely no", "a bit no", "a bit yes", and "definitely yes" [254]. Finally, we asked them to compare both CAs for each dimension.

Interview Data We qualitatively analyzed the video recordings of participants's usage of DAPIE, and the interview data from the children and parents. One of the authors iteratively coded the data through inductive analysis, and the other authors reviewed and verified the coding results.

3.5.4 Results

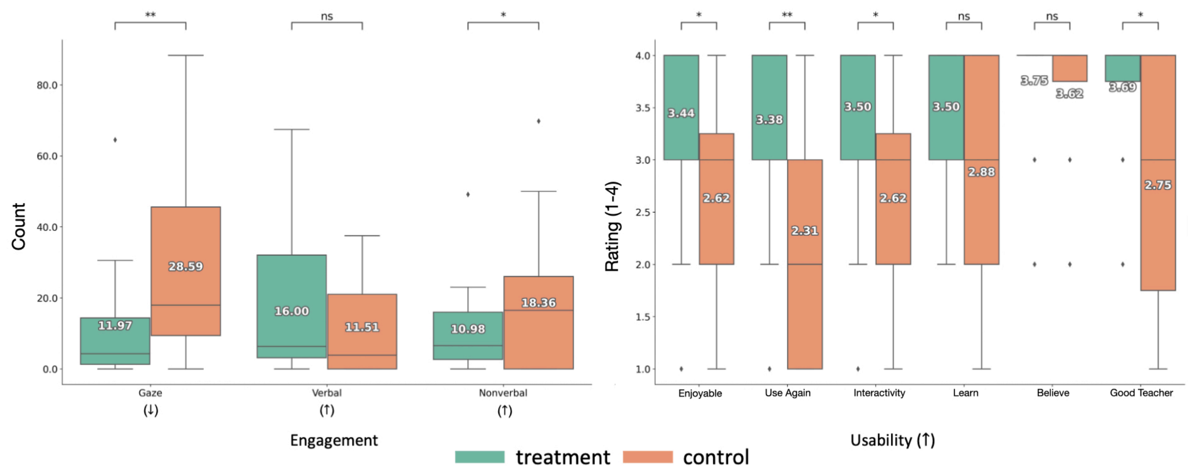

Figure 3.9: Results of engagement and usability analysis. **, *, and ns indicate significance of $p < 0.01$, $p < 0.05$, and $p > 0.05$, respectively.
Bar graphs showing the results for the engagement and usability evaluations. For engagement, gaze counts were significantly (**) higher in control, verbal counts were higher in treatment but the difference was not significant, and nonverbal counts where significantly (*) higher in control. For usability, ratings for "enjoyable", "use again", "interactibility" and "good teacher" were significantly higher in treatment (*), but ratings for "learn" and "believe" where not significantly different despite being higher in treatment.

Children performed better on the immediate understanding assessment and were more engaged in the dialogues when using DAPIE compared to the baseline. To statistically analyze each measure under different conditions, we first conducted a Shapiro-Wilk test to determine if the data was parametric (P) or non-parametric (NP). Then, to compare between conditions, we used a paired t-test (if parametric) and a Wilcoxon signed-rank test (if non-parametric).

Immediate Assessment Score Analysis Out of a maximum of 12 points for the understanding assessment, children's score when they used DAPIE ($M = 7.43, SD = 2.57$) significantly outperformed scores with the baseline ($M = 5.13, SD = 2.52$). The difference equates to correctly answering one more question out of six questions ($p < 0.05$, NP).

We observed that this greater understanding with DAPIE was possibly due to how it provided adaptive explanations (i.e., simplifying, defining difficult words, or providing examples). Through the system, fifteen children received adaptive explanations more than once. We did not ask children whether they understood an explanation after they received each adjustment to not influence their understanding. However, we observed that, on average, the children went through a whole thread of adjusted explanations at least once per dialogue, which might show that children needed all the adjustments once per dialogue. We consider that children who received adaptive explanations were able to digest the information more easily, leading to a better understanding. P14 mentioned, *"DAPIE is more like my dad or kind teachers who explain again more easily when I couldn't understand"*.

Additionally, we observed that DAPIE questions could encourage more active learning. When DAPIE provided diagnosis or elicit questions with multiple choices, some of the children kept moving their cursors while they thought of an answer (C1, C2, C3, C8), and others expressed excitement after getting the correct answer (C1, C6, C12). C7 said, *"I can focus more to get correct answers. I'm happy when I get answers"*. This finding is aligned with prior work that argues that meaningful learning occurs when learners think about the information presented rather than just passively receiving it [37].

Furthermore, children generally felt that the language from DAPIE was easier to understand than the language from the baseline. C4, C5, C9, and C14 mentioned that they liked DAPIE more as its explanations were "easier". For example, C9 said, "*DAPIE is like my friend because words are easier than the other one so it makes me more comfortable*".

Engagement Analysis By analyzing the participants' behaviors, we observed that DAPIE could promote engagement. Children's gaze distraction when using DAPIE ($M = 11.9\%, SD = 0.16$) was statistically lower than when they used the baseline ($M = 29\%, SD = 0.28, p < 0.01$). Also, participants made verbal comments more frequently when using DAPIE ($M = 18.7\%, SD = 0.11$) than when they used the baseline ($M = 11.5\%, SD = 0.13$), but there was no statistical difference ($p > 0.05$, NP). However, children's non-verbal comments when using DAPIE ($M = 11\%, SD = 0.12$) were statistically lower than with the baseline ($M = 18.36\%, SD = 0.19, p < 0.05$). In addition to these behaviors, we also analyzed overall interaction time and observed that children used DAPIE for longer ($M = 7.03\,minutes,\ SD = 3.03$) than the baseline ($M = 2.80\,minutes,\ SD = 1.33$).

Fifteen children chose to listen to the explanation on detailed units when guiding questions were given. Although this could make the dialogues longer, the children focused on the explanations and did not blindly skip or ignore these dialogue turns. Thus, with DAPIE, participants saw an average of three times more turns (20) than with the baseline (7) while also being more engaged. Moreover, when DAPIE provided diagnosis and guiding questions, some participants reacted with a smile (C3, C6, C12, C13) or said their answer out loud while also providing rationales (C1, C6, C14), showing their engagement and curiosity. Thus, children tended to be more engaged in the interactive dialogue, which allowed the child to decide whether they wanted more details or not. This finding aligns with prior work that showed granting agency to a child can increase engagement in an activity [143].

For gaze distraction, we observed that it usually occurred when the child wanted to say something to their parents or when they could not concentrate on the dialogue (e.g., distracted by surrounding noises or lost interest). With DAPIE, we observed that the former happened more frequently. Several of the children looked at their parents to explain their rationale, boast about being correct, and talk about other related topics. However, with the baseline, we observed more of the latter behavior where some children looked at other things while the system was talking and clicked the "Okay" button as soon as it was displayed.

Usability Analysis Overall, the children felt that the dialogues from DAPIE were more enjoyable and trustworthy than the baseline's (Fig. 3.9). In terms of enjoyment, children felt that DAPIE was significantly more enjoyable and interactive. They expressed that they would like to use DAPIE again ($M = 3.38, SD = 1.05$) significantly more than the baseline ($M = 2.31, SD = 1.10, p < 0.05$). In terms of trust, children felt DAPIE was significantly a better teacher ($M = 3.69, SD = 0.58$) than the baseline ($M = 2.75, SD = 1.15, p < 0.05$). Children also felt that they learned new things with DAPIE ($p = 0.07$) and that the system was more trustworthy ($p = 0.60$), but these differences were not significant.

Children noted DAPIE's easier language, questions and corrections, and various explanations as reasons for their positive reactions. For example, C5 said *"DAPIE is a good teacher. It's easier, and it makes me understand new information"*. C6 mentioned that they wanted to use DAPIE again because it provided "a lot of stories", and C9 felt he learned new things from the system because it provided more detailed information, which he wanted. While most children liked the interactive experience, C12 noted that the extra effort involved in answering and clicking could be burdensome.

Regarding trustworthiness, which showed the smallest difference between the two conditions, children generally believed that both systems were trustworthy since they both provided new information, which made the systems appear smart. However, several children commented that DAPIE seemed smarter because it acted like a teacher—e.g., helping them understand, correcting their answers, providing questions. On the other hand, some children also perceived the baseline as more trustworthy and intelligent since it talked in longer sentences.

Parents' Perception In their interviews, parents mentioned that DAPIE was more interactive and provided explanations that were easier to digest, which aligned with their children's comments. Also, parents were surprised that their children could focus on longer dialogues and expressed that this was due to the interactivity. Others mentioned that, by watching how their children enjoyed interacting with DAPIE, they also learned about how they should interact with their child.

Some parents emphasized the benefits of the adaptive explanation provided by DAPIE. They explained how the system acted similarly to how parents or teachers might adapt explanations for their children. For example, parents mentioned that the questions provided by DAPIE were similar to the questions they wanted to give their child when they interacted with the baseline. P14 said, "*When DAPIE explained, I wanted to ask whether my child knows the meaning of these words like "experts" and "product," but [the system] asked these questions to my daughter, so I like it.*"

Parents emphasized the benefits of DAPIE's more accessible explanations and its correction feedback, and related this to their own challenges in delivering scientific information. They mentioned that explaining science to their children is challenging as it is difficult to understand the information and to transform it into simpler expressions (P9, P11, P12, P13). For example, P9 said, *"My son asks me these questions quite frequently, but I couldn't always know the answers, so I googled the question to get information and change it in a way that my child could understand. This process is challenging, so sometimes I can't care about whether my son understands or not."* Moreover, P8, P14, and P16 said that they were able to learn about new information from DAPIE, and that interacting with DAPIE could serve as a bonding activity for parents and children. P8 mentioned, *"It reminded me of what I learned in school. Interestingly, we can learn together"*. However, some parents also mentioned that it was challenging to detect AI-related errors in the turns generated by DAPIE. In fact, several parents were unable to recognize that the dialogues were generated by an AI model at first, and thus failed to recognize any significant errors in the explanations.

3.6 Discussion and Future Work

AI-based Interactive Dialogues for Children and Parents In a real-life setting, DAPIE can be beneficial for answering children's questions when parents are unable or unavailable. Parents may struggle to understand the information needed to answer children's diverse and unpredictable questions [162] and, as expressed by parents in our study, they must dedicate significant effort to make the information comprehensible for their children. Through our system, children can satisfy their curiosity by accessing and consuming explanations on the internet, whenever they need them, without parents' effort.

Furthermore, we believe that parents and DAPIE can collaborate and combine their expertise to produce more meaningful experiences for their children: parents as experts in interacting with their children, and DAPIE as an expert on the domain. DAPIE can easily retrieve an explanation, simplify it, and design initial interventions—time-consuming and burdensome tasks for parents. By handing off this burden, parents can then focus on their child by observing how they interact with the system and noticing any difficulties. Future extensions of DAPIE could allow the parent to mediate in the dialogue and offer this knowledge to DAPIE, which it can then use to generate more personalized interventions. This presents an effective case of human-AI collaboration that leverages the respective strengths of the human and the AI.

AI Errors: Propagate Outputs but Intercept Errors While children were generally positive about generated turns, our AI-based pipeline could occasionally produce unsatisfactory or nonsensical turns due to its inability to propagate outputs across turns. As turns are considered separately when generating questions, the pipeline could lose context about the dialogue and generate confusing questions. In one dialogue, a sentence used the pronoun "these" to refer to an entity in the previous turn, and the pipeline generated a diagnosis question with "these", "that" and "those" as answer options. This question frustrated child participants as the options were nonsensical but, if they did not choose "these", they were told that they were wrong. Further, as the pipeline does not consider what was provided in prior turns, it could generate similar questions across multiple turns. For example, one dialogue explained aspects of gravity in each turn and provided several diagnosis questions where the answer was "gravity"—causing children to feel tired and bored. While these failure cases indicate that the pipeline should propagate information across turns, we observed that propagating AI's outputs could lead to error propagation. Thus, to prevent errors from propagation, the pipeline must also incorporate modules to detect, filter and/or correct failed generations before they are propagated. Additionally, the system can include simple feedback buttons (e.g., "bored", "bot was not smart") for children to explicitly express intent, and for the system to mitigate the cost of errors by providing other explanations.

Beyond Young Children and Beyond Question Answering Although our system targets young children (ages five to seven), we believe that our guidelines and the general structure of our pipeline can generalize to support older children and even adults. As learning theory [236, 199] suggests reducing support (i.e., fading) according to children's ability, the pipeline can be adjusted to make more challenging and cognitively engaging dialogues for older children. For example, the pipeline could provide longer explanations per turn or simplify less to help children learn new terms. Also, instead of the recall-based diagnosis questions, which bored some children in our study, the pipeline could generate self-explanation questions which challenge the child to explain what they have just learned.

A significant merit of our computational pipeline is that it maintains the core information in the original answer, but *wraps* it with interactivity. Beyond supporting children and question answering, this functionality can be extended to enable a new form of reading support. In learning contexts, our pipeline could be extended to assist in the active reading of documents (e.g., textbooks) by creating document-based dialogue turns that guide readers to other relevant parts of the document, diagnose their understanding, and provide adaptive interventions. Whereas current document-based QA models (e.g., Qasper [50]) focus on facilitating information-seeking, this approach could generate dialogues that focus on cognitively engaging users in the reading activity.

Limitations and Future Work As our pipeline generates dialogues while retaining core information, parents in our study said that they rarely noticed any harms or dangers in the dialogues. However, LLMs can exhibit biases [21] and our pipeline could propagate or even amplify various biases (e.g., gender, race, and culture). For these reasons, safeguards are needed to prevent negative impact on children. For example, our system could apply NLP techniques to recognize and mitigate biases [223] or, instead of generating on-the-fly, allow parents to verify dialogues before their children can access them.

Our study had several limitations. First, the study compared our full system to a baseline that provides sentences one-by-one. We adopted this design to evaluate the comprehensive experience supported by our system, but this makes it difficult to discern the effect of each component (e.g., simplification, questions). Second, as we assessed children's understanding immediately after each dialogue, the effect of the system on children's long-term retention is unclear. Third, although our target scenario is for when children ask "why" and "how" questions, we did not allow children in our study to ask the questions to strictly control the experiment design. Future work could explore how children ask questions with our system through a deployment study. Finally, our participant pool was skewed regarding race, age, and usage of voice assistants. Future work could conduct studies with younger children and more diverse demographics.

Chapter 4. PaperWeaver: Providing Personalized Explanations about New Papers for Researchers

This chapter presents the second example of cognitively aligned AI systems: PaperWeaver, a system that provides personalized research paper descriptions tailored to researchers' inferred prior knowledge. The content in this chapter adapts, updates, and extends material originally published in CHI 2024 [**?**]. All uses of "we," "our," and "us" refer to the coauthors of that publication.

Whereas Chapter 3 (DAPIE) demonstrated how asking questions can reveal children's potential learning challenges and maintain engagement through dialogue, this chapter addresses a different context: knowledge workers. In professional or research settings, repeated questioning may be disruptive or burdensome. This chapter explores how user's cognitive states can be inferred more implicitly and unobtrusively, offering insights into designing cognitively aligned AI systems that support knowledge understanding across different user types and contexts.

4.1 Motivation and Contributions

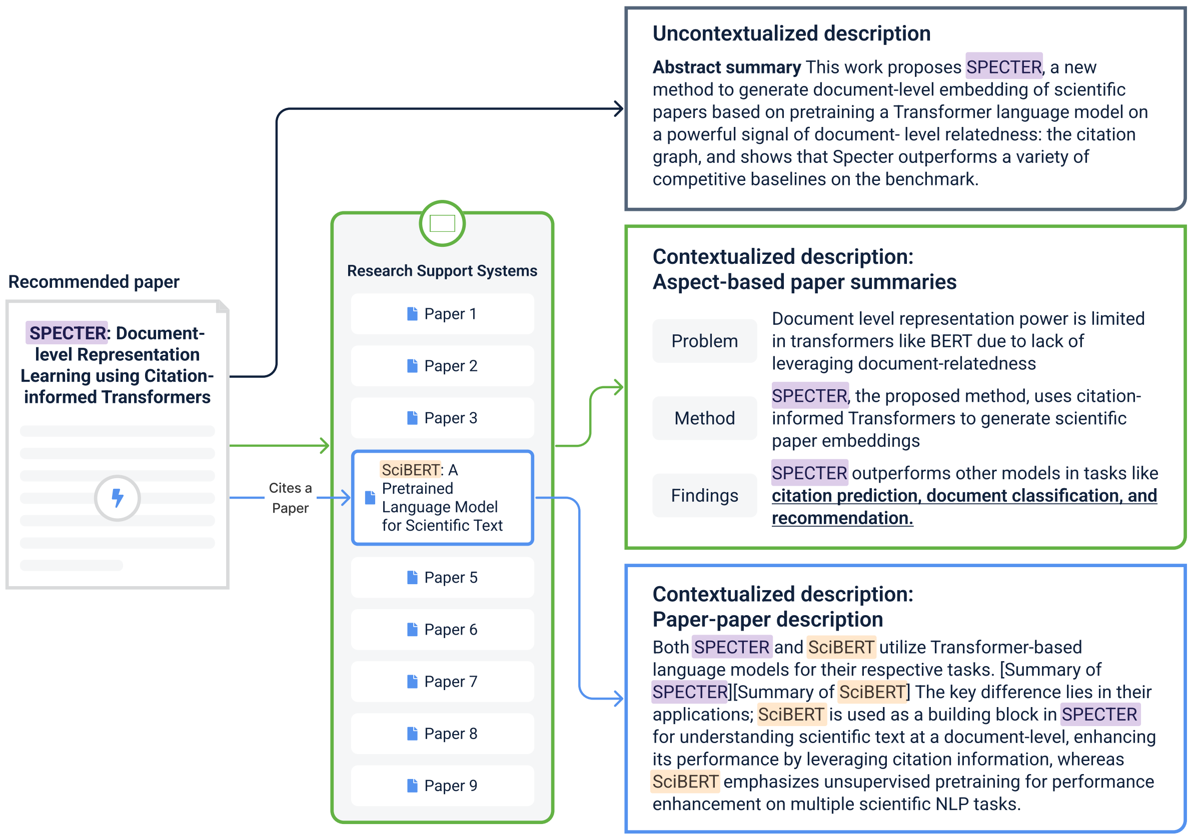

Figure 4.1: In contrast to existing topical paper alert systems that show descriptions with no personalized context for the recommended papers, PaperWeaver provides contextualized descriptions that surface the relevance of recommended papers and anchor them to familiar user-collected folders to help users better make sense of recommended papers.

Managing an ever-increasing accumulation of knowledge has long been a challenge for scholars [20]. With the recent proliferation of published materials, researchers face an even bigger challenge of keeping up with the literature [84, 179, 247]. Fundamentally, scholars need to both discover new and relevant papers and contextualize them to their own research interests. One popular practice recently is to leverage recommender systems that can help researchers retrieve potentially relevant, such as arXivist[1], arXiv Sanity[2] or Google Scholar[3]. Typically, these systems allow users to create "paper alerts" by providing a short description of a specific research topic and a set of seed papers as examples of papers of interest. For example, Semantic Scholar[4] allows users to save sets of collected papers under named topical folders. Users would then receive periodic paper alerts that contain a list of recently published papers similar to the collected papers. These alerts help researchers to quickly narrow down from all recent publications to a small set of potentially relevant papers, allowing them to more easily stay up to date on research topics that are of interest to them.

However, when receiving s set of potentially relevant papers in an alert, researchers still need to more deeply inspect each paper to understand its relevance. This typically involves making meaningful connections between newly encountered information and their existing knowledge of the literature—a process that can incur high cognitive costs for researchers. To illustrate this process with an example, a researcher working on *Research Support Systems* may encounter a relevant new paper titled "SPECTER: Document-level Representation Learning using Citation-informed Transformers" [43]. However, it can be challenging for the researcher to realize its relevance since the paper title only contains information about how it uses Transformers to generate document embeddings. Only when the researcher carefully examines the full abstract can they learn that the paper further described how the pre-trained model can be "easily applied to develop downstream applications," and that the paper implemented a "research paper recommender system" for its evaluation. Further, if the researcher spent additional efforts to examine the content of the recommended paper, they might discover that the related work section described how the proposed method was built on SciBERT[17]—a familiar paper they had recently saved and read—but extends its capability from embedding sentences to long documents. This manual process is effortful but important because failure to identify meaningful connections between new and existing knowledge will lead to overlooking new papers relevant to a researcher's interests.

Further, existing paper recommender systems typically only show a list of titles and abstract summaries for the recommended papers with little information on *how* they were relevant to the topic of the folder or the set of seed papers that the recommendations were based on. Findings from our formative study (Section 4.2) suggest that the title of a paper often lacks enough details that help researchers understand the paper's relevance yet the abstract is often too long to skim through for paper alerts. Prior work has explored an approach that generates one-sentence TL;DR (too long; didn't read) summaries [28] that are easier to consume but they lack contextualization to the folder topic and the collected papers in them. Specifically, since the summaries are not tailored to a user's folder context, parts of the abstracts that showed how the recommended papers were relevant to the folder can sometimes be omitted. While there has been research exploring ways to better contextualize paper recommendations by surfacing personalized social signals (e.g., based on a user's prior interaction or publication history [96]), they do not describe how the content of the papers is relevant to the users. As a result, users are left to decide whether to carefully examine the recommended papers to find potential connections, with no guarantee that the effort will pay off. As shown in the above example, this sensemaking process can be effortful for users, and failure to effectively triage paper recommendations could result in overlooking important paper recommendations and reducing the effectiveness of such systems.

[1] https://arxivist.com/
[2] https://arxiv-sanity-lite.com/
[3] https://scholar.google.com/
[4] https://www.semanticscholar.org/

In this work, we investigate what types of information in paper alerts help scholars deeply understand the relevance of recommended papers to their topical folders. In a formative study with seven researchers, we investigated challenges in identifying relevance from existing paper alerts and desired alternative descriptions of the recommended papers. Our formative findings suggested that scholars desire paper alerts that were contextualized to their folders compared to only showing titles, abstracts, and uncontextualized summaries. Participants found that descriptions about the recommended papers that revealed connections of a recommended paper to the user's folder context helped them understand the recommended paper more effectively since it spotlights where to focus on among many aspects of the paper. We also found that presenting the comparison and contrasting descriptions of multiple papers allowed participants to understand how the recommended papers build on prior work they had already collected in their folders. Anchoring unfamiliar papers with collected familiar papers also reduced the cognitive load of processing new information.

Based on our formative findings, we propose PAPERWEAVER, a new paper alert system that can enrich existing paper recommender systems by generating descriptions of how each recommended paper relates to a user's interests and their collection of papers. Our system is built on an existing document recommender system. PAPERWEAVER leverages recent advancements in Large Language Models (LLMs) for text generation to support users in several ways. First, PAPERWEAVER generates a compact topic description of a set of user-collected papers. This compact topic description provides users with a quick summary of the collected papers. It is user-editable and useful for generating future descriptions for each recommended paper contextualized to a user's interest. Second, to help users understand how a recommended paper is relevant to their research context, PAPERWEAVER generates two types of complementary contextual descriptions: *contextualized aspect-based summaries* and *paper-paper descriptions* (Fig. 4.1). *Contextualized aspect-based summaries* leverage the generated topic descriptions to extract statements on the problems, methods, and findings of papers that are highly relevant to the topic of interest. *Paper-paper descriptions* summarize how a recommended paper relates to collected papers, which are more familiar to the users. If the recommended paper cites collected papers, PAPERWEAVER summarizes these citation descriptions and, if not, it synthesizes relationships by comparing and contrasting aspects from the papers. Motivated by how multiple alternative descriptions can improve understanding of complex scientific topics [2, 3], we design an interactive paper alert interface where users can explore multiple descriptions for recommended papers (Fig. 4.2).

To evaluate PAPERWEAVER, we conducted a within-subjects study (N = 15) with researchers who were interested in receiving paper recommendation alerts. To ensure participants were motivated in the study, the paper alerts used in the study were generated based on their actual set of collected papers. We compared PAPERWEAVER with a strong baseline similar to existing paper alert systems but additionally enriched with uncontextualized summaries and extracted related work sections from the recommended papers. Our user study results showed that participants were able to better understand the nuanced relevance of the recommended papers and were able to triage them more confidently with PAPERWEAVER. Further, participants were able to capture richer relationships between recommended and collected papers in their notes when using PAPERWEAVER compared to the baseline.

The contributions of this work are as follows:

- Qualitative findings from a formative study employing design probes with researchers that identified user challenges around making sense of recommended papers and the need for contextualized summaries.

- PAPERWEAVER, a tool that provides additional contextualized descriptions for a set of recommended papers, tailored to the user-collected papers. PAPERWEAVER uses an LLM-based pipeline that synthesizes content from both recommended and user-collected papers.

- Findings from a user study (N = 15) that demonstrated how using PAPERWEAVER facilitates sense-making of paper recommendations and aids in uncovering useful relationships between recommended and collected papers.

4.2 Formative Study

We conducted a formative study with two goals. Firstly, to learn about researchers' current challenges when reading existing paper alerts that consist of a list of titles and abstract. Secondly, to gain a deeper understanding of the types of information that would be more useful in a paper alert setting to inform the design of PAPERWEAVER.

4.2.1 Procedure and Apparatus

We conducted a targeted recruitment of seven participants who are graduate students expressed interest or have worked on one of two predefined topics: "Reading support tools for scientific articles" or "Scholarly document processing." Four participants identified their discipline as human-computer interaction (HCI) and three as natural language processing (NLP).[5] To seed the two folders with papers that the participants are likely to be familiar with, we chose relevant papers published by their research groups. To generate paper recommendations for each of the two topics, we used a publicly available paper recommender API with the seed papers as inputs.[6]

We designed five different types of contextualized text descriptions as design probes to gather feedback. The design probes covered different ways of describing the recommended papers in the context of a paper alert. The descriptions were generated either manually or using a previous scientific abstract summarization method called TL;DR [28] and an LLM with prompts that focused on including different types of information, such as relevance to the folder via the folder name or relevance to the seed papers (List of design probes in Supplementary Materials).

To simulate realistic paper alert usage, we asked participants to consider the pre-defined folders as their own. In the first 15 minutes of the interview, participants reviewed 10 seed papers on the topic they had chosen to ensure that they are familiar with the papers we had collected for them. We then probed their experience with current paper alert systems by showing them recommended papers in the standard presentation as a list of titles and abstracts, and asked them about their goals and pain points. Afterward, we conducted a think-aloud interview by walking them through a set of design mock-ups with the design probes for 40 minutes. We probed the costs and benefits based on their reactions to different types of descriptions. Participants were also asked to participate in co-design by revising the descriptions as they saw fit. The interviews were screen recorded, and the first author conducted a thematic analysis to capture qualitative insights [24, 44]. The rest of this section lists the common themes from the interviews and describes the design probes when relevant.

4.2.2 Formative Study Findings

Challenges in Making Sense of Existing Paper Alerts When we showed a design probe similar to existing paper alert systems with a list of titles and abstracts for the recommended papers, all participants mentioned that the biggest challenge is to figure out how all the recommended papers are relevant to the topic of their folders. It is hard to quickly decide whether they should save and read the recommended papers from just their titles and abstracts. In order to effectively identify relevance, participants pointed to how they need to carefully read the full abstracts or even the papers themselves. However, doing so would take more time and effort than they typically spend on consuming paper recommendation alerts. The cost of carefully reading a list of full abstracts is so high that many participants resort to using a less effective strategy where they spot mentions of certain keywords in the

[5]Two participants joined in person and the rest joined remotely through Google Meet. This study was approved by our internal review board. Each participant was paid 30 USD for one hour of their time.

[6]https://api.semanticscholar.org/api-docs/recommendations

titles or whether it was published by a familiar author (P1, 3, 4, and 6). At the same time, they acknowledge that this strategy leads to overlooking relevant papers. Participants pointed to two reasons: (1) They might not know all relevant authors in the space which suggests previous social-signal-based approaches could be insufficient [96]. (2) The keyword spotting only allows them to find high-level topical relevance (e.g., "scientific papers") but does not allow them to find deeper connections hidden in the abstracts (e.g., "semantic embeddings that can support citation predictions" [43]).

Surfacing Relevant Aspects when Summarizing Recommended Papers We explicitly asked participants to compare different types of description in three mock-ups, each with different descriptions to the same set of paper recommendations: full abstracts, "TL;DR" abstract summaries generated using [28], and abstract summaries that focused on including information relevant to the folder name generated using an LLM. When comparing with the full abstracts, participants reacted positively with the two shorter abstract summaries over reading the full abstracts that were longer and less concise. Additionally, they preferred the LLM-generated contextualized summaries over the uncontextualized TL;DR summaries. Specifically, participants felt that they could more effectively understand why the recommended papers were relevant while the TL;DR summaries often omitted parts of the abstract that were relevant to the folder. For example, P2 and P5 liked the contextualized sentence: "*You may be interested in this paper because it addresses the issue of trustworthiness and factuality in question answering systems, which can be relevant to processing scientific documents*[i.e., the folder name]," which highlights parts of the abstract most relevant to the folder. Additionally, we observed that participants less familiar with the topic reacted particularly strongly with this type of explanation while more experienced participants felt the descriptions were accurate but sometimes trivial because they had more background knowledge in the domain.

Comparing and Contrasting Recommended and User-collected Papers Participants emphasized that the main goal when making sense of paper recommendations is to situate recently published papers within their own research context. With this goal, participants mentioned that the design probe with descriptions that focused on showing connections between recommended and collected papers helped them recognize recommended papers that directly build on a paper they had collected. Additionally, anchoring recommended papers to a familiar collected paper helped them more effectively understand how they were situated within the literature. For example, P3 found a description that connected two papers by a similar research problem to be useful, and said "By aligning the [similar] motivations of Foundwright [i.e., [183]; a recommended paper] with the motivation of Citesee[i.e., [32]; a collected paper], I don't need to find and read two [separate] sentences that explain the problem in the [two] abstracts."

Bringing New Perspectives to Previously Collected Papers One interesting observation was that participants found value not only in seeing information about the recommended papers, but their connections to papers in the folders also provided new insights about them (P1, P3, P6). For example, when seeing a design mock that showed the citation context (i.e., citing sentences) in a recently published recommended paper that cited one of the papers in the folder, participants felt that they had "rediscovered" the collected papers and how the descriptions "makes me feel like the collected papers are still relevant." and that they "[re]introduced papers that I already saved but had not read deeply" providing them with "new opportunities to better understand them and learn new perspectives about them" (P1).

4.2.3 Design Goals

Based on insights from the formative interviews, we formalize the following Design Goals for our system:

[DG1] Describe details about the recommended papers in a way that helps users understand how they are relevant to the topic of their research context (e.g., folders) to avoid overlooking relevant recommended papers.

[DG2] Help users make connections between the recommended and collected papers by comparing and contrasting them.

[DG3] Reveal new aspects of previously collected papers to keep their understanding of the papers up-to-date and remind users of unread collected papers that have become more relevant with new surfaced connections.

4.3 PAPERWEAVER

Based on our design goals, we present PAPERWEAVER (Fig. 4.2), a paper alert system that contextualizes recommended papers' descriptions based on a user's topical folder information. PAPERWEAVER enables users to explore these descriptions while reading paper alerts. With lessons learned from the formative study, our computational method leverages a combination of an LLM-generated text and extracted related work sections from the recommended papers. PAPERWEAVER extracts folder-specific aspects from a recommended paper to surface how it is relevant to the user's context (DG1). The system describes how the recommended paper is similar to and different from a topical folder paper by identifying the relationships between papers (DG2). To remind users about papers they have previously collected in the folder, PAPERWEAVER includes information about both the recommended and the collected papers in the descriptions (DG3). Given a paper recommendation and a set of collected papers in a named topical folder, the system provides "*paper-paper descriptions*" that compare and contrast the recommended paper anchored to a relevant collected paper. To generate this type of descriptions, the system uses a pair of an abstract and a citance as input when available. For recommended papers that do not cite any

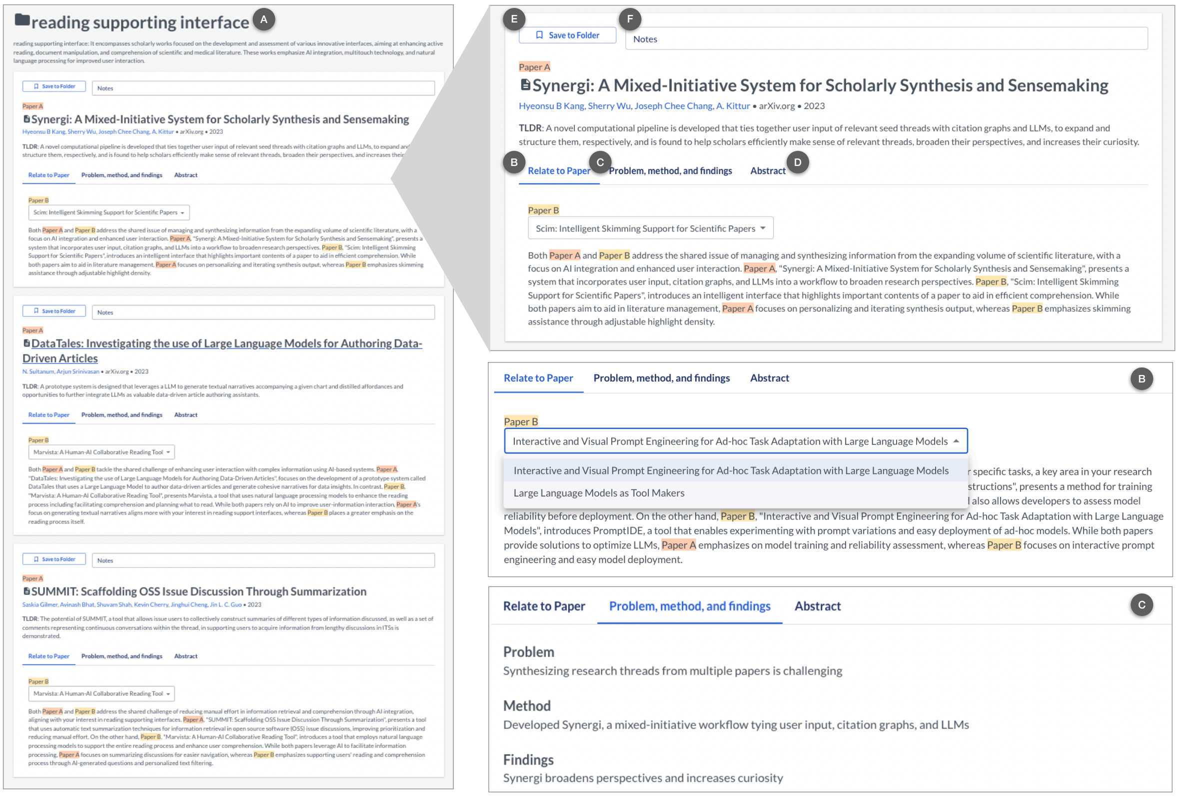

Figure 4.2: PAPERWEAVER's paper alert interface. The page shows the folder title and description (A) and a list of recommended papers. Each recommended paper is shown on a paper card with the title, authors, venue, and publication year. The bottom of the card features descriptions in three tabs: *Relate to Paper* (B) allows users to select a paper from existing papers in the folder with the description related to the recommended paper. These descriptions are *paper-paper Descriptions Based on Citances* and *Paper-paper Descriptions via Generated Pseudo-citances* that focus on specific relationships between two papers. The description assigns the label "Paper A" to the recommended paper and "Paper B" to the existing paper in the library and highlights the references in different colors; *Problem, method, and findings* (C) describes the recommended papers in three aspects— problem, method, and findings—related to the folder context with *contextualized Aspect-based Paper Summaries*; *Abstract* (D) shows the unmodified abstract of the paper. The user can save a paper to the library (E) and take notes about the recommended paper (F).

of the collected papers, PAPERWEAVER uses an LLM to identify latent relationships from their abstracts. Finally, PAPERWEAVER also generates "*contextualized aspect-based paper descriptions*" that summarizes a recommended paper's abstract in a way that reflects the paper's relevance to the topic of the given folder.

4.3.1 Example User Scenario

Imagine a computer science researcher who started working on the topic of *Reading-supporting interfaces* a few weeks ago. She has collected a list of papers related to this topic in a folder. Her familiarity with each paper in the folder varies. She has read some of the papers during her prior project. For some other papers, she only saved them because she saw relevant keywords in the title of the paper but has not had the time to read them yet. Because this is her active project, she is looking out for new papers relevant to the topic. She signed up for a paper alert service that regularly provides her with a list of papers based on papers that she has collected. However, she felt that the information provided was lacking. She often had to wade into a paper's abstract and its full text to see whether the paper was relevant to her research interests. This activity requires more time and attention than she usually has when reading paper alerts.

Looking for a more effective way to process paper alerts, she gives PAPERWEAVER a try. She starts using the system by creating a folder titled *Reading-supporting interfaces* and adding the list of papers she has collected to the folder. PAPERWEAVER automatically expanded on the folder title to provide an overview summary of the source papers she collected. Although she is mostly satisfied with the automated summary, she removes the keyword *medical literature* in the description and adds another keyword *Large Language Models* to better reflect her research interests. She also notices a keyword *multitouch technology*, which reveals an aspect she has not thought of before. PAPERWEAVER saves the updated description and uses this description throughout the rest of the process.

After setting up the folder, the researcher receives a paper alert email when there is a new set of recommended papers. By clicking on the link in the email, she is directed to an interactive paper alert interface (Fig. 4.2). She bookmarks this link for future reference. She quickly goes through the recommended papers and immediately saves papers with obvious relevance (e.g., "The Semantic Reader Project: Augmenting Scholarly Documents through AI-Powered Interactive Reading Interface"[150]) based on the provided title and metadata (authors and TL;DR). For a paper with non-obvious connections, she explores the contextualized descriptions of the paper. She first opens the *Problem, method, and findings* tab to see a summary of the paper by the folder context broken down into problem, method, and findings. For example, for the recommended paper "Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking" [97], she can see that PAPERWEAVER-provided method of the paper is relevant to her interest in LLM ("Method: Developed Synergi, a mixed-initiative workflow tying user input, citation graphs, and LLMs") and the paper's problem is "Problem: synthesizing research threads from multiple papers is challenging". Seeing such connections with *Reading-supporting interfaces*, she saves the paper to her library folder. Feeling curious about the Synergi paper, she decides to explore the paper-paper relationship descriptions under the *Relate to Paper* tab. She chooses to compare the recommended paper with a paper "CiteRead: Integrating Localized Citation Contexts into Scientific Paper Reading" [188] that she is already familiar with. She learns that the system in the Synergi paper requires more active engagement from users compared to the system in the CiteRead paper. Now, she notes interactivity as one dimension of the reading interfaces design space. She also sees that there is a description for the Synergi paper and a paper she has collected in passing. By reading the description, she now recalls that "Scim: Intelligent Skimming Support for Scientific Papers" [63] supports paper reading by automatically highlighting important content in a paper and, unlike Synergi, there is no personalization component in the system. Compared to other traditional paper alert systems that only show a list of titles and, at most, abstracts, PAPERWEAVER helps the researcher figure out how each of the recommended papers connects to her topic of interest, as well as their nuanced relationships to papers she had already collected.

4.3.2 Methods for Generating Contextualized Descriptions

To generate contextualized descriptions like those presented in the scenario, we developed an LLM-based pipeline (Fig. 4.3) that processes the user's folder information and the recommended paper's content to generate descriptions for each recommended paper. LLMs can make meaningful improvements in comprehension and summarization, particularly for long, complex documents that demand a high degree of accuracy [5]. This capability enables PAPERWEAVER to identify relevant aspects with the given context in papers and synthesize descriptions from multiple aspects. PAPERWEAVER generates three types of descriptions: (1) *contextualized aspect-based paper summaries*, (2) *paper-paper descriptions based on citances*, and (3) *paper-paper descriptions via generated pseudo-citances*. We describe how PAPERWEAVER generates a compact summary of collected papers in a topical folder (§4.3.2) and uses the folder summary to generate contextualized aspect-based summaries (§4.3.2), contextualized

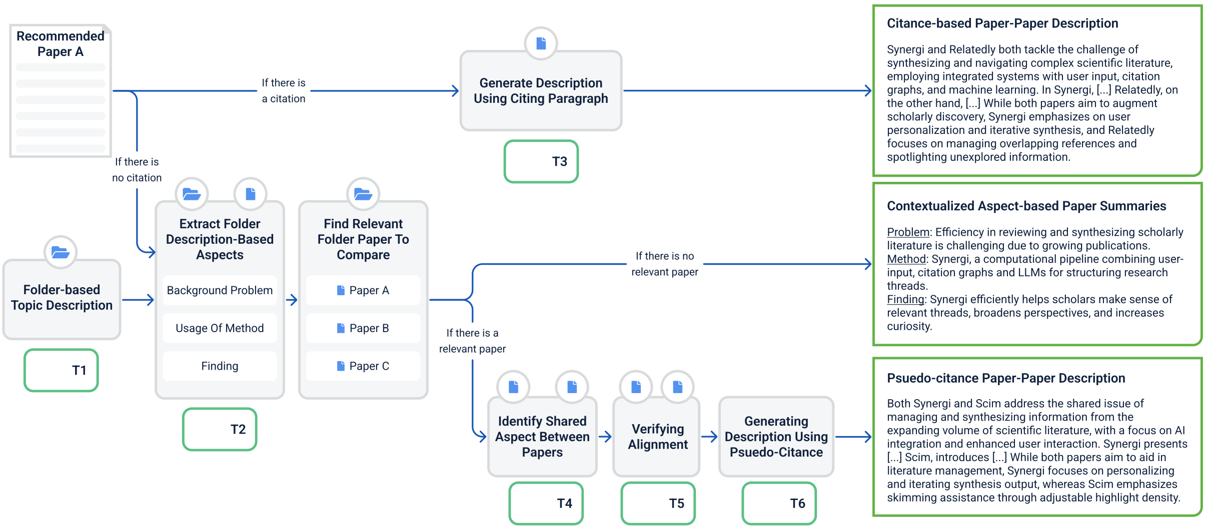

Figure 4.3: Overview of PAPERWEAVER's pipeline to generate contextualized descriptions. If there is a citation from the recommended paper to a collected paper, PAPERWEAVER generates *paper-paper descriptions based on citances* using citing paragraph. If there are no citances, PAPERWEAVER synthesizes pseudo-citing sentences that shows the relationship between recommended paper and relevant collected paper to make *paper-paper descriptions via generated pseudo-citances*. If there are no relevant collected papers, PAPERWEAVER generates *contextualized aspect-based paper summaries* with the folder's overall context.

paper-paper relationships when recommended papers cite one or more collected papers (§4.3.2), and when they do not (§4.3.2).

Suggested Topic Description To ensure the three types of descriptions (detailed in the following subsections) are relevant to the topic of the folder, PAPERWEAVER allows users to provide a compact description for their folders in addition to the folder name. This topic description is used in all subsequent LLM prompts to convey the user's interests. To lower the effort of writing folder descriptions, PAPERWEAVER uses an LLM to generate a default description based on collected papers already saved in the folder. We adapted our prompt design from the prior *LAMP* approach[206], which creates a user profile prompt with the information from the user's own paper. Our prompt takes a task instruction and a list of titles of papers included in the folder as inputs Then, our prompt instructs an LLM to generate a description including the folder title, shared goals of collected papers, and a list of topical keywords (Fig. 4.2A; folder title and description). This default folder description was then shown to the user. The user could further edit them to reflect their interests beyond the papers that they had collected.

Contextualized Aspect-based Paper Summaries Among the various aspects of the recommended paper, we extract the aspects (*i.e.,* rhetorical structure elements indicating problem, method, findings [168]) that the user who has curated this library folder might find relevant. The problems, methods, and findings are typically the main pillars of most papers [31]. At the same time, these aspects can describe research in a specific and comprehensive way. To extract a set of problems, methods, and findings in the context of the library folder, our method takes a title and an abstract of the recommended paper and the folder description that represents the user's research interest as inputs. We guide an LLM to identify as many relevant problems from the recommended paper as possible. Then, we describe the specific methods applied by the paper for each problem and elaborate on the specific findings identified by applying each method. Our method uses these aspects not only as the summary of a single recommended paper, we further build on the extracted aspects to align two papers to create pseudo-citances as described in §4.3.2.

Paper-paper Descriptions Based on Citances For this description type, we exploit citations between recommended papers and collected papers. Citation sentences (*i.e.,* citances) are widely used as proxies of the relationship between the citing and the cited papers [157]. Among the intents of citances (*i.e.,* background, method, or results), citances with the background intent give more context about a problem, concept, approach, topic, or importance of the problem in the field [42]. Citances with the background intent often contain information about how a citing paper presents a new approach and how it compares to a cited paper. Since our formative study revealed that participants regarded "build on" relationship to be important, we prioritize background citances as classified by [42] when selecting a citance to generate a description.

A citance by itself is hard to understand for users without any context. Once a citance was selected in a recommended paper, we extracted the paragraph which the citance was in (*i.e.,* citing paragraph) to obtain additional context around it. Rather than showing the citing paragraph that only has partial content of the recommended paper, we employ an LLM to generate a compact but detailed description that describes both the recommended paper and its relationship to the cited collected paper as mentioned in the citing paragraph. Additionally, to support our DG3 of helping users learn new aspects of collected papers, we added a short summary of the cited collected paper. To obtain this description, we designed inputs of the prompt to include the titles and abstracts of both the recommended and the cited paper, as well as the citing paragraph.

Paper-paper Descriptions via Generated Pseudo-citances The method described in the above section (§4.3.2) does not apply to all recommended papers. Authors typically do not comprehensively cite all relevant papers. Some relevant papers are omitted from the citances even when they are relevant to some of the collected papers. We build on prior work that used the problem-method-findings schema to show similarities and differences between papers in a search engine setting. Our method generates structured summaries using a similar approach to describe paper recommendations anchored to previously collected papers. In our design, this type of description features two core relations primitives between papers: 1) *comparisons*, which provide concise descriptions of the most salient similarities along either the problem or the method aspect and 2) *contrasts*, which surface differences along a different aspect (*e.g.,* two papers that tackled the same research problem but used different methods). This structure has been shown to facilitate scholarly sensemaking and inspirations (§2.2.3.)

Rather than concatenating each paper's aspects that are similar or different with extractive techniques, our method aims to provide well-aligned comparisons and contrasts between two papers. This method is motivated by the formative study where participants mentioned that simple concatenation was not helpful in making sense of the relationships between two papers. To create this description, our method follows the following process: (1) find relevant papers, (2) identify shared aspects for each pair of the recommended paper and a collected paper to find similarities and differences between them, (3) verify whether the shared aspect is aligned with both papers and (4) generate a structured summary.

Here, we explain our approach in more detail, using the case where the recommended paper and the collected paper share similar "problems". Our approach first selects the top-5 most similar collected papers to a recommended paper based on the abstract similarity using Flag embeddings [249], a state-of-the-art text embedding model. Then, with the aim of finding similarities and differences between papers, we use the method described in §4.3.2 to extract multiple problem-method aspect pairs from each paper's abstract. By providing the titles and the aspects of all five relevant collected papers and the recommended paper as inputs, we instruct an LLM to (1) identify the top-5 papers that have problems that are the most similar to the problems in the recommended paper, (2) list all of the identified pairs (one from the given paper and the other from a collected paper), and (3) describe one shared problem that could encompass the two identified problems. To confirm whether the shared problem encompasses problems of both the recommended paper and the chosen collected paper, our approach prompts an LLM to verify whether the shared problem is addressed in each paper. This was done by providing each paper's title, abstract, and the generated shared problem. Finally, by inputting respective contrasting methods employed in the recommended and the folder paper, along with the generated shared problem and the abstracts of both papers, the LLM generates the structured summary. This summary includes comparing and contrasting sentences with short summaries of two papers. While this process explains how to generate a description for two papers with similar problems, a similar process applies to a description of two papers with similar methods.

4.3.3 Paper Alert Interface

For each folder, PAPERWEAVER takes the folder title, the folder description, and a list of papers in the folder as inputs. We use the publicly available Semantic Scholar Paper Recommendation API[7] to retrieve a list of recommended papers for a given set of folded papers. For each folder, a dedicated web page allows users to view recommendations displayed on detailed cards, featuring information about the paper (title, authors, venue, and year) and a machine-generated TL;DR summary [28]. The title of a paper links to the corresponding paper details page on Semantic Scholar where users can access the pdf file of the paper (if available) together with other information such as the paper's citations and references. Users can explore different descriptions displayed in three tabs:

- *Related to Paper* (Fig. 4.2B) shows paper-paper descriptions that are focused on showing relationships between the recommended paper and a specific paper previously saved in the folder. The descriptions shown in this tab are *paper-paper descriptions based on citances* and *paper-paper descriptions via generated pseudo-citances*. The two papers mentioned in the text are highlighted in different colors to indicate which paper the description refers to. Users can use the dropdown control to select descriptions for the relationships between the recommended paper and a specific paper.
- *Problem, method, and findings* (Fig. 4.2C) shows the *contextualized aspect-based paper summary* for the recommended paper broken down into three aspects.
- *Abstract* (Fig. 4.2D) shows the original full abstract of the paper.

Through the paper alert interface, users can save a recommended paper (Fig. 4.2E) to their folder and write a note (Fig. 4.2F) about the recommended paper for future references.

4.3.4 Implementation Details

PAPERWEAVER interface is a standard web application. The back-end was implemented in Python using Flask for an HTTP server and PostgresSQL for a database. The front-end was written in TypeScript using React framework. PAPERWEAVER retrieves the paper's metadata (title, authors, abstract, etc.), TL;DR and citances from the public Semantic Scholar API [8]. We obtain the extracted plain text for papers using S2ORC, an open-source PDF-to-text extraction pipeline and a corpus of processed 81.1M academic papers across multiple disciplines [152]. We use GPT4-0613 through OpenAI API[9] for all text generation with an LLM.

4.4 User Study

We conducted a within-subjects laboratory study to investigate whether PAPERWEAVER can help readers efficiently understand the relevance of new papers, triage them, and deeply understand their relevance to users' own

[7] https://api.semanticscholar.org/api-docs/recommendations
[8] https://api.semanticscholar.org/api-docs/graph#tag/Paper-Data/operation/get_graph_get_paper
[9] https://api.openai.com/v1/chat/completions

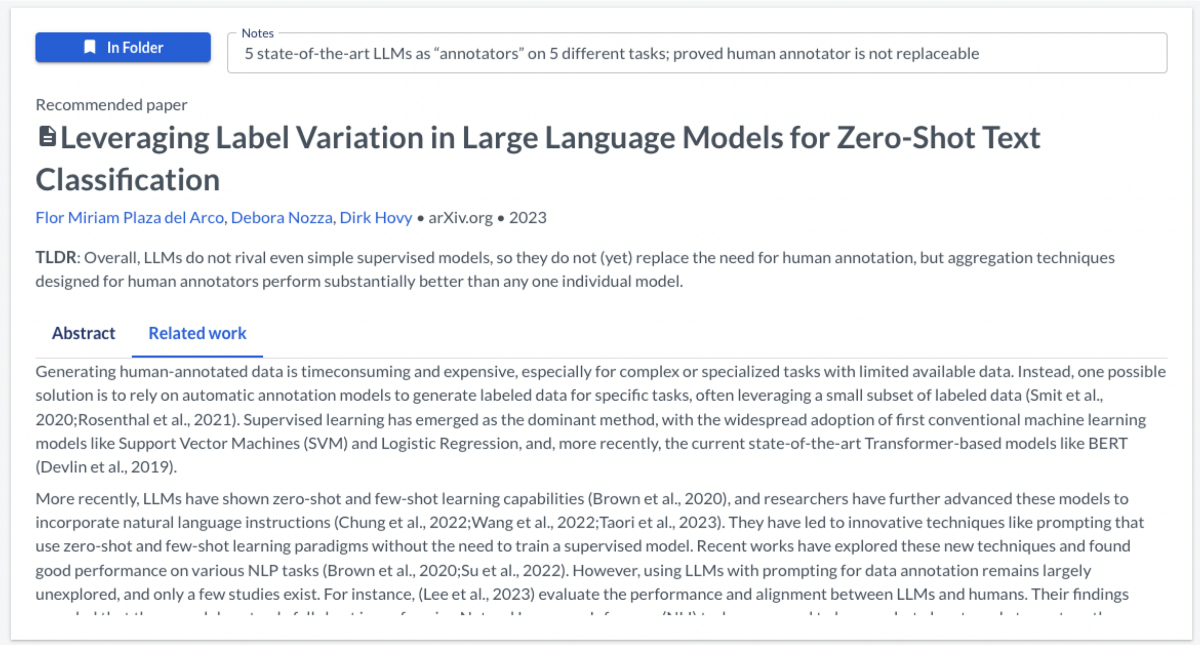

Figure 4.4: In the baseline condition, participants also see the recommended papers in a similar paper card to PAPERWEAVER paper alert interface. Instead of the three tabs with different types of explanation, there are two tabs: *Abstract* that shows the abstract of the recommended paper and *Related work* that shows the text of the related work section of the recommended paper. The rest of the functionalities remain the same across conditions.

context.

4.4.1 Research Questions

To this end, we formulated our research questions as follows:

[RQ1] How do relevance descriptions augmented by user-collected papers' information help a user understand how recommended papers are relevant to the user's research context?

[RQ2] How do relevance descriptions augmented by user-collected papers' information help a user triage the paper?

[RQ3] How do relevance descriptions augmented by relationships help a user understand the relationship between user-collected and recommended papers?

[RQ4] How do relevance descriptions augmented with a set of user-collected papers help a user understand new aspects of the collected papers?

4.4.2 Study Design

Participants We recruited 15 researchers who are Ph.D./MS students in the CS domain for the study through snowball sampling and the authors' social media (Twitter). Two of the participants were Master's students, seven were junior (first- and second-year) doctoral students, six were senior (third-, fourth-, and fifth-year) doctoral students. Our study focused on Ph.D./MS students as they may receive the greatest benefit from explaining why those papers are connected to their personal research context. Ph.D./MS students need to keep up with the large amount of literature the most. All participants reported that they read paper alerts, from monthly to daily.

Conditions For the study, participants read paper alerts and listed what they learned from them using either PAPERWEAVER or a baseline system. Like in PAPERWEAVER, each recommended paper in the system for the baseline condition (Fig 4.4) contains its TL;DR, abstract, and the link to the paper's details (the paper details page on Semantic Scholar). Additionally, we provide excerpts of the related work section of the recommended papers in the baseline condition. The related work section usually shows relationships between the paper and prior work. Researchers are likely to check this section to look for relationships between the recommended paper and other papers. While both PAPERWEAVER and the baseline system give the link to the full paper to inspect the paper further, having the related work section on the baseline condition helps balance the amount of readily available information between the two conditions. In the PAPERWEAVER condition, *contextualized aspect-based paper summaries* and *paper-paper descriptions* are additionally provided. The ordering of the conditions was counterbalanced to mitigate potential ordering effects.

Paper Alert Setup We asked participants to create their own library folders (paper collections) on Semantic Scholar and save papers into them before the study. To balance the quality and familiarity of folder content between the conditions, we asked participants to follow the instructions:

1. (Scenario and Number of Topic Folders) Pick three research topics for curation, and fix the use scenario (*e.g.,* surveying the literature to write a related work section in your paper).
2. (Topic Familiarity) Topics should be similarly familiar to you.
3. (Number and Familiarity of Papers) Collect five papers you fully read, three papers that you have read somewhat, and two papers that you have seen but not read.

We chose two folders from the three participants brought with them based on the following criteria, to reduce the significance of potential quality and familiarity differences between the papers actually used in the experiment: We generated recommendations for each folder and ranked them based on the content similarity using the average of cosine similarity between library papers' and recommended papers' SPECTER [43] embeddings; pick the top two. Regarding folder descriptions, participants saw an example and edited them before the study.

Procedure The study was conducted remotely using video conferencing software. After a brief introduction to the overall study, participants were given a short tutorial of the system (either treatment or control, counterbalanced). They interacted with the interface for 2 minutes to familiarize themselves with the interaction features using fake data. After the tutorial, participants were given 12 minutes to read the descriptions on 8 recommended papers. After they read the descriptions, they were provided with survey questions. Once the survey was completed, participants were given 10 minutes to write down a list of what they learned from reviewing the descriptions and wanted to retain for future use. Participants were asked to use their own language to describe the things and were informed the quality of writing was irrelevant. The steps above were repeated for the second condition. Finally, the experimenter conducted a semi-structured interview about participants' overall experience. The study lasted for ~90 minutes, and participants were compensated $80 USD for their time. The study was approved by our internal review board.

4.4.3 Measures

Overview In relation to the research questions described in §4.4.1, we measured participants' subjective responses to corresponding survey questions. To triangulate the survey responses, two of the authors manually coded what participants wrote down as their learning from reading the descriptions in terms of whether it contained elaboration on the relationship between a recommended paper and the saved papers participants began with or their curiosity for further investigation. We also employed a survey asking participants' general experience with each system. Finally, we conducted a sanity check on the degree of factual hallucination among the descriptions generated in PAPERWEAVER, to quantify its current limitations and point to future work directions.

Targeted Survey Responses We collected participants' responses (7-point Likert scale) to the following five questions from the survey: "*The system helped me understand how the recommended papers are relevant to my research interest.*" (RQ1), "*The system helped me decide which papers are worth saving.*" (RQ2), "*I was confident in deciding which papers were relevant and worth saving.*" (RQ2), "*The system helped me understand how the recommended papers relate to the papers I have already saved.*" (RQ3), and "*The system helped me learn something new about the papers I have already saved.*" (RQ4).

General Experience Survey Responses To measure perceived workload, the survey also included five questions (excluding the physical demand question) from the NASA-TLX questionnaire [82], where a more compact 7-point scale, mapped to the original 21-point scale, was instrumented [202].

Annotating Participants' Learning To deeply analyze participants' learning from reviewing the descriptions, we constructed an annotation regime in five dimensions to capture the quantity of facts and relationships described in participants' notes, and the level of elaboration around relationship- or curiosity-related details. Specifically, the five dimensions were: 1) the number of facts about papers; 2) the number of groups of papers synthesized by participants; 3) the number of relationships described between a paper and a group of papers or between groups of papers; 4) the level of factual details on the relationships; 5) mentions of curiosity and/or authentic motivation for further investigation. The first three dimensions were coded using an interval scale, ranging from 0; the last two dimensions were coded using a binary scale, where 0 represented 'lacking elaboration'. We sampled a set of randomly sampled 15 notes participants wrote from the full set of 169 notes for two annotators (two of the authors who were not experimenters in the study). The annotators coded and discussed these samples over two rounds to reach a consensus. The following is the rubrics developed for annotation:

Step 1 Identify a paper or a group of papers described in the comment.

Step 2 Identify whether there is a relationship described between a paper and a group of papers, among a group of papers, or between groups of papers.

Step 3 Once at least *one* relationship is found, find whether there are factual details or concepts elaborated about the relationship in the comment.

Step 4 Similarly with *Step 3*, identify if participants expressed surprise, curiosity, or motivation for future investigation in the text (*e.g.,* "I'm not sure how that difference manifests yet.", "I'm surprised by the results, I'll look into the papers later.").

Step 5 Count the number of distinct facts, using the clause or sentence boundaries.

The annotators annotated 20 shared samples (randomly drawn) blind-to-condition to check for inter-rater reliability. This resulted in Krippendorff's α's ranging between 0.46 (the binary measure of factual detailedness in descried relationships) and 0.78 (the number of groups of papers synthesized by participants).

Accuracy of LLM-generated Descriptions Additionally, as LLM can generate hallucinations (*i.e.,* factually incorrect information), we collected annotations of factual correctness for 60 descriptions for each of *contextualized aspect-based description* and *paper-paper description* used in the user study. Since evaluating the factual correctness of our description requires expertise in these domains, we recruited two experts per domain for the HCI and NLP domains for annotation. Each expert annotated the data for recommended papers in their domain of expertise. For each description, experts rated its factual correctness on a three-point scale based on counting the number of factually incorrect mentions in descriptions (3: None, 2: 1-2, 1: 3 or more) and wrote how they are incorrect. Two experts in each domain independently annotated 40 descriptions per expert (20 per each type), where 20 descriptions get two annotations (60 samples are evaluated in total per domain). Krippendorff's α is 0.55 for HCI paper descriptions and 0.41 for NLP ones.

4.4.4 Analysis

We analyzed the survey data (§4.4.3 and §4.4.3) by first conducting the Shapiro-Wilk test to determine if the data required a parametric (P) or non-parametric (NP) testing procedure and proceeded with a paired t-test (when parametric) and a Wilcoxon signed-rank test (when non-parametric) accordingly. In order to account for the between-subject variability in participants' notes data (§4.4.3), we employed a Mixed Linear Model Regression. The dependent variable in the model was each of the five dimensions of annotations, random effects were the participant IDs, and the fixed effects in the model were the experimental conditions. We used Restricted Maximum Likelihood (REML) for parameter estimation to ensure robust estimates of the variance and covariance parameters in cases when the homoscedasticity and normality of data assumptions are violated. We report on the results of the regression analyses as regression coefficients (β's) and p-values, with significance indicated at the α = .05 level. We also report the fixed effects sizes (semi-partial R^2, which represent the % outcome variance, controlling for the predictors and random effects in the model) when applicable.

4.5 Findings

Our results showed that PAPERWEAVER helped participants understand the relevance of recommended papers and make more effective relevance judgments. Additionally, PAPERWEAVER promoted the discovery of more connections between papers that were relevant to the topic of the folder, and led to writing more detailed notes that contained rich connections between papers.

4.5.1 General Behavioral Differences

During the study, we asked participants to think-aloud and observed how they interacted with the two systems. In both conditions, participants typically did a quick pass over the recommended paper list to filter out few recommended papers that seemed obviously irrelevant. When interacting with the baseline system, participants relied on the titles and the TL;DR summaries for this quick triaging process, while when interacting with PAPERWEAVER they additionally considered the *contextualized aspect-based description*. P2 mentioned that even though the two summaries were similar in length, TL;DR summaries typically focused on the research problems that were not always relevant to their folder's topic. In contrast, the *contextualized aspect-based description* provided by PAPERWEAVER often surfaced parts of the abstracts that were relevant. At the same time, participants in both conditions said that after this first pass they still needed to examine most of the papers more carefully to understand how they are relevant and to confidently judge which ones to save.

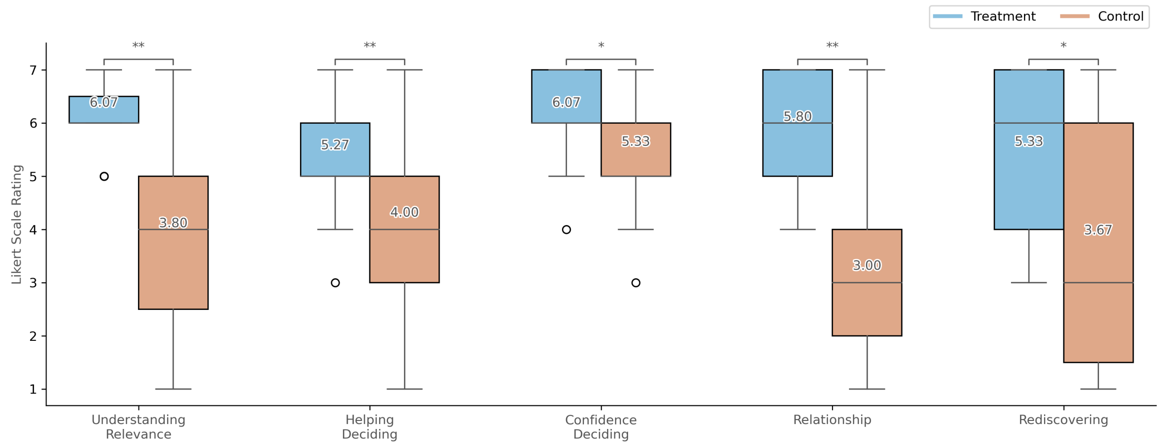

Figure 4.5: Results of the post-survey showed that participants found various benefits when using PAPERWEAVER compared to a strong baseline. **, *, and ns indicate significance of $p < 0.01$, $p < 0.05$, and $p > 0.05$, respectively.

When interacting with the baseline system, participants continued to read the full abstract or open the papers to see its figures but said this process was effortful. For example, P9 said that *"I tried to understand [the recommended] papers but reading all of the abstracts is overwhelming."* Further, even when they tried to carefully examine the recommended papers in the baseline condition, they often failed to identify the connections. For example, P1 mentioned that *"I cannot understand why this paper is recommended to me. It seems they are talking about just their own topic"* and P15 said that *"Abstract alone cannot answer 'how does this paper relevant to my research context?'*.

In contrast, when interacting with PAPERWEAVER, participants relied on *paper-paper description* after this first pass. In this case, many participants mentioned that they can better understand *"how this paper is different from the paper that they've already known"* and *"what new contributions are there [in the recommended papers]"* that were relevant to the collected paper. Interestingly, after reading the *paper-paper description*, participants often still continued to examine the abstracts. Participants mentioned the goal was to both verify the LLM-generated *paper-paper description* with the original sources and to "gain deeper context" around the *paper-paper description* once they became interested in a recommended paper (more details around this behavior in §6.4).

4.5.2 Exploring Papers Broadly by Understanding Relevance

Based on the post-survey, participants felt that they could understand how the recommended papers were relevant to their own research interest significantly better in PAPERWEAVER (M = 6.07, SD = 0.68) than the baseline (M = 3.80, SD = 1.64, p = 0.0013, NP). We also found evidence that PAPERWEAVER supported decision-making. Specifically, participants felt PAPERWEAVER helped them decide which papers were worth saving (PAPERWEAVER: M = 5.27, SD = 1.06, baseline: M = 4.00, SD = 1.55, p = 0.0024, P) and were significantly more confident in their decisions (PAPERWEAVER: M = 6.07, SD = 0.85, baseline: M = 5.33, SD = 1.07, p = 0.0124, NP). Qualitative insights revealed that these perceptions were due to how PAPERWEAVER contextualized explanations for each user based on their topical folder. According to our participants, both *contextualized aspect-based description* and *paper-paper description* effectively highlighted which parts of the abstracts or papers they should focus on (P9, P13). Specifically, while *contextualized aspect-based description* surfaced *"explicitly relevant aspects dispersed in multiple sections [in the paper, relevant] to my folder context"*. Participants also appreciated how *paper-paper description* connected recommended and collected papers with detailed and insightful

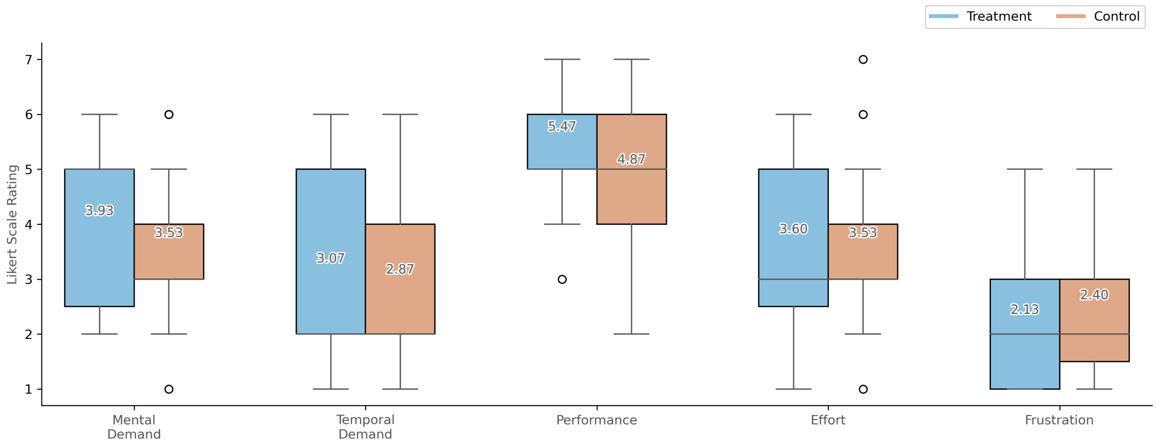

Figure 4.6: Based on a NASA-TLX survey, participants perceived similar workload when using PAPERWEAVER and the baseline conditions. **, *, and ns indicate significance of $p < 0.01$, $p < 0.05$, and $p > 0.05$, respectively.

relationships with *"narrower research interest perspective"* when compared to uncontextualized TL;DR summaries. For example, P15 mentioned that *"I don't have time to read all the abstracts but TL;DRs are too high-level. So with the [baseline], I do not have evidence to choose what to save. With the descriptions in* PAPERWEAVER*, I was more confident in my decision because I can get more understandable evidence from the explanations. For example, how the recommended papers [were relevant but] tackled different problems than the paper I knew already)"*.

4.5.3 Deeper Insights on Both Recommended and Previously Collected Papers

To gain a deeper understanding of the knowledge participants had gained from interacting with PAPERWEAVER and the baseline, we further analyzed the notes participants took during the study in the two conditions (while being blind to which conditions the notes came from). We found that when using PAPERWEAVER, participants on average captured significantly more notes that described connections between papers (PAPERWEAVER: $M = 2.21$, $SD = 1.81$, baseline: $M = 1.07$, $SD = 0.96$, $p = 0.0459$, NP). At the same time, when describing the relationships, participants included similar levels of details after accounting for between- and within-participant variability ($\beta = .113$, $p = .13$). This suggests that the bottleneck of learning connections between papers is a recall problem, and that it is easier to capture these connections with PAPERWEAVER compared to looking at multiple paper abstracts in the baseline condition. These results based on participant behavior also corroborate their perceived understanding of the information that was presented in the two conditions. Specifically, in the post-survey, participants reported that they could understand how the recommended papers connect to their collected papers significantly better with PAPERWEAVER (PAPERWEAVER: $M = 5.80$, $SD = 1.04$. baseline: $M = 3.00$, $SD = 1.67$, $p = 0.0021$, NP).

One possible trade-off for the positive learning effect could require significantly higher cognitive demands on the users. However, based on our NASA-TLX survey, participants did not report a higher workload when comparing using PAPERWEAVER and the baseline. On the other hand, qualitative data do show anecdotal evidence of higher mental and temporal demand for some participants when using PAPERWEAVER. For example, P6 and P9 pointed to descriptions from PAPERWEAVER that peaked their interest in the recommended papers and urged them to explore them in more detail. P4 further commented that *"I felt mentally more demanded using* PAPERWEAVER *because I actively compare the new papers with other [collected] papers. This might be a positive side effect of showing relationships."* These comments suggest that even for participants who felt using PAPERWEAVER was more cognitively demanding, they perceived it positively because it prompted them to become more actively engaged with the paper alert, motivating them to more deeply process the information in front of them.

Finally, in addition to understanding the recommended papers, our analysis of the post-survey data also revealed that participants felt PAPERWEAVER helped them *"learn something new about the papers that they have already saved in the folder"* significantly more than the baseline (PAPERWEAVER: M = 5.33, SD = 1.49. baseline: M = 3.67, SD = 2.18, p = 0.011, NP). In the interviews, participants pointed to different benefits when seeing descriptions that covered relevant collected papers in the folders, including refreshing their memories about papers they had previously read and gaining new knowledge and perspectives about them based on how recommended papers described them in their related work sections. In addition, we also found some participants who had saved papers in their folder that they had not yet read, in this case, they describe how the descriptions from PAPERWEAVER helped them *rediscover* previously collected papers with a *renewed interest.*

4.5.4 Accuracy of LLM-Generated Descriptions

One known issue with current LLMs is that they can be prone to hallucinations (*cf.* [230, 14, 19]), and we did find that some descriptions contained one or two mistakes based on manual evaluation. In general, *contextualized aspect-based description* are more extractive and contain fewer mistakes than *paper-paper description* which relies on training knowledge to align the extracted aspects. Specifically, for *contextualized aspect-based description*, 8% of samples has one or two mistakes (HCI: 6%, NLP: 10%, based on 30 samples each), and *paper-paper description*, 20% (HCI: 16 %, NLP: 23%, based on 30 samples each). One fundamental question here is that - is this an acceptable level of accuracy in the context of paper alerts? Results from our user study around participants' strategies when interacting with PAPERWEAVER offered some insights.

Firstly, like most retrieval systems, results from document recommender systems can also contain errors (i.e., documents falsely classified as relevant) [43]. Because of this, we found that when interacting with paper alerts, participants were already in the mindset of verifying the machine-generated document recommendations, in both conditions. We also found participants go through the list of recommended papers in multiple passes, typically using the titles first to rule out clear irrelevant recommended papers, then reading the generated descriptions in the second pass, and finally verifying them with the abstract or the content of the papers. For example, representative quotes from P6 and P8 mentioned *"validate the connections between them [papers] in the abstract"* and P2 provided more details around their strategy: *"[I read abstracts] not only to see deeper levels of information but also to verify the content of the descriptions."*

Secondly, many participants also explicitly mentioned that they see the PAPERWEAVER descriptions as "*supplementary material*" that help them triage the recommended papers, suggesting that participants could appropriately adjust their levels of reliance on the LLM-generated descriptions. P2 said *"I used the descriptions from [*PAPERWEAVER*] as supplementary material going from titles and TL;DR to abstract,"* representative quote from P8 additionally pointed to benefits around awareness and discovery "PAPERWEAVER*'s descriptions triggered me to become interested in two highly relevant [recommended] papers and have curiosity about them so that I can [decide to] read the abstracts and [then the] whole papers to get more information. These descriptions are supplementary bridges from the titles and TL;DRs to the abstracts rather than replacing them."*

Finally, results based on post-survey and NASA-TLX questionnaire suggested that PAPERWEAVER did not increase perceived workload even though participants were actively verifying LLM-generated content and that they were able to judge the relevance of the recommended papers more confidently and captured richer relationships between papers in their notes. We acknowledge that our participants were computer science researchers who currently might be more familiar with recent advancements in LLMs than researchers in other domains, although the increasing popularity of LLM-powered end-user tools also provides them increasing opportunities to interact with LLMs in other scenarios. Nevertheless, our results suggest that interfaces that make use of LLM-generated text should always provide both adequate indication when generated text is presented to its users and allow users to freely turn on or turn off generated content. Moreover, it is essential to design effective mechanisms for users to verify content efficiently. Inspired by AngleKindling's system design [185] that shows the connection between LLM-generated angles and source material to support journalists, one approach involves making specific sections of the paper, particularly relevant to the LLM-generated descriptions, accessible—such as through highlighting or indicating whether there is clear citations with a hyperlink to source papers. We also can indicate whether there is clear source material that the output is grounded on by marking citations at the end of each sentence.

4.6 Discussion

Limitations We conducted our study with computer science graduate students who might not represent the broad spectrum of academic domains. While our approach is motivated by a general use case and designed to accommodate a broad range of academic domains, further evaluation with researchers from other academic domains, especially less technology-oriented ones, can help us understand how the generated descriptions can be applied broadly. Another limitation is our use of the Problem-Method-Findings schema. While this schema fits a large portion of papers, it might not cover some other papers such as survey papers or systematic review papers [85]. Further, researchers might want to apply their own schema (e.g., medical researchers would be interested in specific aspects of clinical trials). We discuss how to extend the schematic digests beyond Problem-Method-Findings in §4.6. Finally, the measurements in our user study are more focused on subjective measures because it is inherently challenging to get robust and valid measures that score participants' understanding of the relevance between papers and measure their triaging performance due to the evolving notion of relevance personalized to each individual. Standardization of this notion would require careful examinations of prolonged interaction scenarios beyond the scope of the laboratory experiment conducted here. While we obtained objective metrics in the experiment that can be evaluated independently of each researcher's context, as a future work, conducting a longitudinal deployment study could involve participants assessing their own written outputs over time, allowing for evaluations that consider evolving individual contexts.

Pairwise vs. Multiple-Paper Descriptions In this work, we focused on describing pairs of recommended and collected papers to help users contextualize unfamiliar papers with familiar papers. In our current design, each description only covers two papers: a recommended paper and a collected paper. This design decision was based on our formative study where we observed that participants with different familiarity with the topics all found benefits in seeing descriptions that covered two papers, but less experienced participants felt that descriptions covering three or more papers were too complex for paper alerts. Future work could explore ways to allow users to further customize their paper alerts by adjusting the complexity of the descriptions. One opportunity for future work in this direction is to look to research that focuses on automatic related work section generation [180] or multi-document summarization [74] that aimed to generate descriptions for many documents. One interesting and relevant use case we observed in the user study was one participant who had saved a survey paper in their folder. In this case, the description PAPERWEAVER generated allowed participants to compare and contrast recommended papers with different research threads that were described in the survey paper's abstract and was perceived positively by the participants.

Longer-Term Usage and Evolving Folder Descriptions In the user study, we observed that participants collected richer notes that captured information that connects multiple papers as opposed to about a single paper. Since PAPERWEAVER leverages a user's folder name and description in its prompts to generate a description that can better reflect a user's knowledge about the folder topic, an interesting future direction is to allow users to update their folder name and descriptions based on what they had learned in a paper alert. This information can be used to update the folder description to represent the user's current knowledge and this can, in turn, improve subsequent paper alert generations. In this sense, the user's folder can serve as an evolving external representation of the user's understanding on a specific research topic. However, the longer-term effects of accumulating additional notes need further investigation.

Extending Schematic Digests beyond *Problem-Method-Findings* Participants in our studies have also commented on how they would like to further customize the information presented in paper alerts, pointing to avenues for future work. First, some participants commented that they wanted to be able to surface information along certain other aspects of the schema, such as differences or similarities in evaluation regimes and their outcomes (*e.g.,* positive, negative); what types of study designs were run and how they were conducted (*e.g.,* controlled lab studies, field deployment studies, RCTs); approaches to developing AI models in a given problem domain; and the design of interaction features in proposed systems. Such aspects of schemas were customized to different participants, suggesting that while the default information provided in PAPERWEAVER's problem-method-findings schema served as a useful entry into recommended papers, a deeper inspection following users' triage would benefit from further digesting papers along user-defined secondary schemas.

However, the first-and-secondary division of schemas that adapts to users' interaction with paper digests over time also suggests that there is an important gradient of specificity that may have to be designed to be adaptively adjustable based on user interaction, for example by utilizing a form of passive sensing over users' intent based on their interaction. Clear examples of this are when participants stated that they wanted to "see more" in the aspect-based paper summaries, demonstrating an intent for lower-level details; another intent that participants may express would be wanting to "adjust" an explanation provided for a recommended paper, for example, as pointed out by P4, when the problem-method-findings schema in our approach did not provide useful information for survey papers because the problem and method descriptions were presented at too high of a level, abstracting away useful details of any individual papers or groups of papers synthesized by the authors of the survey paper.

Contextualized Paper Descriptions beyond Paper Alerts Finally, while we focused on the scenario of helping users make sense of paper recommendation alerts, the proposed pipeline can potentially be generalized to other scenarios where users need to make sense of unfamiliar papers. For example, using one's publications as "collected papers" and generating descriptions for papers from another author to explore common research interests and facilitate collaborations. Another opportunity is to enrich a user's experience when *reading* a related work sections in a paper (*cf.* CiteSee [32], Threddy [98], CiteRead [188]) by generating alternative descriptions about the cited papers based on papers already familiar to the current user. Beyond comparing with existing papers, users can also compare their own draft of a paper with new papers that might be relevant to them to get insights or new perspectives for framing the contrast and comparisons between papers when organizing related work sections.

4.7 Conclusion

This work presents PAPERWEAVER, an enriched paper alert system that provides contextualized text descriptions of recommended papers based on user-collected papers. PAPERWEAVER leverages an LLM-based computational method to infer users' research interests from their collected papers and extracts contextualized aspects in recommended papers that are relevant to the inferred user's interest. Our method establishes relationships between recommended and collected papers by comparing and contrasting these contextualized aspects. Through a within-subjects user study (N = 15), we found that PAPERWEAVER helped researchers more confidently make sense of paper recommendations and discover more useful relationships between recommended and collected papers in the folder when comparing with a baseline interface that enriched existing paper alert systems with abstract summaries and extracted related work sections.

Chapter 5. QASA: Answering Advanced Questions on Scientific Articles.

This chapter presents the first example of evaluation methods that assess how well AI models align with human cognitive process. This chapter proposes the QASA: a novel benchmark that evaluates AI models' question-answering capabilities on scientific articles. This chapter has adapted, updated, and rewritten content from a paper at ICML 2023 [129]. All uses of "we", "our", and "us" in this chapter refer to coauthors of the aforementioned papers.

5.1 Motivation and Contributions

Reasoning differentiates human intellectual capabilities from low-level intelligence. Dual process models theorize that cognitive reasoning is a two-stage process where the first stage performs associative thinking and the second stage performs logical reasoning [242, 233, 60]. Within the context of Question Answering (QA), the first stage extracts associative pieces of knowledge based on lexical matching and other cognitive heuristics, inductively expanding potential evidences. Then the second stage consciously finds evidential rationales, deductively converging to the answer via systematic compositions of the evidences. This process uniquely characterizes advanced human reasoning, posing a non-trivial challenge to machine learning QA systems.

Reading Comprehension (RC) is one type of reasoning task that can formulate various questions and answers. SQuAD [193], NewsQA [232], DROP [58], and Natural Questions [120] have been proposed. While competing on their model performance significantly improves machine answering capabilities, these datasets consist of factoid QAs mostly in the form of "what", "when", "where", or "who". Thus extracting short spans from the relevant context can easily provide correct answers, but the trained models can barely answer "how" and "why" questions.

Recent work on open-domain QA [104, 79, 149, 91, 92] exploits the *Retrieve-then-read* approaches, where the system first retrieves relevant documents from a large corpus then reads out concrete answers. These approaches target shallow questions that are often inferable relying only on the first stage rather than jointly using the both stages. Some reasoning tasks like bAbI and its permuted version [244, 191] require logically correct spatial reasoning. However, the artificial nature of their QAs minimally leverages the second stage as their reasoning tasks do require neither rich retrieval of associative information from the first stage nor systematic composition of the final answers [124].

Our think-aloud study reveals that reading scientific articles not only raises surface questions but also induces testing and deep questions that require full-stack reasoning. In addition, carefully answering surface questions turns out to involve both first and second stage reasoning, requiring significantly more elaborated efforts compared to what previous datasets and models implicitly assumed. To answer for such naturally advanced questions, we propose the **Q**uestion **A**nswering on **S**cientific **A**rticles (**QASA**), a novel QA benchmark and an approach that realize the full-stack cognitive reasoning from the first to the second stages. Our QASA benchmark differs from existing ones on the following aspects:

- Based on our think-aloud study, we design a schema for advanced questions as *surface*, *testing*, and *deep* questions, then collecting balanced QA pairs form the authors of research papers as well as from expert readers.
- We guide readers and authors to ask questions while reading the *whole paper* rather than gathering only extractive questions from paper abstracts.

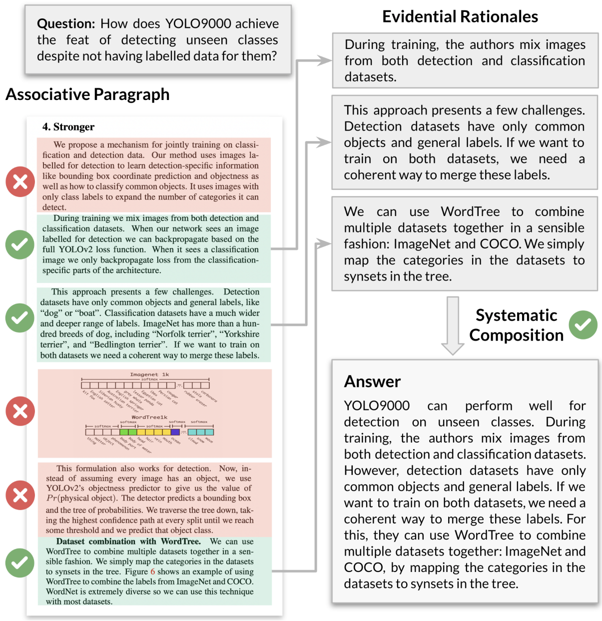

Figure 5.1: An example of QASA. A question that the reader/author asks about the paper while reading the paper. To formulate the answer, one classifies whether the paragraph contains evidence to answer the question. Evidential rationales are written for each evidential paragraph and are systematically composed into a comprehensive answer.

- Readers and authors are asked to propose their *multifaceted long-form* answers to the collected questions, then *composing* a comprehensive final passage than simply summarizing evidential rationales with added fluency.

Our QASA benchmark contains 1798 QA pairs on AI/ML papers where the questions are asked by regular readers of AI/ML papers and answered by AI/ML experts. Each paper has 15.1 questions on average, up to a maximum of 29 questions for a single paper. We collect 39.4% of deep reasoning level questions based on our own question schema. And, maximum 9 evidential rationales are leveraged to compose the final answer.

Our QASA approach models the full-stack reasoning process via state-of-the-art large language models. We decompose the process into three subtasks: *associative selection* (to extract relevant information from paragraphs), *evidential rationale-generation* (to grasp only evidential rationale from each extracted paragraph), and *systematic composition* (to stitch evidential rationales into a comprehensive answer without redundancy). Modeling each subtask by pretrained large language model with existing datasets, we demonstrate that our best test-bed outperforms the state-of-the-art InstructGPT (OpenAI's text-davinci-003) by 5.11 Rouge-1 points. We further verify that directly generating an answer from selected paragraphs causes performance drop, opening a crucial insight for tackling advanced question answering.

The relevant research consists of three categories: QA for academic research papers, long-form QA, and query-based multi-document summarization. Table 5.1 highlights our method against existing approaches in each groups.

5.2 Proposed Task

In this section, we propose a new task for question answering over scientific articles. The core idea of our proposed task is to answer the questions based on multiple evidence snippets that are spread over a long research paper. Specifically, we denote a question as q, an answer as a, and paragraphs in the paper as $P = \{p_1, ...p_N\}$. A one-step approach to process N paragraphs would be adopting length-scalable transformer such as LongFormer

Method	Associative selection	Evidential rationale-generation	Systematic composition
QASPER	✓	✗	✗
ELI5	✗	✗	✗
ASQA	✓	✗	✓
AQuaMuSe	✗	✗	✓
QASA (ours)	✓	✓	✓

Table 5.1: Comparison of existing datasets and our QASA.

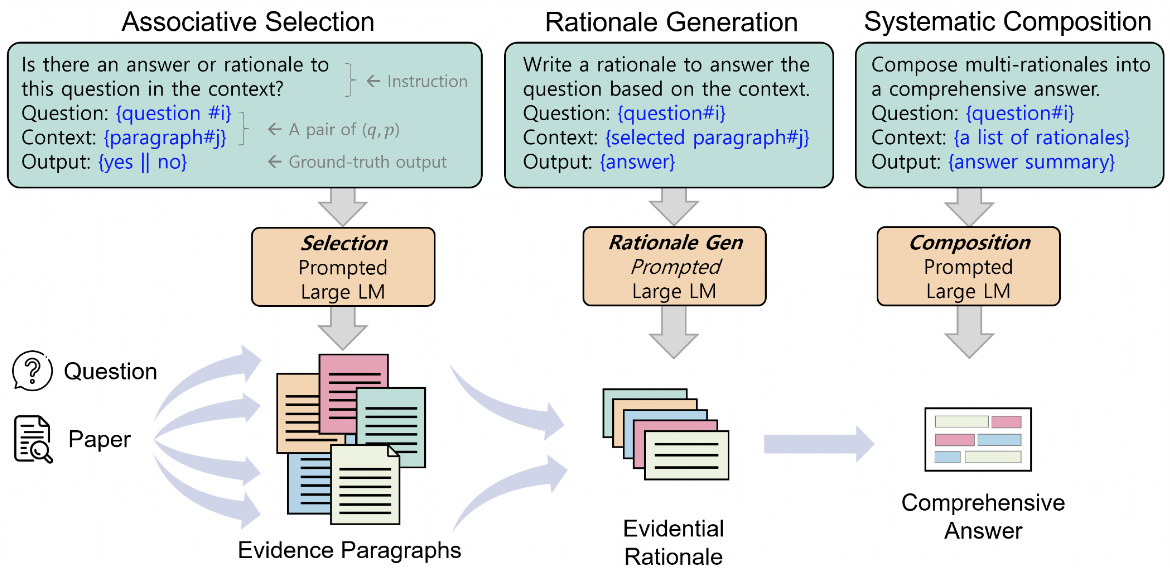

Figure 5.2: An overview of QASA approach. The language model works depending on task-specific instructions.

[18], which enables to encode the multiple snippets at once. In contrast, our advanced questions triggered from research papers requires to connect between rationales for deep reasoning. Hence, we design this problem as multi-step subtasks: (1) *associative selection*, (2) *evidential rationale-generation*, and (3) *systematic composition*. Figure 5.2 shows the overview of our approach.

Associative Selection While research papers have multiple paragraphs (*e.g.,* , 20-60 paragraphs), the first step is to extract associative knowledge from the paragraphs, corresponding to a question. Specifically, given question q and paragraphs $P = \{p_1, ...p_N\}$, we aim to select evidential paragraphs $\bar{P} = \{\bar{p}_1, ...\bar{p}_k\}$ that contains an answer or rationales to question q, where $k \ll N$. While the previous work [192] aims to classify answerability whether a given passage contains answer a to question q, the answer in our task is composed of multiple rationales including a main answer. Our associative selection task can be viewed as the super-task of answerability (*i.e.,* , answerable is evidential, but not the reverse).

Evidential Rationale-Generation In this step, we generate an evidential rationale on each selected paragraph, which could be part of a final long-form answer in the next step. Based on the prior work about discourse structure of answers to complex questions [251], the evidential rationale can be the (1) main answer (*i.e.,* , main content of the answer which directly addresses the question), (2) elaboration (*i.e.,* , sentences which elaborate on the main answer), and (3) auxiliary information (*i.e.,* , background knowledge that could be helpful to the user). Specifically, we denote the evidential rationale that is inferred from (q,p_i) as e_i. That is, from the selected $\bar{P}$ = $\{\bar{p}_1, ...\bar{p}_k\}$, we obtain a list of evidential rationales $\{e_1, ..., e_k\}$.

Systematic Composition To provide concise and readable information to users, the goal of this last step is to systematically compose all the evidential rationales $\{e_1, ..., e_k\}$ into a final comprehensive answer a. Assuming that the answer is composed of multi-rationales, we aim to preserve all the rich rationales in the final answer, except duplicated texts. Specifically, we aggregate the list of texts $\{e_1, ..., e_k\}$ into a single context, and then compose final answer a from the context. The answer a grounded on a given paper could be viewed as comprehensive explanations about question q.

5.3 Building the QASA Dataset

Prior to data collection, we conducted a preliminary study for identifying what kinds of questions are raised when reading papers. Based on our findings, we design a schema to collect diverse and balanced questions with different levels of reasoning. As the source of the QASA, we gather a set of open-access AI/ML papers. To collect advanced questions that require reasoning over evidential rationales, we recruited AI/ML practitioners or researchers who regularly read research papers and conducted two separate sessions from the perspective of both *readers* and *authors*.

5.3.1 Preliminary Study

In the aim of identifying what kinds of questions readers ask while reading, we conduct a think-aloud study (N = 10), a standard approach in human-computer interaction (HCI) for capturing human's intent during a task. Our analysis of 127 questions revealed that 67% of questions required a two-stage process, and the types of reasoning needed in the second stage varied even among these questions. Referring to the prior literature on question taxonomy in the education domain, there are distinct types of questions ranging from surface questions to deeper questions that require more reasoning and interpretation to answer [76, 77]. To gather diverse and balanced types of questions, we design a schema for paper questions by adapting the prior literature in education to a paper reading context and interpreting data collected from our think-aloud study. This schema includes not only questions requiring second stage reasoning, but also a spectrum of reasoning types needed to answer them. The definitions of each question type are shown below and detailed explanations with examples can be found in Table 5.2.

- **Surface questions** aim to verify and understand basic concepts in the content. The answer content is directly related to the words in the question and immediate context. This type includes *verification, distinctive, concept completion* questions.

- **Testing questions** are focused on meaning-making and forming alignment with readers' prior knowledge. These questions aim to find similar examples (*example*), quantify variables (*quantification*), and find meaning and make comparisons across concepts (*comparison*).

- **Deep questions** ask about the connections among the concepts in the content and elicit advanced reasoning in logical, causal, or goal-oriented systems. This type includes *causal antecedent, causal consequence, goal orientation, instrumental/procedural, rationale, expectation* questions.

5.3.2 Papers

To collect papers, we adopt S2ORC [151], a collection of machine-readable full text for open-access papers, and the arXiv[1] paper collection. We only use papers within the CS.AI domain in the arXiv dataset and apply two filtering criteria to the papers in the S2ORC collection: (1) published after 2015 and (2) has more than 100 citations.

5.3.3 Data Collection

With the aim of collecting various advanced questions (surface to deep), we conduct two types of sessions, reader sessions where we collected QAs from general readers and author sessions where authors annotated questions about their own papers. We perform author sessions since authors are the optimal annotators who can make challenging and insightful questions that could be asked by experts, like reviewers—granting greater diversity to the questions in our benchmark. For the reader session, to make the data collection process similar to a real context, we decouple the questioning and answering phase following the collection process of QASPER [49]. For both tasks, we recruited graduate students studying AI/ML and freelancers practicing AI/ML through professional networks and Upwork.[2] For the answering task, we qualify annotators through the exams related to our task and experience in the domain.

Questions To ensure that our questions are realistic, we allow annotators to choose papers that they wanted to read. Additionally to replicate differing reading styles, we asked annotators to follow one of two scenarios: read all the sections in the paper (*i.e.,* , deep reading) or read only certain sections (*i.e.,* , skim reading). To collect diverse types of questions, we provide them with the question schema and asked them to make a balanced number of questions for each type. In the same vein, we also recommend annotators to make at least one question per subsection that they read. When annotating questions, annotators were instructed to write the trigger sentences that raised the question but that did not contain the answer. While they were not used in this work, trigger sentences could be used in future research for question generation from long-form text and to complement the ambiguity of questions that occur in long-form text.

Answers To collect answers, we ask answerers to choose papers from the papers that the questioners worked on. We guide answerers to compose their answers into a comprehensive passage based on the their own-generated evidential rationales from the selected paragraphs. To let them follow our guideline more easily, we provide the annotators with the answering interface when answering the questions. They were shown the question, the full paper, the name of the section that triggered this question, and ten paragraphs that are the most relevant to the question. We provide top ten paragraphs by following that existing IR research [30] adopted a pooling method, where top ranked documents are selected to create the pool of documents that need to be judged when creating evaluation dataset. Our top-10 relevant paragraphs were chosen with an off-the-shelf embedding model.[3] When answering each question, annotators were asked to do the following subtasks.

First, they are asked to look through the ten relevant paragraphs and, for each, make a binary decision as to whether the paragraph is evidence paragraph. If there is no relevant paragraph chosen to have evidence, annotators could freely choose other paragraphs from the paper as having evidence in addition to the ten paragraphs we provided. Second, for each paragraph that was chosen, annotators are instructed to write evidential rationale from that paragraph. Evidential rationale could be the (1) main answer to the question, (2) elaboration, or (3) auxiliary information [251]. Third, they write a final comprehensive answer by composing the multiple evidential rationales

[1] https://arxiv.org/help/bulk_data
[2] https://upwork.com/
[3] https://api.openai.com/v1/embeddings

that they generated for each evidence paragraph. When the answer cannot be fully answered even after composing multiple evidential rationales, annotators are instructed to answer as much as possible with the available information and then specify which part of the question cannot be answered. When a question is completely unanswerable, we ask annotators to indicate that the author do not provide an explanation for the missing information and to specify what information is missing. Finally, they annotate whether writing the final answer requires to compose multiple evidential rationales (True) or not (False)—*i.e.,* , no complex reasoning is needed and they only simplify text from the paper without adding redundancies.

All annotators who ask questions and write answers conducted a practice session to familiarize themselves with the annotation guidelines. Annotations from the practice sessions were reviewed by two authors, and discrepancies between the annotators and the guidelines were discussed. Additional practice sessions were conducted for annotators with substantial discrepancies. If annotators were judged as not having sufficient background knowledge or understanding of the task even after these sessions, we did not let them participate in the tasks.

Authors We recruited paper authors to annotate QAs for their own papers to cover deeper questions that existing datasets rarely cover. We instructed authors to make only **testing** and **deep** types of questions and to annotate trigger sentences that might cause readers to become curious about that question. The rest of the annotation process is similar to the readers' sessions. We recruited 17 authors whose domains are distributed in CV, NLP, GNN, generative models, and music information retrieval.

5.3.4 QASA Analysis

Representative examples from QASA is in Table 5.2.

Question types In terms of question types, two domain experts manually evaluated 100 randomly sampled questions. 89% of the annotated question types were aligned with domain experts' annotations.[4] To describe the diversity of our dataset, we analyze the distribution of the types of the questions in QASA. Among the three types, 39.4% of the questions are deep questions, 30.0% are testing, and 30.7% are surface-level. Among the deep questions, instrumental sub-type (12%) accounts for most of the deep questions, and comparison (11%) and concept completion (17%) are the most annotated questions for the testing and surface questions, respectively.

Distribution of evidential rationale We also analyze to identify the number of evidential rationales that are needed to answer the questions depending on their types. Among all the questions, 12% of questions are annotated as having no evidential rationales, which means that they are unanswerable questions. Out of the answerable questions, the average number of evidential rationales is 1.67. Surface questions need the most evidential rationales (1.73) while testing questions and deep questions need 1.66 and 1.63, respectively, which implies that our surface-level questions like *"Do the authors claim that bigger datasets would improve the performance and expressiveness of reading comprehension models?"* also need systematic reasoning to answer.

Composition, Correctness, Groundedness On average, 49.6% of answers require annotators to compose evidential rationales, while the rest (50.4% of answers) only need simplifying redundant rationales. To analyze which question type requires the most reasoning to answer, we analyze the ratio of compositionality depending on the question type. Deep questions need composing the most (44.6%) in comparison with testing (29.0%) and shallow (26.4%) questions. To estimate the correctness of the answer annotations and groundedness of the answer

[4]Two domain-experts independently judged these and achieved Cohen's κ scores of 0.91.

Table 5.2: Examples from QASA.

Type	Question	Evidential rationales	Answer	Composition
Surface question	“Do the authors claim that bigger datasets would improve the performance and expressiveness of reading comprehension models?”	[“Mentions that previous approaches in the literature that attempt to use synthetic dataset approaches (which can quickly be used to generated datasets of large sizes) have shown promise for reading comprehension tasks”, “Authors mention that creating a large-scale labelled dataset as one of their contributions.”, “Indicates that attention based models and LSTM models are effective at reading comprehension tasks.”, “Explains that they (the authors) intend to provide a dataset that could be used to benchmark performance on reading comprehension tasks”]	“Based on the information in this paper alone, it is unclear if a bigger dataset would improve the performance of reading comprehension models. While authors explain that a key contribution they make is the creation of a real-world, massive labelled reading comprehension dataset, it is unclear if such a dataset is essential to improve the performance of reading comprehension models - the authors pitch their dataset-building approach also as a way of evaluating performance of these models, which is different from the dataset itself leading to better performance.”	True
Surface question	“Which deep neural network architectures were used for experimental comparison of DeepFool algorithm with existing methods?”	[“The experiments for the DeepFool method involve 6 different classifiers for 3 different datasets: MNIST: 2-layer FC, 2-layer LeNet; CIFAR-10: 3-layer LeNet, NIN; ILSVRC 2012: CaffeNet, GoogLeNet.”, “The paper provides extensive experiments with 3 different datasets and 8 different classifiers.”]	“Although the conclusion of the paper claims that 8 different classifiers were used, we can only see 6 classifiers with different datasets: 2-layer fully-connected network (MNIST), 2-layer LeNet (MNIST), 3-layer LeNet (CIFAR-10), NIN (CIFAR-10), CaffeNet (ILSVRC 2012), and GoogLeNet (ILSVRC 2012).”	False
Testing question	“What is the difference between BERT paper and RoBERTa paper’s point of views? Give an answer in NSP loss and their performance perspective.”	[“In BERT paper, author said that removing NSP can hurt the performance of the model.”, “In RoBERTa paper, author said that removing NSP improves downstream task performance.”]	“In BERT paper, author said that removing NSP can hurt the performance of the model. However, in RoBERTa paper, author said that removing NSP improves downstream task performance. Therefore, point of views in terms of NSP is different between BERT and RoBERTa.”	True
Testing question	“What does "interaction between the pixels to the text embedding through the diffusion process" mean?”	[“During the diffusion process, we predict the noise of an image given a noisy image and text embedding using U shaped network. and this process yield our final image at the last step. and the interaction between the two modality occurs during the noise prediction. And the interaction between the visual and textual features are fused using Cross-attention layers that produce spatial attention maps for each textual token.”, “Authors consider using internal cross-attention in their work, and the cross-attention maps are high-dimensional tensors that bind pixels and tokens extracted from the prompt text.”]	“To answer this question we need to recall the diffusion process, which is in order to predict the noise of an image we have two inputs 1- noisy image and 2- text embedding, and the interaction between the two inputs are fused using Cross-attention layers that produce spatial attention maps for each textual token. and that is what is meant by the interaction between pixels to text embedding.”	False
Deep question	“When defining the reading comprehension task, the authors explain that they wish to estimate $p(a \mid c, q)$. What would a model trained on this task do if the context ’c’ itself had factually incorrect information?”	[“This paragraph explains that the authors wish to estimate the conditional probability of an answer (a) being relevant to a question (q) given some context (c).”, “Explains that the proposed dataset’s objective is to test if a model is able to read and comprehend a document - not test a model’s global knowledge of whether a statement is true or false.”, “Mentions that the task they are building the dataset for is a reading comprehension task.”]	“The authors are training a reading comprehension model. Therefore, if the context “c” has incorrect information, the model is likely to answer based on the factually incorrect information itself. The authors clearly explain that the task their model is being built for and evaluated on is of identifying answers from a given text (i.e. comprehension) and not knowledge of global correctness.”	True
Deep question	“What weaknesses would a dataset that without entity replacement or anonymization have when training a reading comprehension model? Why is this a necessary step in the process?”	[“Explains the difference between an original (unprocessed) data point and anonymized sample through an example. This paragraph points out that the non-anonymous version of the query could potentially be answered by an agent (either human or an ML model) even without reading the paragraph/context, while that would not be possible post-anonymization. This change ensures that the metric being measured is reading comprehension only and not anything else.”, “Explains that they replace all entities with an abstract entity marker.”]	“Since the authors are attempting to build a reading comprehension model, not anonymizing the entities before using the dataset might lead to a situation where models use external information, or statistics on the distribution/frequency of words themselves to guess answers. These steps are needed to ensure that models use the context to answer the questions.”‘	False

annotation, domain experts manually analyzed 100 randomly sampled questions. We find that 90% of the answers are correct and 87% are grounded well on the paper.

5.4 QASA Approach

In this section, we propose a QA approach for QASA over research papers. Our task requires to answer questions based on multiple passages whose supporting evidences are spread over a whole paper. As above-mentioned, our tasks consist of (1) *associative selection*, (2) *evidential rationale-generation*, and (3) *systematic composition*. As shown in Figure 5.2, we train LM models with multi-task instructions, following recent works [41, 243, 7, 207].

5.4.1 Multi-step QA system based on LM

Pre-processing via Retrieval Before the first step of *associative selection*, we consider pre-processing step using a retrieval model to narrow the search space, from a whole paper to top-N related paragraphs (we set N=10). This enables the efficient selection step, while compromising the recall of evidential paragraphs. Specifically, we used the off-the-shelf model provided by OpenAI, and leave the question of improving retrieval for future work. Through the retrieval, we encode all paragraphs in the given paper and a target question into dense vectors, and extract top-N nearest neighbor paragraphs by using cosine similarity.

Finetuning Large Language Model with Multitask Instructions We finetune large language models (LLMs) on a mixture of our subtasks through instruction tuning [243]. As in previous work [243, 7], it is known that instruction tuning makes LMs generalizable on unseen tasks. As shown in Figure 5.2, a single LM takes task input with instruction that indicates each subtask. The output of the previous step is sequentially passed to the next step. However, in the *selection* task, if the model does not select any paragraph as evidence, it also cannot generate rationales or answers. Instead of solving this problem, we used top-3 paragraphs if none were selected, which is left to future work. For task-specific prompts, we used manually-written instructions for each subtask. As state-of-the-art LLMs, we consider the following models:

- T5 [190] (Version 1.1, LM-Adapted): it is pretrained on Common Crawl [190] using Transformer with encoder-decoder architecture.
- T0 [207]: starting from T5, it is further trained on 8 downstream tasks.
- FLAN-T5 [41]: similar with T0, it is further trained with scaling up multi-tasks (1k+) including reasoning tasks.
- GALACTICA [226]: it is pretrained on a large collection of scientific papers, with the decode-only architecture like GPT.

5.4.2 Training Data

No training resources have been proposed that support our full-stack QA, and we therefore exploit public and synthetic data for the purpose of each subtask. Table 5.3 shows a summary of used public data.

For *associative selection*, we adopt answerability labels – whether the pair of question and knowledge is answerable or not. In case of ASQA [220], we treat the pair of (q and p^+) as a positive example, and (question q, randomly-sampled p^-) as negative. For QASPER [49], we leverage pairs of (question q, gold paragraph p^+ that

Task	Dataset
Associative Selection	QASPER, ASQA
Rationale Generation	QASPER
Answer Composition	ASQA, ELI5

Table 5.3: Training Resources for our QA system.

contains an answer). The limitation of adopting these datasets is that they aim to capture the presence of an answer, while we target that of evidential rationales, which may affect recall of rationales.

The *rationale-generation* task requires to generate evidential rationale e from (question q, paragraph p). Unfortunately, to the best of our knowledge, there is no data to support this task. As an alternative source, we used the triplets of (q,p,a) in QASPER [49], where gold knowledge p always contains information about answer a to q. We treat answer a as evidential rationale, since the QA labels in QASPER do not require to reason over multiple passages, which is expected to learn the ability of extracting question-focused evidences from a given paragraph.

Lastly, for *systematic composition*, we adopt long-form QA data with multiple evidences. We select ELI5 and the subset of ASQA, which provide selected evidence passages from the pool of passages for answer generation. That is, this task is to generate answer a inferred from the context (question q, multi-evidences $\{e_1, ...e_n\}$), which requires to consolidate and summarize scattered information.

For synthetic data, recent works distill training data from InstructGPT [240, 86]. Inspired, we distil training examples for QASA from large language models by prompting instruction and an in-context example. We use OpenAI's InstructGPT (text-davinci-003) with the temperature set to 0.1, which is the state-of-the-art model on many NLP tasks. Specifically, we first extract AI and ML papers from arXiv, and generate questions over each paragraph sampled from the papers. Then, given the questions, we test InstructGPT following instructions in our subtasks. While LLMs have a general problem of factual inconsistency, known as *hallucination*, we found that InstructGPT performs well on the rationale generation task (See Table 5.4). Although there is no public data to support rationale generation, we can alleviate the insufficiency through evidential rationales obtained from InstructGPT, which would boost our full-stack QA.

Method	**Associative Selection**			**Rationale Generation**			**Answer Composition**		
	(P)	(R)	(F1)	(R-1)	(R-2)	(R-L)	(R-1)	(R-2)	(R-L)
Pretrained LMs (Accessible Checkpoints or API)									
GALACTICA(6.7B)	18.70	94.28	29.36	7.06	0.50	5.01	8.93	1.03	6.84
T5 (3B)	6.85	6.83	5.78	26.99	11.31	20.64	27.46	16.08	21.85
T0 (3B)	6.92	7.60	6.39	20.19	9.75	17.71	32.75	20.49	29.30
FLAN-T5 (3B)	37.50	38.57	34.64	20.30	11.62	18.36	40.90	27.30	35.78
INSTRUCTGPT (175B)	31.78	51.97	34.72	**41.27**	**24.69**	**33.64**	47.27	28.22	36.09
Finetuned LMs (on Collected Data)									
GALACTICA (6.7B)	31.70	47.39	33.32	8.45	1.07	6.98	13.90	2.44	10.41
T5 (3B)	39.79	56.56	40.71	26.73	13.02	22.64	46.40	29.60	38.55
T0 (3B)	39.04	77.29	45.16	27.86	13.40	23.45	46.78	29.29	38.24
FLAN-T5 (3B)	37.13	59.97	**45.86**	27.63	13.65	23.33	45.59	28.80	37.24
FLAN-T5 (3B, QASPER-ONLY)	30.67	91.43	41.67	23.47	13.13	20.94	40.51	26.88	35.57
FLAN-T5 (3B, ASQA-ONLY)	21.39	**98.97**	32.84	24.72	13.90	21.62	**48.97**	**31.93**	**40.34**
FLAN-T5 (3B, ELI5-ONLY)	**46.80**	36.16	40.80	21.03	11.59	18.74	33.24	18.17	28.59
FLAN-T5 (3B, GPT AUG-ONLY)	31.58	28.82	27.85	29.22	13.69	24.15	48.53	31.19	38.80

Table 5.4: The results of baseline systems on three subtasks in QASA, measured by Precision, Recall, F1 score, and ROUGE scores. The best results in each column are **bold**-faced, and 2nd best results are underlined.

Method	Full-stack QA		
	(R-1)	(R-2)	(R-L)
Pretrained LMs (Accessible Checkpoints or API)			
GALACTICA (6.7B)	15.56	3.65	11.44
T5 (3B)	9.83	0.58	8.01
T0 (3B)	15.60	4.28	12.15
FLAN-T5 (3B)	22.48	9.52	18.45
INSTRUCTGPT (175B)	27.11	11.90	19.75
Finetuned LMs (on Collected Data)			
GALACTICA (6.7B)	20.93	6.16	15.01
T5 (3B)	26.66	11.45	20.73
T0 (3B)	<u>29.75</u>	<u>13.13</u>	<u>22.75</u>
FLAN-T5 (3B)	**32.22**	**14.62**	**24.53**
w/o Rationale Gen	27.73	11.31	19.32

Table 5.5: The results of full-stack QA systems on QASA.

5.5 Experiment

In this section, we evaluate our QASA approach on the proposed benchmark. In our experiment, we apply state-of-the-art LMs as two variants: Pretrained and Finetuned versions. In Sec 5.5.1 and 5.5.2, we automatically evaluate models on three subtasks: (1) *associative selection*, (2) *rationale-generation*, (3) *answer composition*, and their full-stack QA task. To complement automatic evaluations in our generation task, we conduct human evaluation in Sec 5.5.3 and an error analysis in Sec 5.5.4.

5.5.1 Experimental Setting

Evaluation of Subtasks and Full-stack QA For subtask evaluation, we provide oracle (or gold) contexts, in order to evaluate each subtask independently. In the *associative selection* task, we consider both positive paragraphs labeled by humans and negative paragraphs among top-10 retrieved results as candidate pool. For *rationale-generation*, we generate evidential rationale conditioned only on each of gold positive paragraphs. Similarly, for *answer composition*, we provide a list of gold evidential rationales as contexts. In contrast, for the full-stack QA, we consider the results of previous task as input to the next task sequentially, which could propagate the errors of the previous steps. Meanwhile, we conduct an ablation experiment, to directly generate final answers from selected paragraphs without *rationale-generation* (“w/o Rationale Gen” in Table 5.5).

Metric For *associative selection*, we measured the precision (P), recall (R), and F1 score. For *rationale-generation* and *answer composition* tasks, we used a standard text generation metric – ROUGE scores [144].

5.5.2 Main Results

Table 5.4 shows the automatic evaluation results of several QA systems on three subtasks and full-stack QA task.

Which pretrained LM is best? Among the pretrained LMs, INSTRUCTGPT (175B) outperformed others. Especially in the *rationale-generation* task, it shows the best performance among all models. Among T5-based LMs, the number of downstream tasks used during training had a significant impact on the performances in full-stack QA, showing T5 < T0 < FLAN-T5.

Which finetuned LM is best? When comparing finetuned T0, T5, and FLAN-T5, these models show little difference in performances on three subtasks. However, FLAN-T5 outperformed all other LMs on the full-stack QA, even the state-of-the-art model, INSTRUCTGPT (175B). Based on this observation, we suggest the finetuned FLAN-T5 could serve as a good test-bed for QASA.

The effect of training resources we curated For an ablation study, we trained individual Flan-T5 on each one of four datasets (QASPER, ASQA, ELI5, Augmented Data from GPT (or GPT AUG)). Through the comparison, we can observe negative transfer across datasets, *e.g.,* , FLAN-T5 trained on ASQA-ONLY shows the best results in the *answer composition* task, outperforming FLAN-T5 trained on combined data. Meanwhile, training of GPT AUG improved significantly the performances in the *rationale-generation* task, which is essential for our full-stack QA, as other resources do not contain rationales.

Does our task indeed need rationale-generation? For our full-stack QA, while we first generate rationales and then compose them into a final answer, we can directly generate an answer from selected paragraphs, skipping the step of rationale generation. However, as shown in Table 5.5, FLAN-T5 "w/o Rationale Gen" showed poor performance, compared to our three-step approach, which means the rationale generation step is crucial for the full-stack QA.

The failure of Galactica Although GALACTICA was pretrained on a large-scale collection of research papers, it performed worse on overall tasks compared to other models. The low performance of GALACTICA was consistently observed in Singhal et al. [216], compared to PubMedGPT of 2.7B. We empirically found that GALACTICA often answered either "yes" or "no", and terminated the generation, in which case the Rouge score is almost zero.

5.5.3 Human Evaluation

Although automated metrics can measure crucial aspects of our task, they are not guaranteed to closely approximate the judgment of humans, whose satisfaction is an overarching goal of a QA system. Therefore, we performed human evaluations based on the dimensions that should be satisfied in this task.

We conducted a pairwise evaluation scheme where evaluators compare two answers to the same question, inspired by Stelmakh et al. [220]. We provided two responses to each human evaluator, one from ours and the other from InstructGPT. The human evaluators could read the rationales and the generated responses side-by-side. Then, the evaluators were asked to choose the better answer in terms of four criteria: Groundedness, Completeness, Specificity, and Fluency, following prior work [220, 229]. For each data point, we assigned three evaluators to collect three trials of such pairwise judgments. The scoring system awards one point for a win and half a point for a tie in pairwise comparisons. The annotations were collected on 100 QA pairs by 9 experts.

The results of this human evaluation in Figure 5.3 show that the answers from our full-stack QA tend to be more complete and grounded than those from InstructGPT, which is consistent with the results from the automatic evaluation. In contrast, the InstructGPT's answers tend to be more fluent and specific, regardless of the reliability of its generated text. We also added some qualitative examples to show how the answers generated by our approach differ to those by InstructGPT in Table 5.6.

5.5.4 Error Analysis

To gain a deeper understanding of the model's errors, we sample 50 test examples with Rouge-L scores below 10 (*i.e.,* , bottom 25%). We exclude instances that are unanswerable based on the given paper. We then classify

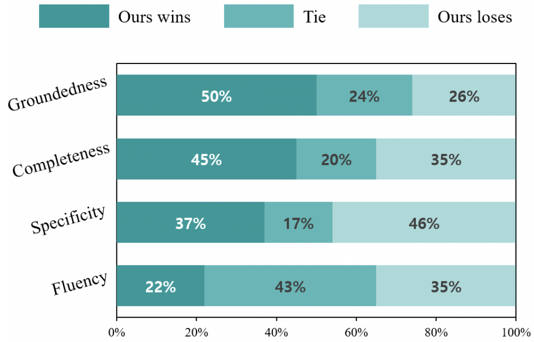

Figure 5.3: The results of human evaluation, comparing Ours to InstructGPT on four dimensions across 100 samples.

Table 5.6: The qualitative examples of answers, evidential paragraphs, and rationales from Ours and InstructGPT

Question	Answer-Ours	Answer-InstructGPT	Evidential Paragraphs	Rationales
Give two examples of public BERT-style english corpora.	*BookCorpus and English Wikipedia are two public BERT-style English corpora.*	*Two examples of public BERT-style English corpora are BookCorpus and OpenWebText. BookCorpus is a large-scale corpus of 11,038 books written in English, while OpenWebText is a collection of over 8 million webpages. Both corpora are used to train and evaluate natural language processing models, such as BERT.*	['We consider five English-language corpora of varying sizes and domains(...) We use the following text corpora:•BookCorpus (Zhu et al., 2015) plus English Wikipedia. This is the original data used to train BERT. (16GB).•CC-News, which we collected from the English portion of the CommonCrawl News dataset (Nagel et al.,, 2016). (...)]	['CC-News, BookCorpus and Wikipedia are BERT-style english corpora.']
How many tokens are changed to [MASK] in BERT training? Give a ratio.	*BERT training uses [MASK] to replace 80% of the tokens.*	*In BERT training, 15% of tokens are changed to [MASK], resulting in a ratio of 1:6.4 (15% of 512 tokens is approximately 80).*	[‘A random sample of the tokens in the input sequence is selected and replaced with the special token [MASK]. (...) BERT uniformly selects 15% of the input tokens for possible replacement. Of the selected tokens, 80% are replaced with [MASK], 10% are left unchanged, and 10% are replaced by a randomly selected vocabulary token.’]	[‘Of the selected tokens 15%, 80% are replaced with [MASK] during training.’]

errors into five categories, ranging from E1 to E5.

E1 refers to cases where the model incorrectly classified the question as unanswerable. E2 is the generation of irrelevant content. E3 is cases where the model provides implicit evidence but fails to generate an explicit answer. E4 refers to cases where the generation is not factually grounded on the source document. Lastly, E5 refers to cases with low completeness, where the generation only covers a partial answer (i.e., a sub-question). Additionally, a low Rouge score does not necessarily indicate a wrong generation. We identify two correct scenarios for this (C1 and C2). C1 refers to cases where the human labels are incorrect. C2 is cases where both the generation and human label are correct, but the lexical overlap between the two texts is low due to the diversity of expressions.

Table 5.7 shows error analysis results. 36% of InstructGPT's answers and 34% of ours belong to C1 and C2: cases with low ROUGE score, but correct. 48% of InstructGPT's answers are cases of refusal to answer (*e.g.,* , "I cannot find any specific information..."), although the context contains relevant evidences. We conjecture that InstructGPT has been trained to avoid answering in uncertain cases for safety. In contrast, our system did not generate such refusal responses, since there is no such example in our training data. 44% of our system's answers are irrelevant to a given question, although the text is grounded on evidence.

Type	Instruct GPT	Our Model
C1: incorrect human label	10%	10%
C2: low lexical overlaps	26%	24%
E1: predict unanswerable	48%	0%
E2: irrelevant generation	8%	44%
E3: failure of answering explicitly	0%	8%
E4: failure of grounding	6%	6%
E5: low completeness	2%	8%

Table 5.7: Error Analysis of InstructGPT and Ours.

5.6 Discussion

Limitation While we proposed a new benchmark for QA task on scientific articles, evaluation is becoming difficult, especially on recently emerging language models (InstructGPT as well as ChatGPT, Bard). Such language models aim not only for accurate responses, but also for longer responses through structured writing. Hence, evaluation metrics using string matching (such as ROUGE) may not represent the overall quality of generated results. The concurrent work showed that none of automatic metrics reliably matches human judgments of overall answer quality [252]. Future work for our QA task could look deeper into adopting multi-faceted evaluations.

Chapter 6. ARXIVDIGESTABLES: Synthesizing Scientific Literature into Tables using Language Models.

This chapter presents the second example of evaluation methods that assess alignment with human cognitive processes: ARXIVDIGESTABLES, a dataset that evaluates language models' capability of synthesizing scientific articles to tables. To examine this capability, we decompose this task into two sub-tasks—(1) schema generation and (2) value generation—and analyze where current models fall short in supporting human cognition. The content in this chapter adapts, updates, and extends material originally published in EMNLP 2024 [171]. All uses of "we," "our," and "us" refer to the coauthors of that publication.

6.1 Introduction

Conducting literature reviews by reading and synthesizing information across a large set of documents is vital for scientists to stay abreast of their fields yet is increasingly laborious as the number of scientific publications grows exponentially [94, 23]. At the core of this sensemaking process is identifying a *schema*, a set of important aspects that are useful for comparing and contrasting prior literature [204]. The results of this process are often presented in the form of *literature review tables*, whose rows are a set of papers and whose columns are a set of aspects that the papers share (Figure 6.1).

In this work, we conceptualize the task of literature review table generation by decomposing it into two sub-tasks: (1) *Schema-generation*: Determining a set of relevant shared aspects given a set of input papers, and (2) *Value-generation*: Determining the value given an aspect and a paper. For example, a table for a set of computer vision papers on video datasets (rows) might have a schema with aspects like "`task`" or "`size`" (columns); cell values under the "`task`" column may say "`VQA`" or "`classification`" (values).

Prior work has largely investigated each of the two sub-tasks independently. In particular, the large body of literature on document-grounded question-answering [121, 51, 129], information extraction [156], and query [262, 253] or aspect-based summarization [258, 1] advances methods that are also suitable for generating values conditioned on an aspect. In our example above, values for aspect "`size`" can be answers to questions like "How many videos are in this dataset?".

In contrast, schema generation from a set of documents remains relatively under-explored, even though it is a crucial and effortful part of the manual literature review process. Prior work like Zhang and Balog [260] infers new schemas from pre-existing ones, while recent work like Wang et al. [239] assumes users can clearly articulate a schema in a short natural language query to infer aspects directly. This paper studies the use of language models for literature review table generation with a focus on unifying these two sub-tasks. This presents us with two research challenges:

First, we note a lack of large-scale, high-quality datasets of literature review tables to serve as a benchmark for this task. Second, similar to challenges faced in summarization and other grounded generation tasks, semantically similar content can be expressed with different surface forms, which makes automatic evaluation difficult even with a high-quality dataset. An example of these surface form differences is in Figure 6.2. To address these challenges:

- In §6.2, we curate and release ARXIVDIGESTABLES,[1] a dataset of 2,228 high-quality literature review tables scraped and filtered from 16 years of ArXiv papers uploaded between April 2007 and November 2023.

[1] DIGESTables stands for **D**ocument **I**nformation **G**athering and **E**xtraction for **S**cientific **T**ables

Figure 6.1: Schematic of our literature review table generation task: (1) synthesize multiple input papers into a table with both (2) a schema (columns) and (3) values. Each row corresponds to an input paper.

These tables compare and contrast a total of 7,542 unique papers using a total of 7,634 columns and 43,905 values. This is the result of extensive filtering on an initial set of around 2.5 million extracted tables to ensure high quality, based on a strict set of desiderata. Finally, we link every table to rich paper content: (1) every input paper (row) has corresponding full text document, and (2) every table has its caption and in-line textual references extracted from the table's source paper for contextual information.

- In §6.5, we present DECONTEXTEVAL, an automatic evaluation framework for comparing model-generated and human-authored tables. Our approach overcomes the difficulty in matching semantically-similar but lexically-different column names by using a language model to expand column names into descriptions grounded in documents. Combining with a small textual similarity model results in a matcher that is nearly twice more precise than prompting Llama 3 (70B), which often hallucinates matches.

We formalize the literature review table generation task (§6.3) and introduce our framework for literature review table generation and detail our implementations using open and closed models (§6.4).

Finally in §6.6, we evaluate LMs on this generation task, addressing two key questions: (1) what contextual information is needed to steer language models to reconstruct human-authored schemas? and (2) are generated aspects that *don't* match gold still useful? For (1), we find that language models have higher recall by conditioning on more context that specifies the purpose of the table (e.g., captions, in-line references, other example tables). For (2), we find that novel aspects not in the reference tables can still be of comparable usefulness, specificity, and insightfulness.

6.2 Creating ARXIVDIGESTABLES

Desiderata To enable research in synthesizing literature review tables, we first collect and curate a set of reference tables to ground our task and enable evaluation. To ensure this data is realistic, high-quality, and focused on supporting literature review, we decide on the following desiderata for including tables in our ARXIVDIGESTABLES dataset:

1. Tables should be ecologically valid—reflecting real syntheses authored by researchers rather than artificial annotation;
2. Tables should be focused on summarizing multiple aspects of a set of papers as opposed to tables for reporting empirical results;

Original Reference Table

	Dataset	Size	Task	Annotations
Paper 1	KoNViD-1k	1200	VQA	114
Paper 2	LIVE-VQC	585	VQA	240
Paper 3	KoNViD-150k	153,841	VQA	5
Paper 4	Sports-1M	1,133,158	Classification	- (auto)

Model Generated Table

	Dataset size	Annotation method	Intended Application	Evaluation Metric
Paper 1	1,200 video sequences	Subjectively annotated	Objective VQA method development	Subjective Mean Opinion Score
Paper 2	585 videos	Subjective video quality scores via crowdsourcing	NR video quality prediction advancement	Subjective video quality scores
Paper 3	153,841 videos	Coarsely annotated set with five quality ratings each	Deep-learning VQA model training	Spearman rank-order correlation coefficient
Paper 4	1 million YouTube videos	N/A	Large-scale video classification and action recognition	Performance improvements over baselines

Figure 6.2: Side-by-side comparison of a reference literature review table from an ArXiv paper [147] and a model-generated table given the same input papers. The generated table has reconstructed two gold aspects: the pink and blue aspects are the same, despite surface form differences (e.g., "Task" vs "Intended Application"). The generated table has also proposed two novel aspects that are still relevant and useful, like "evaluation metric" (green) or "Annotation method" (yellow) not to be confused with reference table's "Annotations".

3. Tables should follow a common structure where each row represents a single document and each column represents a specific aspect.

Based on these goals, we used the procedure below to construct ARXIVDIGESTABLES:

Data Source To ensure our task and benchmark are grounded in realistic cases, we collected a dataset real-world literature review tables from open access ArXiv papers from April 2007 until November 2023. We subsequently filter these tables down to a high-quality set of 2,228 tables that meet our desiderata, as seen in Figure 6.3.

Extracting Tables The first step in our data collection pipeline is to extract the tables from papers published on the ArXiv preprint server. To start, we consider approximately 800,000 papers that have LaTeX source available. We then use unarXive [205] to convert the ArXiv source into XML. From these XML documents, we extract ~2.5 million tables.

Filtering Tables As a first filtering pass, we remove tables that are likely to be misparsed or unusable, filtering those with fewer than 400 or more than 15,000 characters. We also remove tables that have no table cell tags within them. Toward Desiderata 3, we filter out tables that have fewer than two citations, two rows, or two columns. We also remove any tables that have citations in more than one column, as these are often tables where papers are values rather than rows. This leaves approximately 211,000 tables.

Matching Rows to Papers We use heuristics to convert XML-formatted tables into JSON objects that allow us to directly index the tables by paper and aspect. At this stage, the citation information is usually contained

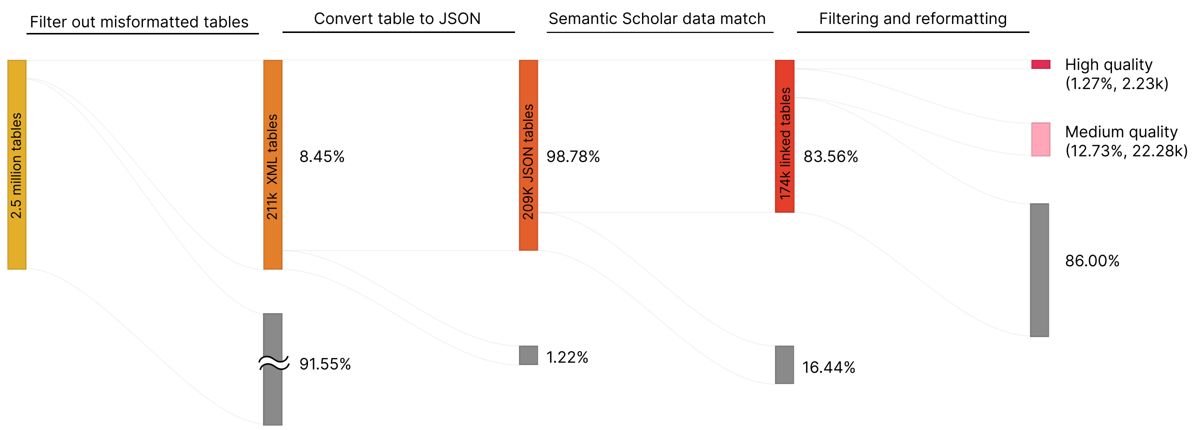

Figure 6.3: Pipeline for curating ARXIVDIGESTABLES involves extensive data cleaning and filtering. The full pipeline filters from 2.5 million starting tables published in 800,000 papers to 2,228 tables published in 1,723 papers. Data pipeline described in §6.2.

within a cell in a table. For instance, an example cell with the header "`Model`" might have the value "`BERT (Devlin et al., 2019)`". We extract the citations from these cells and place them in their own column called "`References`". Rows without citations are assumed to refer to the source paper containing the table. After this step in the process we remove any tables where the algorithm failed and any tables that now have fewer than two rows, leaving 47,876 tables.

Obtaining Table Citation Metadata unarXive [205] helpfully links each citation in the table to a bibliography item. We use endpoints from the Semantic Scholar API [111] to obtain titles and abstracts. This occasionally fails for various reasons (e.g., the bibliography text is missing information, the paper is missing from or could not be found in the Semantic Scholar database). We filter out any tables that have fewer than two matched citations, leaving us with 44,617 tables.

Grounding to Paper Texts To meet Desiderata 2, we want to ensure that the information in the table actually comes from the cited paper. For instance, a common type of table reports experimental results whose values require actual experimentation and cannot be derived from the input papers' text alone. To filter such columns, we remove any that have math symbols or floating point numbers. Additionally, to make sure the generation task is tractable, we remove any rows whose papers do not have publicly-available full texts.

Final Filter and Manual Verification The last step applies a set of stringent filters and manually identifies and corrects any parsing errors. Finally, we produce a set of 2,228 high-quality tables.

Dataset Statistics We present summary statistics in Table 6.1 of our high-quality set of ARXIVDIGESTABLES.[2] We are also interested in the types of aspects represented in the tables, the topics of the columns, and the fields the tables come from. To categorize the table aspects, we use simple heuristics (Table 6.2). We find ~40% of the columns are categorical or boolean, which are more suitable for supporting inter-paper comparisons, while the other ~60% are more descriptive. To obtain column topics, we manually annotate columns in ~50 tables—~38% are about datasets, ~20% are about methods, and the rest are on other topics such as applications or tasks. Finally,

[2] To enable future work to improve on this pipeline, we also release a set of 22, 283 medium-quality tables (see Figure 6.3) with less strict filtering alongside which filters we ran to produce it along with quality metadata.

we use the ArXiv API to obtain which archive a table's paper was submitted to. We find a majority (1,985) of the tables come from computer science publications, with others coming from Physics, Quantitative Biology, Statistics, Math, and other fields.

	Min	Max	Median	Mean	Total
Papers	1	35	3.0	4.944	11016
Aspects	2	13	3.0	3.426	7634

Table 6.1: Number of papers (rows) and aspects (columns) in ARXIVDIGESTABLES. Of the 11,0016 total rows there are 7,542 unique papers.

Aspect Type	% of Cols	Example Value
Category	35.5%	"Open" vs "Proprietary"
Entity	27.3%	"CNN/Daily Mail", "Reddit"
Numeric	21.7%	"10, 000"
Text	9.7%	"...collected via various ..."
Boolean	5.8%	"✓" vs "✗"

Table 6.2: Types of aspects in ARXIVDIGESTABLES's columns.

6.3 Literature Review Table Generation

Equipped with our dataset, we formalize the task of generating literature review tables.

Task Definition We define our *table generation* task as follows: Given an input set of M documents $d_1, \ldots, d_M$, generate a table with M rows and any number of columns $N \geq 2$. Each row $r_1, \ldots, r_M$ corresponds to a unique input document. Each column $c_1, \ldots, c_N$ represents a unique aspect. Taken together, the columns constitute a schema. The table then has $N \times M$ values, with one value in each cell.[3] The cell values should be derived from the input documents.

Generation We consider two main approaches to generate a table given a set of input documents. (1) The schema and values could be *jointly* generated, e.g. in a single call to a language model. This approach is fast, but initial experiments found it more prone to hallucinations and generic column names (e.g., "`Title`" or "`Year`"). (2) The generation process can be *decomposed* into separate schema and value generation steps. This approach is slower but allows us to overcome context window limits and leverage prior work in aspect-based question answering to perform value generation.

Evaluation We evaluate our approaches by determining whether the generated schemas are *useful* and values are *correct*. We consider a generated schema to be useful if its aspects either match those in the corresponding human-authored table in ARXIVDIGESTABLES or if human evaluators rate them to be useful.[4] These two conditions allow us to measure how well systems *reconstruct* reference table aspects (§6.5.1) and evaluate their ability to generate *novel* aspects (§6.6.1). Second, we evaluate correctness of values as we would for any information extraction or QA task: for a pair of aligned columns (and rows), we judge whether the predicted cell value is semantically equivalent to the gold cell value (see §6.5.2).

[3] We leave the case where a cell can contain multiple values to future work.

[4] There are many alternative ways to evaluate usefulness. For example, adding constraints on users' reading time could penalize very detailed tables, while ideation-focused use cases could penalize more generic aspects.

6.4 Experiments

We prompt language models to perform either joint or decomposed generation.

6.4.1 Base Models

We use two language models, one open-weight, Mixtral 8x22 [164], and one closed weight, GPT-3.5-Turbo [177]. To avoid gaming our recall metric, we instruct all models to generate schemas with the same number of aspects as the corresponding reference tables.

6.4.2 Joint Table Generation

We represent input papers using their titles and abstracts, which usually have enough information to form useful schemas and are easier to fit in the context window of models. We use a zero-shot table generation prompt. We treat this condition as our baseline.

6.4.3 Decomposed Table Generation

Step 1: Schema generation Like in joint generation, we represent input papers using their titles and abstracts. We explore a range of prompts, each including a different piece of additional context (detailed in §6.4.3).

Step 2: Value generation Similar to extractive QA, for each aspect-paper pair, we prompt the model to generate a cell value based on the aspect name and the *full text* of the paper. After generating values for each paper given an aspect, we instruct a model to rewrite the values to be shorter and more consistent in style for display in table format. For this step, we use GPT 3.5-Turbo for speed and accuracy [177].

Additional Context To further investigate what contextual information is needed to steer language models to reconstruct human-authored tables, we test the following additional contexts, which could be added to either schema and/or value generation (Figure 6.4). (1) a generated caption where GPT-3.5-Turbo generates a short description that is consistent with all input papers; (2) the gold caption from the reference table; (3) the gold caption and in-text references, which include referencing sentences from the table's source paper; and (4) few-shot in-context examples, consisting of five reference table examples from ARXIVDIGESTABLES retrieved based on cosine similarity between caption embeddings [196].

6.5 Developing an Automatic Metric

Below we describe the design of our automatic evaluation procedure with two components: evaluating the schema and values for a generated table.

6.5.1 Schema Evaluation

Challenges The key challenge in assessing how well a generated table reconstructs a reference table lies in *determining schema alignments*—identifying which columns convey the same information despite different phrasing. Two issues make schema alignment difficult. First, reference tables tend to present information concisely, making column headers and values hard to interpret without additional context (e.g., a column might be named "`VQA`"

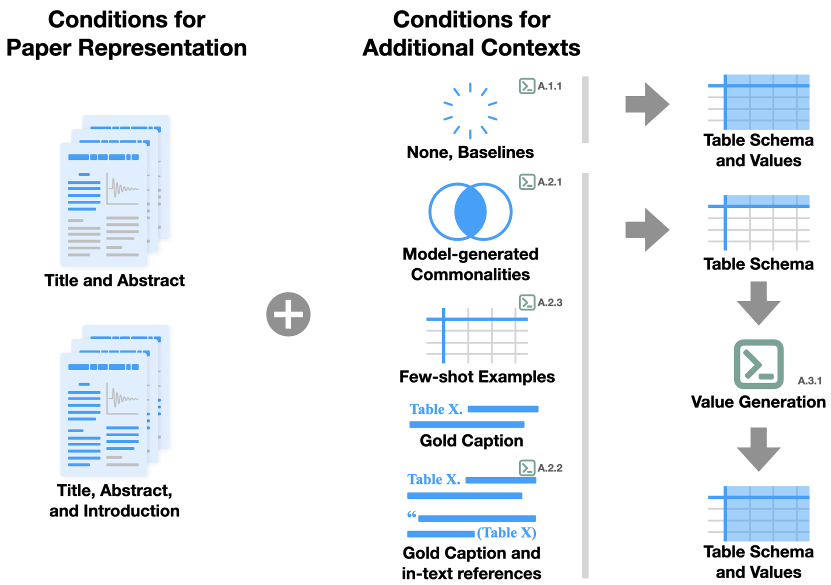

Figure 6.4: Diagram of prompting methods under experiment conditions.

instead of "`video quality assessment`"). Second, information in generated and reference tables might have low lexical overlap despite semantic similarity, a problem also observed in summarization evaluation [145].

Problem Definition To formalize the schema alignment problem, recall that a table schema is a set of N aspects. Given a model-generated table schema, $S^m = \{a_1^m, \ldots, a_N^m\}$, a reference table schema $S^r = \{a_1^r, \ldots, a_N^r\}$, and a threshold $0 \leq t \leq 1$, our goal is to construct a scoring function f to score each pair of aspects, (a_i^m, a_j^r), such that $f(a_i^m, a_j^r) > t$ if and only if human raters would agree that a_i^m and a_j^r convey the same information.

Alignment Framework We propose to define f as the composition of two functions: a featurizer (ϕ), and a scorer (g). The goal of the featurizer is to improve aspect interpretability by incorporating additional context, while the goal of the scorer is to account for meaning-preserving lexical diversity, leading to better schema alignments.

Configurations of f We study three featurizers ϕ: (1) "`name`" only takes the column name as-is, (2) "`values`" concatenates all values under a column to the name, and (3) "`decontext`" prompts a language model[5] to generate a stand-alone description [39, 172], given the column name and its values.

We also study four scoring functions g:

- **Exact Match**, which assigns a score of 1 if $\phi(a_i^m) = \phi(a_j^r)$ and 0 otherwise.
- **Jaccard**, which computes Jaccard similarity of the featurized aspects, with stopwords removed.
- **Sentence Transformers**, which encodes featurized aspects using `all-MiniLM-L6-v2` and computes cosine similarity between them [196].
- **Llama 3**, which prompts Llama 3 (70B) Chat with generated and reference tables, with the column headers replaced by featurized versions, instructions to output aligned columns, and ten in-context examples. All pairs of columns returned by the LLM are assigned a score of 1, and 0 otherwise.

[5] `Mixtral-8x7B-Instruct-v0.1` [164].

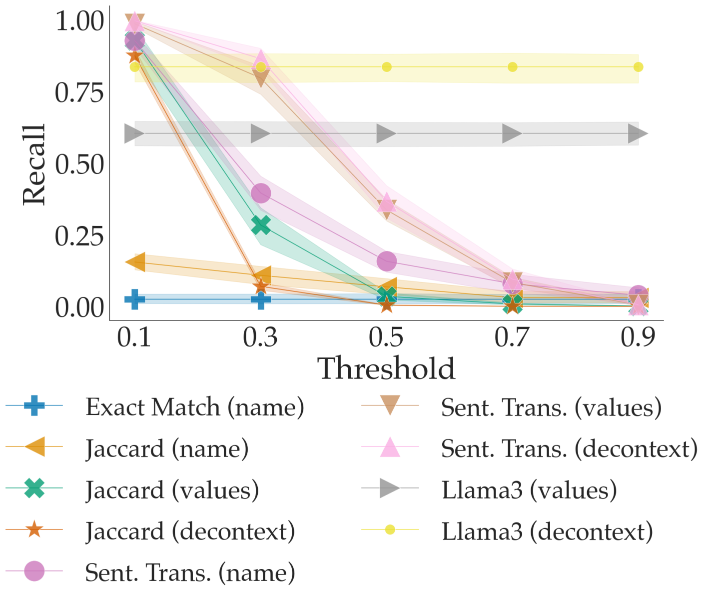

Figure 6.5: Recall averaged over different contexts and systems. The band represents 95% confidence interval. Llama3 scorers have high recall, but low precision. Sentence Transformers (decontext) has the best trade-off.

Calibrating Schema Alignment We first run various combinations of (ϕ, g, t) and compute schema recall (i.e., proportion of reference table aspects matched to generated table aspects) on 25% of the tables in ARXIVDIGESTABLES. In Figure 6.5, we observe a wide range of recall trade-offs: (1) Exact match has very low recall, as expected, serving as our conservative bound. (2) Llama 3 aligners tend to predict many more matches than other configurations despite that half of the in-context examples are tables with no matches. Llama 3 aligners serve as our upper bound. We perform human evaluation on ~50 tables and find that Llama 3 aligners have between 37–55% precision on their predicted matches.[6] (3) Focusing our attention on the configurations that yield recall between these two bounds, we evaluate a range of configurations on the same tables and arrive at DECONTEXTEVAL, our best configuration with ϕ using `decontext` features, g using sentence transformers, and $t = 0.7$; we find DECONTEXTEVAL performs at 70–85% precision with acceptable yield.

6.5.2 Value Evaluation

Automated value evaluation suffers from the same issues that complicate schema evaluation, but one issue specific to value evaluation is reliance on *accurate schema alignments*. If aspects are incorrectly matched by a schema alignment metric, performance on value evaluation might rise/drop undeservedly. Therefore, we propose evaluating value generation in isolation, instead of an end-to-end table evaluation setting.

Specifically, we use the reference table's schemas as input to our value generation module. This ensures that every value in the reference table has a corresponding generated value (barring generation failures), bypassing the need for schema alignment. Following §6.4.3, we consider three settings using different types of contexts: (1) "`Column Names`" only, (2) "`Caption Context`" which adds the table caption, and (3) "`All Context`" which further adds in-text references. We then use the same suite of scorers from §6.5.1 (except Llama 3, which we observed was low-precision) to compute overlap between pairs of generated and reference table's values.

[6]Predicted matches are rated either as incorrect, partially, or completely correct. The lower bound only counts complete matches and the upper bound includes partial matches.

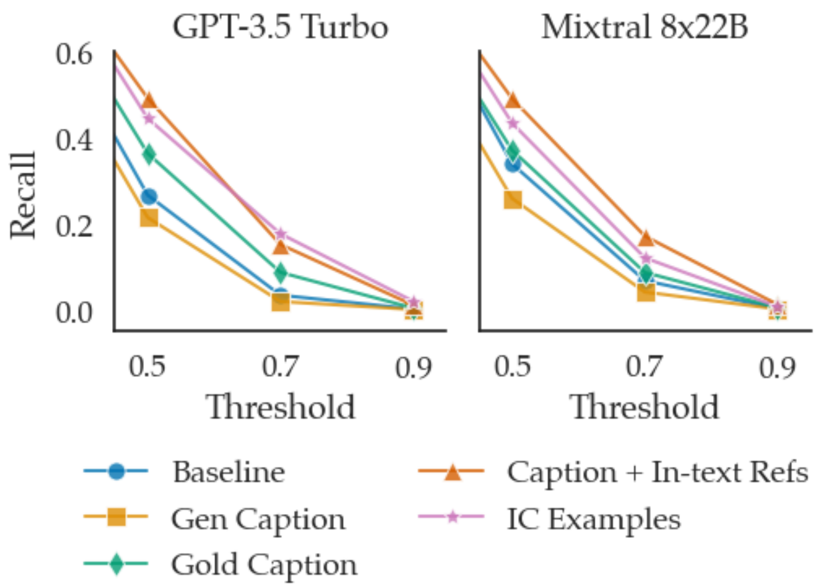

Figure 6.6: Schema recall for GPT-3.5-Turbo and Mixtral 8x22, using various types of additional contexts. All scores are computed using our best metric: sentence transformer-based scorer with decontext featurizer. More context improves recall, but does not lead to completely reproducing reference table schemas.

6.6 Results

6.6.1 Schema Evaluation Results

Automated Evaluation Figure 6.6 shows the ability of GPT-3.5-Turbo and Mixtral 8x22 to reconstruct schemas (as measured via DECONTEXTEVAL) using various types of additional contexts described in §6.4.3. Turning back to the question: *How does the amount of context provided affect table reconstruction?* (1) We see that **low context** prompts (e.g., a baseline with no additional context, caption-only) perform the worst while **high context** prompts (e.g., in-text references, in-context examples) perform best. This trend is fairly stable across systems. (2) Interestingly, though adding context improves reconstruction, it does not make the task trivial — even the best performing systems are far from perfect.

One potential concern for this analysis is that the models we use may have seen the older tables during training, which could inflate performance. To address this, we compute recall separately on subsets of newer and older tables (those from before or after January 2023 constituting 30% and 70% of our data respectively) for the high context prompts. We find that there is minimal difference between these two sets (the newer tables have recalls on average 1–3 percentage points lower).

Human Evaluation Our automated evaluation measures how well LMs can recover the reference tables' aspects, but leaves an additional question: *Are LM-generated novel aspects which do not match with gold aspects also useful?* To investigate this, we collect human assessments of generated aspects. Annotators are provided a generated table and the titles of all input papers. They are then prompted to provide a 5-point Likert scale rating for each of the following aspects: (1) general *usefulness* for understanding the input papers, (2) *specificity* to input papers (i.e., would this aspect be applicable to any other set of papers), and (3) *insightfulness* of the generated aspect (i.e., capturing novelty). We also instruct annotators to only judge based on the quality of the *aspects* only, ignoring the values which are evaluated separately. After collecting these ratings, we separated the rated aspects into two groups—ones that matched a gold aspect (**M**), and ones that did not (**NM**). The annotators were blind to the

	Caption+In-text Ref		Baseline	
	M	**NM**	**M**	**NM**
Useful	3.70 (1.74)	4.07 (1.06)	3.92 (0.69)	3.73 (1.17)
Specific	2.88 (1.26)	3.06 (1.34)	2.85 (1.31)	2.75 (1.35)
Insightful	1.86 (1.04)	1.93 (1.21)	2.34 (1.25)	2.27 (1.19)
# Samples	102	208	64	283

Table 6.3: Mean (SD) ratings from human assessments of generated aspects that match the gold schema (M) with those that do not (NM).

conditions when rating the aspects, and inter-annotator agreement was 0.56 (Krippendorff's α).

Comparing ratings on matched and unmatched aspects, we did not find aspects that matched to be rated significantly higher than ones that did not (Table 6.3; Mann-Whitney U tests). This suggests that novel generated aspects are of comparable quality (*usefulness*, *specificity*, *insightfulness*) to gold aspects or even have a higher quality (*usefulness* of aspects from Caption+In-text References). Moreover, aspects from Caption+In-text Reference are shown to be more *useful* and *specific* than the Baseline's, but were less *insightful*. This suggests an interesting tradeoff between our reconstruction objective, and possibly a different objective like creativity.

Error Analysis Finally, we report some qualitative observations of errors in the generated schemas we used for human evaluation. These point to future areas of improvement. Comparing outputs from the baseline to the Caption+In-text References condition, we find that the latter tends to output more specific aspects. For example, for one table, the Mixtral baseline produces aspects "Model Architecture" and "Application", while the Caption+In-text References Mixtral system generates the more specific aspects "Maximum resolution" and "Training batch size". We also note a few differences between schemas generated in the Caption+In-text references setting the reference tables' schemas, as well as categories of aspects that can pose difficulty for generation in Table 6.5.

6.6.2 Value Evaluation Results

Automated Evaluation. Figure 6.7 shows the performance of GPT-3.5-Turbo on value generation, using various types of additional contexts (described in §6.5.2.) We see that scorers continue to follow the same trend observed during schema alignment, with the sentence transformer scorer being fairly permissive while an exact match is overly strict. Interestingly, unlike schema reconstruction, we observe that incorporating additional context does not seem to improve value generation accuracy; we dig deeper into this during human evaluation. Finally, like schema alignment, models are far from perfect in value generation.

Human Evaluation. We conduct additional human evaluation to investigate whether adding context indeed has no impact on value accuracy, or our automated metrics are not sensitive enough to capture differences. We randomly sample 30 tables and compare gold vs generated values for these tables under all three settings. For each gold-generated value pair, we have two annotators label whether it is a complete match, partial match or unmatched. Partial matches include cases where values are lists of items and the generated value misses or adds some (e.g., "DPO" vs "DPO, PPO"), or cases where the gold and generated values have a hypernymy relationship (e.g., "graph neural networks" vs "GATs"). Inter-annotator agreement is 0.55 (Cohen's κ). Table 6.4 presents results from this assessment, showing that adding additional context leads to a significant improvement in partial matches. However, many matches have no lexical overlap (e.g., "X" vs "No") or require some inference (e.g., "Yes" under a column called "sensors deployed" should match a value like "sensors used to monitor air quality"). This indicates that there

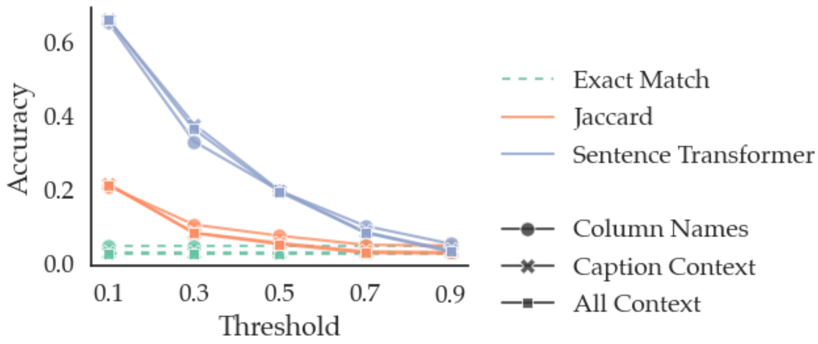

Figure 6.7: Value generation accuracy for GPT-3.5-Turbo using various types of additional contexts, as computed by different scorers.

Setting	Complete	Partial	None
Col. Names	21.13% (75)	22.54% (80)	56.34% (200)
+ Captions	18.84% (65)	31.30% (108)	49.86% (172)
+ IT-Refs	22.65% (77)	31.77% (108)	45.59% (155)

Table 6.4: Proportion of matched gold-generated value pairs for various context settings, according to human assessment.

is scope for further research in developing more sensitive featurizers and scorers for value evaluation.

6.7 Discussion

6.7.1 Limitations

We only study scientific papers from ArXiv. While in theory, scientists in many fields produce literature review tables, we restrict our reference tables to ones that we can scrape from ArXiv. This means many of the papers in our dataset come from fields that are most represented on ArXiv (e.g. computer science) and fewer come from medicine, humanities, or social science publications. Additionally, all of the tables in our high quality set are in English, even though literature review tables may also be used in other languages.

Reconstructing tables is difficult. While DECONTEXTEVAL is effective at matching generated and reference table columns, and we test providing different additional context to steer the table generation models, many generated table columns do not match with the reference columns. Though we presented a human evaluation protocol that showed utility for generated columns that do not match the reference columns, such evaluation is costly. Future work should investigate automatic metrics that correlate with human utility evaluations as well.

6.7.2 Ethical Considerations and Broader Impact

Generated literature review tables might misrepresent authors' work. Generating literature review tables requires taking aspects of papers out of their original context to show them to users. Similar to summarization, this process has the potential to misrepresent the original work either due to the table cell values not having enough context, or less accurate models introducing hallucinations. Additional checks would have to be implemented if such tables were to be deployed in user-facing situations.

Challenge Type	Description
Different Granularity	The generated schema might be a high-level category (e.g. “data types”), while the reference schema includes more specific aspects (e.g. “image”, “text”, “audio”, etc.)
Different topics	The generated schema might have a different variety of topics than the reference schema (e.g. {“model architecture”, “dataset used”, “performance metric”} versus just dataset properties {“color”, “context”})
Complex Aspects	Aspects combine information from multiple cells, which can mislead the value generator. E.g. “dataset size” leads to some values pertaining to training data and others to test data.
Overly Specific	A predicted aspect might only apply to one paper

Table 6.5: Qualitative observations of challenges with generated tables

Literature review tables may discourage reading original sources. The resource we present is meant to encourage the development of methods to construct literature review tables. If the field iterates on this task and develops systems that perform very well, the tables may have all of the information that a given reader wants to see. This could discourage readers from finding the original source of the claims. That said, the rows in the tables in our benchmark do include citations, so readers can trace values back to their sources. However, readers are not guaranteed to follow these citations, so generated tables could encourage poor scholarly practices.

6.8 Conclusion

Language models have the potential to help scientists organize papers during literature review by synthesizing tables with schemas that aid comparison. In this work, we curate ARXIVDIGESTABLES, a dataset of such tables and additional contexts that can be used to evaluate systems' abilities to produce such tables. We present DECONTEXTEVAL, an automatic evaluation framework for comparing model-generated and human-authored reference tables. We then use this evaluation framework to investigate two research questions: what context is needed to reconstruct human-authored tables, and whether generated aspects that don't align with references are also useful, specific and insightful. We release our artifacts to help spur development of literature review table generation systems, and seed potential for their role in evaluating systems' scientific synthesis abilities.

Chapter 7. Discussion

This chapter compares AI alignment with cognitive processes against general alignment and presents design guidelines to build adaptive AI systems for future researchers and practitioners. In addition, I on how adaptation and evaluation, while addressed separately in this thesis, may interact in a reinforcing cycle that supports continuous improvement in cognitively aligned AI systems. I will also discuss some of the limitations of the approaches to enabling adaptation through cognitive AI alignment in this thesis.

7.1 Alignment with Cognitive Process vs. Existing Alignment

AI alignment focuses on ensuring that AI systems act in accordance with human intentions and values. As AI capabilities advance, alignment approaches have evolved from basic alignment (helpful, honest, and harmless) to more sophisticated methods addressing complex challenges in decision-making and ethical considerations. One notable development in alignment approaches is RLHF, which refines model behavior based on human feedback [178].

- **Assumptions about the user.** Traditional alignment often treats the user as a fully specified oracle (e.g., providing optimal prompts) [], whereas cognitive alignment assumes users have incomplete knowledge and may not know what to ask, how to ask it, or how to interpret the answer.
- **Nature of the task.** Standard alignment focuses on task correctness or reward optimization, while cognitive alignment focuses on supporting comprehension and learning process.
- **Evaluation criteria.** Existing benchmarks typically assess surface-level correctness or preference agreement. In contrast, this thesis introduces user-centered and cognitively motivated evaluation frameworks (e.g., multi-step QA reasoning, comparative understanding) that reflect alignment with the user's thought process.
- **Interaction model.** Traditional alignment assumes one-shot or limited interaction, whereas cognitive alignment assumes rich, iterative interaction—where AI systems must infer and adapt over time to evolving user knowledge states.

7.2 Design Guidelines for Adaptive AI Systems

To effectively support each different individual's knowledge understanding, AI systems need to be designed with a deep awareness of how humans process, structure, and build upon knowledge. This section presents three key design guidelines distilled from the systems developed in this thesis: (1) structuring knowledge for interaction, (2) eliciting user knowledge states, and (3) generating adaptive interventions. Together, these guidelines offer a foundation for building AI systems that align with users' cognitive processes across diverse contexts.

7.2.1 Structuring Knowledge for Interaction

The first step in building adaptive AI systems is to structure the underlying information in ways that align with how people comprehend and relate knowledge. Drawing on human learning theories and cognitive science, the work in this thesis design representations (e.g., tables, dialogue trees) that structure knowledge into meaningful units and relationships. These structured representations enable systems to (1) elicit users' relevant knowledge states to their target understanding and (2) generate cognitively meaningful interventions that satisfy users' needs.

The granularity of knowledge units and relationships needs to be adjusted based on specific cognitive processes required by distinct users and contexts. For example, when supporting researchers assessing the relevance of research papers, high-level comparisons across key common dimensions (e.g., problem, method, and findings) are more effective than unstructured summaries. In this case, tables that explicitly reveal similar and different values on these common aspects help surface cognitively salient similarities and differences. In contrast, when supporting children learning scientific concepts, gradual exposure to small, connected information units—structured in a dialogue tree or explanatory chain—better matches their cognitive load capacity.

7.2.2 Eliciting User Knowledge States

To deliver adaptive support, systems need to infer what the user knows. This thesis explores multiple strategies to elicit user's knowledge states, ranging from explicit questioning to implicit inference from interaction patterns. To effectively elicit users' knowledge states, knowledge representations are combined with HCI methodologies to design human-AI interactions that suit users' tasks.

Explicit Interaction In settings where rich, back-and-forth interaction is useful, such as with children—explicitly asking questions is an effective strategy. For example, in the DAPIE system [128], questions are used to diagnose what the child understood, what needs clarification, and what the child wants to know next. These interactions help the system dynamically adapt the content and keep the user cognitively engaged.

Implicit Interaction In contexts that require navigating large volumes of information or performing cognitively demanding tasks, frequent questioning can be burdensome and disruptive. In such situations, it may not be feasible to explicitly elicit all aspects of the user's knowledge state. Instead, AI systems can make use of information already available about the user, such as their reading history, prior interactions, or saved materials, to implicitly infer relevant aspects of their knowledge. For example, PaperWeaver [126] analyzes users' previously read or saved papers to estimate their familiarity with specific research topics, and generates personalized explanations accordingly, without requiring additional input from the user. This approach allows systems to adapt to users' cognitive context while minimizing interaction overhead.

Combined Interaction In many real-world scenarios, especially those involving complex or evolving user goals, effective support requires a combination of both explicit and implicit interaction. Explicit input, such as specifying goals or asking clarification questions, can enhance user control and agency, allowing users to steer the interaction based on their needs. At the same time, leveraging implicit signals, such as interaction history, document selection, or behavioral cues, can reduce the cognitive and interactional burden by enabling the system to infer relevant aspects of the user's knowledge state without requiring constant input. Together, these complementary strategies allow the system to maintain adaptivity while balancing user effort and autonomy. For instance, when helping knowledge workers navigate complex documents, the DocVoyager [130] system combines both types of interaction to suggest personalized reading paths. The system elicits user goals and interests both explicitly (e.g., asking their reading goal) and implicitly (e.g., through previously consumed information), enabling the system to continuously refine its understanding of the user's knowledge state. This combined interaction helps the system infer the user's knowledge state more accurately and adapt its support accordingly, without causing unnecessary interruption.

7.2.3 Generating Adaptive Interventions

To provide meaningful support in cognitively demanding tasks, this thesis introduces computational methods for generating adaptive interventions aligned with each user's cognitive process. Rather than relying solely on language models to produce generic or simplified explanations, the proposed method incorporates structured knowledge representations tailored to each task and draws upon principles from cognitive science and human learning theories to generate adaptive interventions.

As discussed in Section 7.2.1, each understanding task requires different meaningful knowledge units and relational structures. Using natural language processing techniques such as discourse parsing and information extraction, the system transforms source texts into structured knowledge representations that capture these units and their relationships. This representation enables the system to identify a user's current knowledge state and target knowledge. The system then transforms the user's knowledge state in natural language, allowing the language model to take it as input and apply effective explanatory strategies that bridge prior knowledge with new information.

To assess the effectiveness of this approach, the thesis introduces novel tasks and benchmarks. These evaluations focus on the quality of cognitively grounded interventions, such as analogies and comparative explanations between research papers, demonstrating that the proposed methods produce outputs that are both pedagogically meaningful to target users and aligned with human cognitive processes.

7.3 A Cycle of Adaptation and Evaluation

This thesis advances the view that supporting human understanding requires not only building adaptive AI systems, but also continuously evaluating how well these systems align with the way people think, reason, and learn. These two threads, system design and evaluation, are not independent efforts, but part of a reinforcing cycle. Each informs and shapes the other in the pursuit of more cognitively aligned AI.

On the one hand, improvements in system design offer opportunities to examine what cognitive capabilities are missing or underdeveloped in current models. For instance, as systems like DAPIE [128] begin to scaffold understanding through interactive dialogues, they expose limitations in model abilities such as generating pedagogically effective follow-up questions or detecting when to scaffold children's prior knowledge. Traditional evaluation metrics, such as answer correctness or preference agreement, fall short in assessing these capabilities. As a result, DAPIE [128] adopts alternative evaluation criteria and contributes to emerging benchmarks that assess capabilities like asking clarification questions and generating pedagogically meaningful questions.

On the other hand, cognitively aligned evaluation methods can drive the development of AI systems that are more trustworthy, interpretable, and robust. The process-level evaluation frameworks proposed in this thesis—such as QASA, which decomposes tasks into reasoning steps and analyzes capability to generate intermediate outputs—enable system designers to pinpoint where models fall short and identify opportunities for improvement. These insights are not only diagnostic, but suggest new design opportunities: how to revise interfaces, adapt model prompts, or reconfigure interaction flows to better support human knowledge understanding.

Chapter 8. Conclusion

This dissertation introduced how to develop AI systems to interactively align with human cognitive processes.

8.1 Summary of Contributions

This thesis makes contributions in human-AI interaction, adaptive knowledge understanding, and human-centered evaluation.

- Human-AI Interaction: Interaction techniques powered by knowledge representation that models human cognitive processes grounded in human learning theories
- Adaptive knowledge understanding: Interactive systems that provide adaptive interventions through novel LLM-powered pipelines that generates outputs aligned with each individual's cognitive processes
- Human-centered Evaluation: Novel evaluation frameworks and benchmarks that assess how well AI outputs align with human cognitive process by decomposing task into fine-grained cognitive sub-steps

8.2 Future Directions

8.2.1 Ambient Cognitive AI Alignment

As AI systems evolve, they are increasingly expected to act not just as passive tools, but as collaborative partners that actively support human knowledge work. This thesis contributes to this vision by developing cognitively aligned AI systems that adapt to users' cognitive processes and proactively guide understanding. However, the proactive strategies explored in this work were bounded—limited to specific types of user inputs (e.g., explicit answers to system-generated questions or previously consumed content) and constrained by predefined interaction structures.

Looking forward, I aim to expand cognitive alignment toward more flexibly proactive and ambient support. This means developing systems that naturally integrate into users' ongoing workflows and anticipate their evolving cognitive needs without requiring constant prompting. Achieving this vision requires two key advances: (1) developing models that can infer cognitive processes from ambient interaction signals such as exploration patterns, writing behavior, or task history; and (2) designing ambient, context-aware interventions that support users' cognitive processes without interrupting or derailing their current task. This direction moves toward AI collaborators that are deeply attuned to the nuances of human understanding, supporting users subtly and at just the right moments.

8.2.2 Aligning AI Systems with Human Workflow in Complex Tasks

While this thesis focused on aligning AI systems with human cognitive processes in individual understanding tasks, future work will extend this approach to more complex workflows that involve multiple, interrelated tasks. These tasks often require distinct cognitive processes and produce intermediate artifacts such as notes, outlines, or experimental results, each contributing to the broader goal. Therefore, supporting these tasks requires modeling and integrating multiple cognitive processes that arise across different stages of a task.

For example, a full research project spans activities like reviewing prior work, formulating research questions, conducting experiments, and writing papers. I aim to design AI systems that are not limited to supporting a single step, but instead are structured to understand and assist across multiple stages—connecting related tasks, leveraging prior context, and offering support that is consistent with the overall workflow structure.

8.2.3 Sustaining Cognitive Alignment over Time

Real-world engagements with AI systems often unfold over extended periods—spanning weeks, months, or even years. In such long-term interactions, the challenges of alignment shift: users' cognitive goals evolve, their domain knowledge deepens, and their strategies for learning and problem-solving change. As users' cognitive processes that underlie their knowledge work evolve, supporting this temporal dimension of cognitive change introduces a new layer of complexity in alignment.

To extend cognitive alignment into long-term interactions, cognitively aligned systems need to distinguish between stable aspects of a user's cognitive profile, such as foundational knowledge or reasoning styles, and those that evolve, like shifting goals or emerging interests. While this thesis focused on adapting to short-term cognitive signals, long-term alignment requires systems to continually refine their models of users' cognitive processes over time, deciding what to retain, update, or forget.

This direction also introduces new challenges. Long-term cognitive modeling involves collecting and reasoning over rich user traces, raising questions of privacy, user agency, and transparency. Future cognitively aligned systems should not only support evolving understanding, but do so in a way that allows users to trust and manage how their cognitive information is represented and used.

Bibliography

[1] Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2022. ASPECTNEWS: Aspect-Oriented Summarization of News Documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 6494–6506. https://doi.org/10.18653/v1/2022.acl-long.449

[2] Shaaron Ainsworth. 2006. DeFT: A conceptual framework for considering learning with multiple representations. *Learning and instruction* 16, 3 (2006), 183–198.

[3] Shaaron Ainsworth. 2008. The educational value of multiple-representations when learning complex scientific concepts. In *Visualization: Theory and practice in science education*. Springer, 191–208.

[4] Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-Domain Question Answering Goes Conversational via Question Rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 520–534.

[5] Anthropic. 2023. *Introducing Claude 2.1*. https://www.anthropic.com/index/claude-2-1 Accessed: 2023-11-21.

[6] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwan, Yoshua Bengio, Danqi Chen, Philip H. S. Torr, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. 2024. Foundational Challenges in Assuring Alignment and Safety of Large Language Models. arXiv:2404.09932 [cs.LG] https://arxiv.org/abs/2404.09932

[7] Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. 2021. ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning. In *International Conference on Learning Representations*.

[8] D. Ausubel. 2000. The Acquisition and Retention of Knowledge: A Cognitive View.

[9] David G. Ausubel. 1963. Cognitive Structure and the Facilitation of Meaningful Verbal Learning1. *Journal of Teacher Education* 14, 2 (1963), 217–222. https://doi.org/10.1177/002248716301400220 arXiv:https://doi.org/10.1177/002248716301400220

[10] David P. Ausubel. 1962. A Subsumption Theory of Meaningful Verbal Learning and Retention. *The Journal of General Psychology* 66, 2 (1962), 213–224. https://doi.org/10.1080/00221309.1962.9711837 arXiv:https://doi.org/10.1080/00221309.1962.9711837 PMID: 13863333.

[11] David Paul Ausubel. 2012. *The acquisition and retention of knowledge: A cognitive view*. Springer Science & Business Media.

[12] Fan Bai, Junmo Kang, Gabriel Stanovsky, Dayne Freitag, and Alan Ritter. 2023. Schema-Driven Information Extraction from Heterogeneous Tables. *ArXiv* abs/2305.14336 (2023). https://api.semanticscholar.org/CorpusID:258841398

[13] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL] https://arxiv.org/abs/2212.08073

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

[15] Hritik Bansal, John Dang, and Aditya Grover. 2024. Peering Through Preferences: Unraveling Feedback Acquisition for Aligning Large Language Models. arXiv:2308.15812 [cs.LG] https://arxiv.org/abs/2308.15812

[16] Tal Baumel, Matan Eyal, and Michael Elhadad. 2018. Query Focused Abstractive Summarization: Incorporating Query Relevance, Multi-Document Coverage, and Summary Length Constraints into seq2seq Models. https://doi.org/10.48550/ARXIV.1801.07704

[17] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:202558505

[18] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. https://doi.org/10.48550/ARXIV.2004.05150

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

[20] Ann M Blair. 2010. *Too much to know: Managing scholarly information before the modern age*. Yale University Press.

[21] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[22] Elizabeth Baraff Bonawitz and Tania Lombrozo. 2012. Occam's rattle: children's use of simplicity and probability to constrain inference. *Developmental psychology* 48, 4 (2012), 1156.

[23] 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 (2021), 1–15. https://nature.com/articles/s41599-021-00903-w

[24] Richard E. Boyatzis. 1998. Transforming Qualitative Information: Thematic Analysis and Code Development.

[25] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/ARXIV.2005.14165

[26] Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. *arXiv preprint arXiv:1909.05358* (2019).

[27] Sonia Q. Cabell, Laura M. Justice, Anita S. McGinty, Jamie DeCoster, and Lindsay D. Forston. 2015. Teacher–child conversations in preschool classrooms: Contributions to children's vocabulary development. *Early Childhood Research Quarterly* 30 (2015), 80–92. https://doi.org/10.1016/j.ecresq.2014.09.004

[28] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S. Weld. 2020. TLDR: Extreme Summarization of Scientific Documents. *ArXiv* abs/2004.15011 (2020). https://api.semanticscholar.org/CorpusID:216867622

[29] William Cai, Hao Sheng, and Sharad Goel. 2020. MathBot: A Personalized Conversational Agent for Learning Math.

[30] Ben Carterette, E. Kanoulas, and Emine Yilmaz. 2010. Low cost evaluation in information retrieval. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (2010).

[31] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. SOLVENT: A Mixed Initiative System for Finding Analogies between Research Papers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 31 (Nov. 2018), 21 pages. https://doi.org/10.1145/3274300

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

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

[34] Mingda Chen, Sam Wiseman, and Kevin Gimpel. 2021. WikiTableT: A Large-Scale Data-to-Text Dataset for Generating Wikipedia Article Sections. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 193–209. https://doi.org/10.18653/v1/2021.findings-acl.17

[35] Yi Cheng, Kate Yen, Yeqi Chen, Sijin Chen, and Alexis Hiniker. 2018. Why Doesn't It Work? Voice-Driven Interfaces and Young Children's Communication Repair Strategies. In *Proceedings of the 17th ACM Conference on Interaction Design and Children* (Trondheim, Norway) *(IDC '18)*. Association for Computing Machinery, New York, NY, USA, 337–348. https://doi.org/10.1145/3202185.3202749

[36] Michelene T.H. Chi, Miriam Bassok, Matthew W. Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science* 13, 2 (1989), 145–182. https://doi.org/10.1016/0364-0213(89)90002-5

[37] Michelene T. H. Chi and Ruth Wylie. 2014. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist* 49, 4 (2014), 219–243. https://doi.org/10.1080/00461520.2014.965823 arXiv:https://doi.org/10.1080/00461520.2014.965823

[38] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2174–2184. https://doi.org/10.18653/v1/D18-1241

[39] 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 (2021), 447–461. https://doi.org/10.1162/tacl_a_00377

[40] Michelle M Chouinard, Paul L Harris, and Michael P Maratsos. 2007. Children's questions: A mechanism for cognitive development. *Monographs of the society for research in child development* (2007), i–129.

[41] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).

[42] Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural Scaffolds for Citation Intent Classification in Scientific Publications. *ArXiv* abs/1904.01608 (2019). https://api.semanticscholar.org/CorpusID:102483154

[43] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2270–2282. https://doi.org/10.18653/v1/2020.acl-main.207

[44] Lynne M. Connelly. 2013. Grounded theory. *Medsurg nursing : official journal of the Academy of Medical-Surgical Nurses* 22 2 (2013), 124, 127.

[45] Kathleen H. Corriveau and Katelyn E. Kurkul. 2014. "Why Does Rain Fall?": Children Prefer to Learn From an Informant Who Uses Noncircular Explanations. *Child Development* 85, 5 (2014), 1827–1835. http://www.jstor.org/stable/24033022

[46] Catherine Crain-Thoreson, Michael P Dahlin, and Terris A Powell. 2001. Parent-child interaction in three conversational contexts: Variations in style and strategy. *New directions for child and adolescent development* 2001, 92 (2001), 23–38. https://doi.org/10.1002/cd.13

[47] Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Zhao, Aida Amini, Mike Green, Qazi Rashid, and Kelvin Guu. 2022. Dialog Inpainting: Turning Documents to Dialogs. In *International Conference on Machine Learning (ICML)*. PMLR.

[48] Judith H. Danovitch, Candice M. Mills, Kaitlin R. Sands, and Allison J. Williams. 2021. Mind the gap: How incomplete explanations influence children's interest and learning behaviors. *Cognitive Psychology* 130 (2021), 101421. https://doi.org/10.1016/j.cogpsych.2021.101421

[49] 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 *North American Chapter of the Association for Computational Linguistics*.

[50] 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 *NAACL*.

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

[52] Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. (01 2010).

[53] Paul Denny, Hassan Khosravi, Arto Hellas, Juho Leinonen, and Sami Sarsa. 2023. Can We Trust AI-Generated Educational Content? Comparative Analysis of Human and AI-Generated Learning Resources. arXiv:2306.10509 [cs.HC] https://arxiv.org/abs/2306.10509

[54] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[55] Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MSˆ2: Multi-Document Summarization of Medical Studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 7494–7513. https://doi.org/10.18653/v1/2021.emnlp-main.594

[56] Griffin Dietz, Zachary Pease, Brenna McNally, and Elizabeth Foss. 2020. Giggle Gauge: A Self-Report Instrument for Evaluating Children's Engagement with Technology. In *Proceedings of the Interaction Design and Children Conference* (London, United Kingdom) *(IDC '20)*. Association for Computing Machinery, New York, NY, USA, 614–623. https://doi.org/10.1145/3392063.3394393

[57] Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. 2017. "Hey Google is It OK If I Eat You?": Initial Explorations in Child-Agent Interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children* (Stanford, California, USA) *(IDC '17)*. Association for Computing Machinery, New York, NY, USA, 595–600. https://doi.org/10.1145/3078072.3084330

[58] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2368–2378.

[59] Gerald G. Duffy, Laura R. Roehler, Michael S. Meloth, and Linda G. Vavrus. 1986. Conceptualizing instructional explanation. *Teaching and Teacher Education* 2, 3 (1986), 197–214. https://doi.org/10.1016/S0742-051X(86)80002-6

[60] Jonathan St BT Evans. 2012. Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning* 18, 1 (2012), 5–31.

[61] Mary Evans, Shelley Moretti, Deborah Shaw, and Maureen Fox. 2003. Parent Scaffolding in Children's Oral Reading. *Early Education and Development - EARLY EDUC DEV* 14 (07 2003), 363–388. https://doi.org/10.1207/s15566935eed1403_5

[62] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3558–3567. https://doi.org/10.18653/v1/P19-1346

[63] Raymond Fok, Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Andrew Head, Marti A. Hearst, and Daniel S. Weld. 2022. Scim: Intelligent Skimming Support for Scientific Papers. *Proceedings of the 28th International Conference on Intelligent User Interfaces* (2022). https://api.semanticscholar.org/CorpusID:254591867

[64] Brandy N. Frazier, Susan A. Gelman, and Henry M. Wellman. 2009. Preschoolers' Search for Explanatory Information within Adult: Child Conversation. *Child Development* 80, 6 (2009), 1592–1611. http://www.jstor.org/stable/25592097

[65] Brandy N Frazier, Susan A Gelman, and Henry M Wellman. 2016. Young children prefer and remember satisfying explanations. *Journal of Cognition and Development* 17, 5 (2016), 718–736.

[66] Krzysztof Z. Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 794–806. https://doi.org/10.1145/3490099.3511138

[67] Kyle Yingkai Gao and Jamie Callan. 2017. Scientific Table Search Using Keyword Queries. *ArXiv* abs/1707.03423 (2017). https://api.semanticscholar.org/CorpusID:9964828

[68] Radhika Garg, Hua Cui, Spencer Seligson, Bo Zhang, Martin Porcheron, Leigh Clark, Benjamin R. Cowan, and Erin Beneteau. 2022. The Last Decade of HCI Research on Children and Voice-Based Conversational Agents. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 149, 19 pages. https://doi.org/10.1145/3491102.3502016

[69] Susan A Gelman. 1988. The development of induction within natural kind and artifact categories. *Cognitive Psychology* 20, 1 (1988), 65–95. https://doi.org/10.1016/0010-0285(88)90025-4

[70] Andrea Gemelli, Emanuele Vivoli, and Simone Marinai. 2023. CTE: A Dataset for Contextualized Table Extraction. arXiv:2302.01451 [cs.CL]

[71] Dedre Gentner and Russell Landers. 1985. ANALOGICAL REMINDING: A GOOD MATCH IS HARD TO FIND.. In *Unknown Host Publication Title*. IEEE, 607–613.

[72] Mary L Gick and Keith J Holyoak. 1980. Analogical problem solving. *Cognitive psychology* 12, 3 (1980), 306–355.

[73] Mary L. Gick and Keith J. Holyoak. 1983. Schema induction and analogical transfer. *Cognitive Psychology* 15, 1 (1983), 1 – 38. https://doi.org/10.1016/0010-0285(83)90002-6

[74] John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Lu Wang, and Arman Cohan. 2022. Towards multi-document summarization in the open-domain. https://api.semanticscholar.org/CorpusID:258865156

[75] Usha Goswami. 2001. *Analogical Reasoning in Children*. 437 – 470.

[76] Arthur C Graesser, Natalie Person, and John Huber. 1992. Mechanisms that generate questions. *Questions and information systems* 2 (1992), 167–187.

[77] Arthur C Graesser and Natalie K Person. 1994. Question asking during tutoring. *American educational research journal* 31, 1 (1994), 104–137.

[78] Tanishq Gupta, Mohd Zaki, Devanshi Khatsuriya, Kausik Hira, N M Anoop Krishnan, and Mausam. 2023. DiSCoMaT: Distantly Supervised Composition Extraction from Tables in Materials Science Articles. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 13465–13483. https://doi.org/10.18653/v1/2023.acl-long.753

[79] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*. PMLR, 3929–3938.

[80] Graeme S. Halford. 2009. *Children's understanding: The development of mental models*. Erlbaum.

[81] Paul L Harris. 2012. *Trusting what you're told: How children learn from others*. Harvard University Press.

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

[83] Hayato Hashimoto, Kazutoshi Shinoda, Hikaru Yokono, and Akiko Aizawa. 2017. Automatic Generation of Review Matrices as Multi-document Summarization of Scientific Papers. In *BIRNDL@SIGIR*. https://api.semanticscholar.org/CorpusID:44200100

[84] Paul Hemp. 2009. Death by information overload. *Harvard business review* 87 9 (2009), 82–9, 121. https://api.semanticscholar.org/CorpusID:584292

[85] Julian PT Higgins, Sally Green, et al. 2008. Cochrane handbook for systematic reviews of interventions. (2008).

[86] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor. *arXiv preprint arXiv:2212.09689* (2022).

[87] Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating Innovation Through Analogy Mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. ACM, New York, NY, USA, 235–243. https://doi.org/10.1145/3097983.3098038

[88] Tom Hope, Ronen Tamari, Daniel Hershcovich, Hyeonsu B Kang, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2022. Scaling creative inspiration with fine-grained functional aspects of ideas. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15. https://dl.acm.org/doi/10.1145/3491102.3517434

[89] Jiangbo Hu, Camilla Gordon, Ning Yang, and Yonggang Ren. 2020. "Once Upon A Styoar": A Science Education Program Based on Personification Storytelling in Promoting Preschool Children's Understanding of Astronomy Concepts. *Early Education and Development* 32 (05 2020), 1–19. https://doi.org/10.1080/10409289.2020.1759011

[90] Kayoko Inagaki and Giyoo Hatano. 1987. Young Children's Spontaneous Personification as Analogy. *Child Development* 58, 4 (1987), 1013–1020. http://www.jstor.org/stable/1130542

[91] Gautier Izacard and Édouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 874–880.

[92] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299* (2022).

[93] Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Joshua B. Tenenbaum, and Bernhard Schölkopf. 2022. When to Make Exceptions: Exploring Language Models as Accounts of Human Moral Judgment. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=uP9RiC4uVcR

[94] Arif E Jinha. 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned publishing* 23, 3 (2010), 258–263. https://onlinelibrary.wiley.com/doi/abs/10.1087/20100308

[95] Joan N Kaderavek, Ying Guo, and Laura M Justice. 2014. Validity of the children's orientation to book reading rating scale. *Journal of Research in Reading* 37, 2 (2014), 159–178.

[96] Hyeonsu Kang, Rafal Kocielnik, Andrew Head, Jiangjiang Yang, Matt Latzke, Aniket Kittur, Daniel Weld S., Doug Downey, and Jonathan Bragg. 2022. From Who You Know to What You Read: Augmenting Scientific Recommendations with Implicit Social Networks. (2022), 1–16.

[97] Hyeonsu Kang, Sherry Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. 2023. Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, 19 pages. https://doi.org/10.1145/3586183.3606759

[98] Hyeonsu B Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An Interactive System for Personalized Thread-based Exploration and Organization of Scientific Literature. *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (2022). https://api.semanticscholar.org/CorpusID:251402552

[99] Hyeonsu B Kang, Sheshera Mysore, Kevin J Huang, Haw-Shiuan Chang, Thorben Prein, Andrew McCallum, Aniket Kittur, and Elsa Olivetti. 2022. Augmenting Scientific Creativity with Retrieval across Knowledge Domains. In *Second Workshop on Bridging Human-Computer Interaction and Natural Language Processing at NAACL 2022*. arXiv. https://doi.org/10.48550/ARXIV.2206.01328

[100] Hyeonsu B. Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. 2022. Augmenting Scientific Creativity with an Analogical Search Engine. *ACM Trans. Comput.-Hum. Interact.* (mar 2022). https://doi.org/10.1145/3530013 Just Accepted.

[101] Hyeonsu B Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. 2022. Augmenting scientific creativity with an analogical search engine. *ACM Transactions on Computer-Human Interaction* 29, 6 (2022), 1–36. https://dl.acm.org/doi/10.1145/3530013

[102] Hyeonsu B Kang, Nouran Soliman, Matt Latzke, Joseph Chee Chang, and Jonathan Bragg. 2023. ComLittee: Literature Discovery with Personal Elected Author Committees. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 738, 20 pages. https://doi.org/10.1145/3544548.3581371

[103] Hyeonsu B Kang, Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. 2023. Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–19. https://dl.acm.org/doi/10.1145/3586183.3606759

[104] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6769–6781.

[105] Harmanpreet Kaur, Doug Downey, Amanpreet Singh, Evie Yu-Yen Cheng, Daniel S. Weld, and Jonathan Bragg. 2022. FeedLens: Polymorphic Lenses for Personalizing Exploratory Search over Knowledge Graphs *(UIST '22)*.

[106] Deborah Kelemen. 2019. The Magic of Mechanism: Explanation-Based Instruction on Counterintuitive Concepts in Early Childhood. *Perspectives on Psychological Science* 14 (04 2019), 174569161982701. https://doi.org/10.1177/1745691619827011

[107] James Kennedy, Séverin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) *(HRI '17)*. Association for Computing Machinery, New York, NY, USA, 82–90. https://doi.org/10.1145/2909824.3020229

[108] Tae Soo Kim, Yoonjoo Lee, Minsuk Chang, and Juho Kim. 2023. Cells, Generators, and Lenses: Design Framework for Object-Oriented Interaction with Large Language Models. In *Proceedings of the 36th*

Annual ACM Symposium on User Interface Software and Technology (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 4, 18 pages. https://doi.org/10.1145/3586183.3606833

[109] Yea-Seul Kim, Jessica Hullman, Matthew Burgess, and Eytan Adar. 2016. SimpleScience: Lexical Simplification of Scientific Terminology. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 1066–1071. https://doi.org/10.18653/v1/D16-1114

[110] Alison King. 1994. Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American educational research journal* 31, 2 (1994), 338–368.

[111] Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140* (2023). https://arxiv.org/abs/2301.10140

[112] Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking Cognitive Biases in Large Language Models as Evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 517–545. https://doi.org/10.18653/v1/2024.findings-acl.29

[113] Keti Korini, Ralph Peeters, and Christian Bizer. 2022. SOTAB: the WDC schema. org table annotation benchmark. In *CEUR Workshop Proceedings*, Vol. 3320. RWTH Aachen, 14–19. https://ceur-ws.org/Vol-3320/paper1.pdf

[114] Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to Progress in Long-form Question Answering. arXiv:2103.06332 [cs.CL] https://arxiv.org/abs/2103.06332

[115] Christina Krist, Christina V. Schwarz, and Brian J. Reiser. 2019. Identifying Essential Epistemic Heuristics for Guiding Mechanistic Reasoning in Science Learning. *Journal of the Learning Sciences* 28, 2 (2019), 160–205. https://doi.org/10.1080/10508406.2018.1510404 arXiv:https://doi.org/10.1080/10508406.2018.1510404

[116] Christoph Kulgemeyer. 2018. Towards a framework for effective instructional explanations in science teaching. *Studies in Science Education* 54, 2 (2018), 109–139. https://doi.org/10.1080/03057267.2018.1598054 arXiv:https://doi.org/10.1080/03057267.2018.1598054

[117] Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. AQuaMuSe: Automatically Generating Datasets for Query-Based Multi-Document Summarization. https://doi.org/10.48550/ARXIV.2010.12694

[118] Katelyn E Kurkul, Eleanor Castine, Kathryn Leech, and Kathleen H Corriveau. 2021. How does a switch work? The relation between adult mechanistic language and children's learning. *Journal of Applied Developmental Psychology* 72 (2021), 101221. https://doi.org/10.1016/j.appdev.2020.101221

[119] Katelyn E Kurkul and Kathleen H Corriveau. 2018. Question, explanation, follow-up: A mechanism for learning from others? *Child Development* 89, 1 (2018), 280–294.

[120] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 452–466. https://doi.org/10.1162/tacl_a_00276

[121] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 452–466. https://doi.org/10.1162/tacl_a_00276

[122] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.

[123] Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, and Jilin Chen. 2023. Improving Diversity of Demographic Representation in Large Language Models via Collective-Critiques and Self-Voting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10383–10405. https://doi.org/10.18653/v1/2023.emnlp-main.643

[124] Moontae Lee, Xiaodong He, Wen-tau Yih, Jianfeng Gao, Li Deng, and Paul Smolensky. 2016. Reasoning in Vector Space: An Exploratory Study of Question Answering. (2016).

[125] Yoonjoo Lee, John Joon Young Chung, Tae Soo Kim, Jean Y Song, and Juho Kim. 2022. Promptiverse: Scalable Generation of Scaffolding Prompts Through Human-AI Hybrid Knowledge Graph Annotation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 96, 18 pages. https://doi.org/10.1145/3491102.3502087

[126] Yoonjoo Lee, Hyeonsu B Kang, Matt Latzke, Juho Kim, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. PaperWeaver: Enriching Topical Paper Alerts by Contextualizing Recommended Papers with User-collected Papers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19. https://dl.acm.org/doi/10.1145/3613904.3642196

[127] Yoonjoo Lee, Hyeonsu B Kang, Matt Latzke, Juho Kim, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. PaperWeaver: Enriching Topical Paper Alerts by Contextualizing Recommended Papers with User-collected Papers. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 19, 19 pages. https://doi.org/10.1145/3613904.3642196

[128] Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 2023. DAPIE: Interactive Step-by-Step Explanatory Dialogues to Answer Children's Why and How Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–22.

[129] Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. 2023. QASA: advanced question answering on scientific articles. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 787, 17 pages. https://proceedings.mlr.press/v202/lee23n.html

[130] Yoonjoo Lee, Nedim Lipka, Zichao Wang, Ryan Rossi, Puneet Mathur, Tong Sun, and Alexa Siu. 2025. DocVoyager: Anticipating User's Information Needs and Guiding Document Reading through Question Answering. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 196, 8 pages. https://doi.org/10.1145/3706599.3719846

[131] Cristine H Legare. 2014. The contributions of explanation and exploration to children's scientific reasoning. *Child Development Perspectives* 8, 2 (2014), 101–106. https://doi.org/10.1111/cdep.12070

[132] Cristine H Legare and Tania Lombrozo. 2014. Selective effects of explanation on learning during early childhood. *Journal of experimental child psychology* 126 (2014), 198–212.

[133] Gaea Leinhardt, Kevin Crowley, and Karen Knutson. 2015. *Building islands of expertise in everyday family activity*. Routledge.

[134] Gaea Leinhardt and Michael Steele. 2005. Seeing the Complexity of Standing to the Side: Instructional Dialogues. *Cognition and Instruction - COGNITION INSTRUCT* 23 (03 2005), 87–163. https://doi.org/10.1207/s1532690xci2301_4

[135] Dani Levine, Amy Pace, Rufan Luo, Kathy Hirsh-Pasek, Roberta Michnick Golinkoff, Jill de Villiers, Aquiles Iglesias, and Mary Sweig Wilson. 2020. Evaluating socioeconomic gaps in preschoolers' vocabulary, syntax and language process skills with the Quick Interactive Language Screener (QUILS). *Early Childhood Research Quarterly* 50 (2020), 114–128.

[136] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.

[137] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[138] Huihan Li, Tianyu Gao, Manan Goenka, and Danqi Chen. 2021. Ditch the Gold Standard: Re-evaluating Conversational Question Answering. *ArXiv* abs/2112.08812 (2021). https://api.semanticscholar.org/CorpusID:245218415

[139] Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-turn Comparisons. https://doi.org/10.48550/ARXIV.1909.03087

[140] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957* (2017).

[141] Joseph CR Licklider. 1960. Man-computer symbiosis. *IRE transactions on human factors in electronics* 1 (1960), 4–11.

[142] Paul Light and George Butterworth. 2017. *Chapter 7. Desituating cognition through the construction of conceptual knowledge*. Routledge.

[143] Mike E.U. Ligthart, Mark A. Neerincx, and Koen V. Hindriks. 2020. Design Patterns for an Interactive Storytelling Robot to Support Children's Engagement and Agency. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 409–418. https://doi.org/10.1145/3319502.3374826

[144] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[145] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[146] Ching Liu, Juho Kim, and Hao-Chuan Wang. 2018. *ConceptScape: Collaborative Concept Mapping for Video Learning*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173961

[147] Hongbo Liu, Mingda Wu, Kun Yuan, Ming Sun, Yansong Tang, Chuanchuan Zheng, Xing Wen, and Xiu Li. 2023. Ada-DQA: Adaptive Diverse Quality-aware Feature Acquisition for Video Quality Assessment. In *Proceedings of the 31st ACM International Conference on Multimedia* (Ottawa ON, Canada) *(MM '23)*. Association for Computing Machinery, New York, NY, USA, 6695–6704. https://doi.org/10.1145/3581783.3611795

[148] Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=BTKAeLqLMw

[149] Ye Liu, Kazuma Hashimoto, Yingbo Zhou, Semih Yavuz, Caiming Xiong, and S Yu Philip. 2021. Dense Hierarchical Retrieval for Open-domain Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 188–200.

[150] 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, F.Q. Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Michael Kinney, Aniket Kittur, Hyeonsu B Kang, Egor Klevak, Bailey Kuehl, Michael Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Stuart Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita R Rao, Paul L Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline M Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. 2023. The Semantic Reader Project: Augmenting Scholarly Documents through AI-Powered Interactive Reading Interfaces. *ArXiv* abs/2303.14334 (2023). https://api.semanticscholar.org/CorpusID:257766269

[151] 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

[152] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Michael Kinney, and Daniel S. Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:215416146

[153] Silvia B. Lovato, Anne Marie Piper, and Ellen A. Wartella. 2019. Hey Google, Do Unicorns Exist? Conversational Agents as a Path to Answers to Children's Questions. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children* (Boise, ID, USA) *(IDC '19)*. Association for Computing Machinery, New York, NY, USA, 301–313. https://doi.org/10.1145/3311927.3323150

[154] Xinyi Lu, Simin Fan, Jessica Houghton, Lu Wang, and Xu Wang. 2023. ReadingQuizMaker: A Human-NLP Collaborative System that Supports Instructors to Design High-Quality Reading Quiz Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 454, 18 pages. https://doi.org/10.1145/3544548.3580957

[155] Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A Large-scale Dataset for Extreme Multi-document Summarization of Scientific Articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 8068–8074. https://doi.org/10.18653/v1/2020.emnlp-main.648

[156] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 3219–3232. https://doi.org/10.18653/v1/D18-1360

[157] Kelvin Luu, Xinyi Wu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A. Smith. 2020. Explaining Relationships Between Scientific Documents. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:236459799

[158] Jessica Maghakian, Paul Mineiro, Kishan Panaganti, Mark Rucker, Akanksha Saran, and Cheng Tan. 2023. Personalized Reward Learning with Interaction-Grounded Learning (IGL). In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=wGvzQWFyUB

[159] Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9004–9017. https://doi.org/10.18653/v1/2023.emnlp-main.557

[160] Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2021. MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases. *arXiv preprint arXiv:2005.00352* (2021).

[161] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661* (2020).

[162] Candice M. Mills, Judith. H. Danovitch, Victoria N. Mugambi, Kaitlin R. Sands, and Candice Pattisapu Fox. 2022. "Why do dogs pant?": Characteristics of parental explanations about science predict children's knowledge. *Child Development* 93, 2 (2022), 326–340. https://doi.org/10.1111/cdev.13681 arXiv:https://srcd.onlinelibrary.wiley.com/doi/pdf/10.1111/cdev.13681

[163] Candice M Mills, Kaitlin R Sands, Sydney P Rowles, and Ian L Campbell. 2019. "I want to know more!": Children are sensitive to explanation quality when exploring new information. *Cognitive Science* 43, 1 (2019), e12706.

[164] Mistral AI. 2024. Mixtral of Experts. https://mistral.ai/news/mixtral-of-experts/. Accessed on March 25, 2024.

[165] Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. SciGen: a Dataset for Reasoning-Aware Text Generation from Scientific Tables. In *NeurIPS Datasets and Benchmarks*. https://api.semanticscholar.org/CorpusID:244906308

[166] Roxana Moreno and Richard Mayer. 2007. Interactive Multimodal Learning Environments. *Educ Psychol Rev* 19 (09 2007), 309–326. https://doi.org/10.1007/s10648-007-9047-2

[167] Sonia K. Murthy, Kyle Lo, Daniel King, Chandra Bhagavatula, Bailey Kuehl, Sophie Johnson, Jon Borchardt, Daniel S. Weld, Tom Hope, and Doug Downey. 2022. ACCoRD: A Multi-Document Approach to Generating Diverse Descriptions of Scientific Concepts. *ArXiv* abs/2205.06982 (2022). https://api.semanticscholar.org/CorpusID:248811750

[168] Sheshera Mysore, Timothy J. O'Gorman, Andrew McCallum, and Hamed Zamani. 2021. CSFCube - A Test Collection of Computer Science Research Articles for Faceted Query by Example. *ArXiv* abs/2103.12906 (2021). https://api.semanticscholar.org/CorpusID:232335540

[169] 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. 2021. WebGPT: Browser-assisted question-answering with human feedback. *CoRR* abs/2112.09332 (2021). arXiv:2112.09332 https://arxiv.org/abs/2112.09332

[170] Preslav I Nakov, Ariel S Schwartz, Marti Hearst, et al. 2004. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR*, Vol. 4. Citeseer, 81–88.

[171] Benjamin Newman, Yoonjoo Lee, Aakanksha Naik, Pao Siangliulue, Raymond Fok, Juho Kim, Daniel S Weld, Joseph Chee Chang, and Kyle Lo. 2024. ArxivDIGESTables: Synthesizing Scientific Literature into Tables using Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 9612–9631. https://doi.org/10.18653/v1/2024.emnlp-main.538

[172] Benjamin Newman, Luca Soldaini, Raymond Fok, Arman Cohan, and Kyle Lo. 2023. A Question Answering Framework for Decontextualizing User-facing Snippets from Scientific Documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino,

and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3194–3212. https://doi.org/10.18653/v1/2023.emnlp-main.193

[173] Nielsen. 2018. (Smart) speaking my language: Despite their vast capabilities, smart speakers are all about the music.

[174] J. Novak. 2002. Meaningful learning: The essential factor for conceptual change in limited or inappropriate propositional hierarchies leading to empowerment of learners.

[175] Joseph D. Novak and Alberto J. Cañas. 2006. *The Theory Underlying Concept Maps and How to Construct and Use Them.* research report 2006-01 Rev 2008-01. Florida Institute for Human and Machine Cognition. http://cmap.ihmc.us/Publications/ResearchPapers/TheoryCmaps/TheoryUnderlyingConceptMaps.htm

[176] Angela M. O'Donnell, Donald F. Dansereau, and Richard H. Hall. 2002. Knowledge Maps as Scaffolds for Cognitive Processing. *Educational Psychology Review* 14, 1 (01 Mar 2002), 71–86. https://doi.org/10.1023/A:1013132527007

[177] Open AI. 2022. Introducing ChatGPT. https://openai.com/index/chatgpt/. Accessed on March 25, 2024.

[178] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL] https://arxiv.org/abs/2203.02155

[179] Elisabeth Pain. 2016. How to keep up with the scientific literature. *Science* (2016). https://api.semanticscholar.org/CorpusID:158399837

[180] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding Literature Reviews with Existing Related Work Sections. *arXiv preprint arXiv:2302.06754* (2023).

[181] Anusri Pampari, Preethi Raghavan, Jennifer J. Liang, and Jian Peng. 2018. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In *Conference on Empirical Methods in Natural Language Processing*.

[182] Dimitris Pappas, Ion Androutsopoulos, and Haris Papageorgiou. 2018. BioRead: A New Dataset for Biomedical Reading Comprehension. In *International Conference on Language Resources and Evaluation*.

[183] Haekyu Park, Gonzalo A. Ramos, Jina Suh, Christopher Meek, Rachel Ng, and Mary Czerwinski. 2023. FoundWright: A System to Help People Re-find Pages from Their Web-history. *ArXiv* abs/2305.07930 (2023). https://api.semanticscholar.org/CorpusID:258685533

[184] Markus Peschl, Arkady Zgonnikov, Frans A. Oliehoek, and Luciano C. Siebert. 2021. MORAL: Aligning AI with Human Norms through Multi-Objective Reinforced Active Learning. arXiv:2201.00012 [cs.LG] https://arxiv.org/abs/2201.00012

[185] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. AngleKindling: Supporting Journalistic Angle Ideation with Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 225, 16 pages. https://doi.org/10.1145/3544548.3580907

[186] Antoine Ponsard, Francisco Escalona, and Tamara Munzner. 2016. PaperQuest: A visualization tool to support literature review. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2264–2271.

[187] Jason Portenoy, Marissa Radensky, Jevin D West, Eric Horvitz, Daniel S Weld, and Tom Hope. 2022. Bursting Scientific Filter Bubbles: Boosting Innovation via Novel Author Discovery. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 309, 13 pages. https://doi.org/10.1145/3491102.3501905

[188] 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*. 707–719.

[189] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[190] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG]

[191] Janarthanan Rajendran, Jatin Ganhotra, Satinder Singh, and Lazaros Polymenakos. 2018. Learning End-to-End Goal-Oriented Dialog with Multiple Answers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

[192] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 784–789.

[193] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2383–2392.

[194] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* 7 (2019), 249–266. https://doi.org/10.1162/tacl_a_00266

[195] Darrel A Regier, William E Narrow, Diana E Clarke, Helena C Kraemer, S Janet Kuramoto, Emily A Kuhl, and David J Kupfer. 2013. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. *American journal of psychiatry* 170, 1 (2013), 59–70.

[196] 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*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[197] Melissa N. Richards and Sandra L. Calvert. 2017. Measuring young U.S. children's parasocial relationships: toward the creation of a child self-report survey. *Journal of Children and Media* 11, 2 (2017), 229–240. https://doi.org/10.1080/17482798.2017.1304969 arXiv:https://doi.org/10.1080/17482798.2017.1304969

[198] D. Robinson and Kenneth A. Kiewra. 1995. Visual argument: Graphic organizers are superior to outlines in improving learning from text. *Journal of Educational Psychology* 87 (1995), 455–467.

[199] Laura R Roehler and Danise J Cantlon. 1997. Scaffolding: A powerful tool in social constructivist classrooms. (1997).

[200] Rod D. Roscoe and Michelene T. H. Chi. 2007. Understanding Tutor Learning: Knowledge-Building and Knowledge-Telling in Peer Tutors' Explanations and Questions. *Review of Educational Research* 77, 4 (2007), 534–574. https://doi.org/10.3102/0034654307309920 arXiv:https://doi.org/10.3102/0034654307309920

[201] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. *QuizBot: A Dialogue-Based Adaptive Learning System for Factual Knowledge*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300587

[202] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. 2018. Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 159 (jan 2018), 23 pages.

[203] Rosemary Russ, Rachel Scherr, David Hammer, and Jamie Mikeska. 2008. Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education* 92 (05 2008), 499 – 525. https://doi.org/10.1002/sce.20264

[204] Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 269–276. https://doi.org/10.1145/169059.169209

[205] Tarek Saier, Johan Krause, and Michael Färber. 2023. unarXive 2022: All arXiv Publications Pre-Processed for NLP, Including Structured Full-Text and Citation Network. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE Computer Society, Los Alamitos, CA, USA, 66–70. https://doi.org/10.1109/JCDL57899.2023.00020

[206] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. LaMP: When Large Language Models Meet Personalization. *arXiv preprint arXiv:2304.11406* (2023).

[207] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. 2021. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*.

[208] Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing Design Biases of Datasets and Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 9080–9102. https://doi.org/10.18653/v1/2023.acl-long.505

[209] Sami Sarsa, Paul Denny, Arto Hellas, and Juho Leinonen. 2022. Automatic Generation of Programming Exercises and Code Explanations Using Large Language Models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research - Volume 1* (Lugano and Virtual Event, Switzerland) *(ICER '22)*. Association for Computing Machinery, New York, NY, USA, 27–43. https://doi.org/10.1145/3501385.3543957

[210] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 857–868. https://doi.org/10.1145/3196709.3196772

[211] Darsh J. Shah, L. Yu, Tao Lei, and Regina Barzilay. 2021. Nutri-bullets Hybrid: Consensual Multi-document Summarization. In *North American Chapter of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:235097327

[212] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. 2012. Metro maps of science. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1122–1130. https://dl.acm.org/doi/10.1145/2339530.2339706

[213] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. 2024. Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. arXiv:2406.09264 [cs.HC] https://arxiv.org/abs/2406.09264

[214] Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. 2023. Measuring Inductive Biases of In-Context Learning with Underspecified Demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 11289–11310. https://doi.org/10.18653/v1/2023.acl-long.632

[215] Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. *ArXiv* abs/2211.13308 (2022).

[216] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large Language Models Encode Clinical Knowledge. *arXiv preprint arXiv:2212.13138* (2022).

[217] Catherine Snow. 1983. Literacy and language: Relationships during the preschool years. *Harvard educational review* 53, 2 (1983), 165–189.

[218] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A Roadmap to Pluralistic Alignment. In *ICML*. https://openreview.net/forum?id=gQpBnRHwxM

[219] Petros Stavropoulos, Dimitris Pappas, Ion Androutsopoulos, and Ryan T. McDonald. 2020. BioMRC: A Dataset for Biomedical Machine Reading Comprehension. *ArXiv* abs/2005.06376 (2020).

[220] Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid Questions Meet Long-Form Answers. *arXiv preprint arXiv:2204.06092* (2022).

[221] Hariharan Subramonyam, Colleen Seifert, Priti Shah, and Eytan Adar. 2020. *TexSketch: Active Diagramming through Pen-and-Ink Annotations*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376155

[222] Nicole Sultanum, Christine Murad, and Daniel Wigdor. 2020. Understanding and supporting academic literature review workflows with litsense. In *Proceedings of the International Conference on Advanced Visual Interfaces*. 1–5.

[223] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1630–1640. https://doi.org/10.18653/v1/P19-1159

[224] Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is Unfair: On Social Bias in Language Model-Based Metrics for Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3726–3739. https://doi.org/10.18653/v1/2022.emnlp-main.245

[225] Phillip Swazinna, Steffen Udluft, and Thomas Runkler. 2023. User-Interactive Offline Reinforcement Learning. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=a4COps0uokg

[226] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085* (2022).

[227] Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. 2024. Interactive AI Alignment: Specification, Process, and Evaluation Alignment. arXiv:2311.00710 [cs.HC] https://arxiv.org/abs/2311.00710

[228] Anuj Tewari and John Canny. 2014. What Did Spot Hide? A Question-Answering Game for Preschool Children. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1807–1816. https://doi.org/10.1145/2556288.2557205

[229] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang 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. arXiv:2201.08239 [cs.CL]

[230] H Holden Thorp. 2023. ChatGPT is fun, but not an author. , 313–313 pages.

[231] Barbara Tizard and Martin Hughes. 2008. *Young children learning*. John Wiley & Sons.

[232] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. 191–200.

[233] Takeo Tsujii and Shigeru Watanabe. 2009. Neural correlates of dual-task effect on belief-bias syllogistic reasoning: a near-infrared spectroscopy study. *Brain research* 1287 (2009), 118–125.

[234] Araceli Valle and Maureen Callanan. 2006. Similarity Comparisons and Relational Analogies in Parent-Child Conversations About Science Topics. *Merrill-Palmer Quarterly* 52 (01 2006), 96–124. https://doi.org/10.1353/mpq.2006.0009

[235] Stella Vosniadou and Marlene Schommer. 1988. Explanatory Analogies Can Help Children Acquire Information from Expository Text. Technical Report No. 460.

[236] Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

[237] Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. 2022. SQuALITY: Building a Long-Document Summarization Dataset the Hard Way. https://doi.org/10.48550/ARXIV.2205.11465

[238] Shufan Wang, Fangyuan Xu, Laure Thompson, Eunsol Choi, and Mohit Iyyer. 2022. Modeling Exemplification in Long-form Question Answering via Retrieval. 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, 2079–2092. https://doi.org/10.18653/v1/2022.naacl-main.151

[239] Xingbo Wang, Samantha L Huey, Rui Sheng, Saurabh Mehta, and Fei Wang. 2024. SciDaSynth: Interactive Structured Knowledge Extraction and Synthesis from Scientific Literature with Large Language Model. *arXiv preprint arXiv:2404.13765* (2024). https://arxiv.org/abs/2404.13765

[240] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-Instruct: Aligning Language Model with Self Generated Instructions. *arXiv preprint arXiv:2212.10560* (2022).

[241] Zichao Wang, Andrew S. Lan, Weili Nie, Andrew E. Waters, Phillip J. Grimaldi, and Richard G. Baraniuk. 2018. QG-net: a data-driven question generation model for educational content. In *Proceedings of the*

Fifth Annual ACM Conference on Learning at Scale (London, United Kingdom) *(L@S '18)*. Association for Computing Machinery, New York, NY, USA, Article 7, 10 pages. https://doi.org/10.1145/3231644.3231654

[242] Peter C Wason and J St BT Evans. 1974. Dual processes in reasoning? *Cognition* 3, 2 (1974), 141–154.

[243] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* (2021).

[244] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698* (2015).

[245] Angelica Willis, Glenn Davis, Sherry Ruan, Lakshmi Manoharan, James Landay, and Emma Brunskill. 2019. Key Phrase Extraction for Generating Educational Question-Answer Pairs. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale* (Chicago, IL, USA) *(L@S '19)*. Association for Computing Machinery, New York, NY, USA, Article 20, 10 pages. https://doi.org/10.1145/3330430.3333636

[246] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. *Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376781

[247] Chris Woolston. 2019. PhDs: the tortuous truth. *Nature* 575 (2019), 403 – 406. https://api.semanticscholar.org/CorpusID:207986664

[248] Xueqing Wu, Jiacheng Zhang, and Hang Li. 2022. Text-to-Table: A New Way of Information Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 2518–2533. https://doi.org/10.18653/v1/2022.acl-long.180

[249] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. arXiv:2309.07597 [cs.CL]

[250] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808* (2020).

[251] Fangyuan Xu, Junyi Jessy Li, and Eunsol Choi. 2022. How Do We Answer Complex Questions: Discourse Structure of Long-form Answers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 3556–3572. https://doi.org/10.18653/v1/2022.acl-long.249

[252] Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. *arXiv preprint arXiv:2305.18201* (2023).

[253] Yumo Xu and Mirella Lapata. 2020. Coarse-to-Fine Query Focused Multi-Document Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 3632–3645. https://doi.org/10.18653/v1/2020.emnlp-main.296

[254] Ying Xu, Valery Vigil, Andres S. Bustamante, and Mark Warschauer. 2022. "Elinor's Talking to Me!":Integrating Conversational AI into Children's Narrative Science Programming. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 166, 16 pages. https://doi.org/10.1145/3491102.3502050

[255] Ying Xu and Mark Warschauer. 2020. A Content Analysis of Voice-Based Apps on the Market for Early Literacy Development. In *Proceedings of the Interaction Design and Children Conference* (London, United Kingdom) *(IDC '20)*. Association for Computing Machinery, New York, NY, USA, 361–371. https://doi.org/10.1145/3392063.3394418

[256] Ying Xu and Mark Warschauer. 2020. Exploring Young Children's Engagement in Joint Reading with a Conversational Agent. In *Proceedings of the Interaction Design and Children Conference* (London, United Kingdom) *(IDC '20)*. Association for Computing Machinery, New York, NY, USA, 216–228. https://doi.org/10.1145/3392063.3394417

[257] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. Harnessing Biomedical Literature to Calibrate Clinicians' Trust in AI Decision Support Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 14, 14 pages. https://doi.org/10.1145/3544548.3581393

[258] Xianjun Yang, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Xiaoman Pan, Linda Petzold, and Dong Yu. 2023. OASum: Large-Scale Open Domain Aspect-based Summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4381–4401. https://doi.org/10.18653/v1/2023.findings-acl.268

[259] Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. RRHF: Rank Responses to Align Language Models with Human Feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=EdIGMCHk4l

[260] Shuo Zhang and Krisztian Balog. 2018. On-the-fly Table Generation. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (2018). https://api.semanticscholar.org/CorpusID:44090123

[261] Siyan Zhao, John Dang, and Aditya Grover. 2024. Group Preference Optimization: Few-Shot Alignment of Large Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=DpFeMH4l8Q

[262] Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. 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, 5905–5921. https://doi.org/10.18653/v1/2021.naacl-main.472

Acknowledgment

I would not have made it this far without the unwavering support of so many people who have been part of this journey.

- **My advisor, Juho Kim**, for helping me become not only a better researcher but also someone who dares to dream bigger and follow my heart. Thank you for trusting me and guiding me through every step of this journey.
- **My incredible dissertation committee—Eytan Adar, Alice Oh, Sherry Wu, and Minjoon Seo**—thank you for your thoughtful feedback, encouragement, and for being there not only for my dissertation but also for my growth and future.
- **My internship mentors—Pao Siangliulue, Joseph Chee Chang, Jonathan Bragg, Kyle Lo, Daniel S. Weld, Alexa F Siu, Tong Sun, Kyungjae Lee, and Moontae Lee**—thank you for helping me grow into a better collaborator and a more independent researcher.
- **My internship friends**, thank you for making even the short three-month stints feel memorable and full of joy.
- **All KIXLAB members**, for creating a community where I could grow both intellectually and personally. **Minsuk, John, and Jean**, thank you for your mentorship and thoughtful advice. **Taesoo**, you've been the best collaborator and lunchmate—I'm grateful for all the exciting ideas we explored together. **Kihoon and Yoonsu**, our Pokémon breaks truly kept me (in)sane. **Hyunwoo, Yoonseo, and Seoyoung**, you helped me settle in and feel at home when I first joined the lab. **Saelyne**, you'll always be my best conference roommate, and **DaEun**, thank you for being a reliable co-lab lead. **Hyungyu and Eunyoung**, I'm glad we could share this graduation milestone together. And to all the past and current members of KIXLAB—thank you for the memories, the laughter, and your constant support.
- **My friends Junseon, Sunkyoung, Seah, Seunghyun, Soojung**, thank you for pulling me out of the lab to breathe, laugh, and drink. Our conversations were my refuge during the most hectic days of PhD life.
- **My parents Jaeho and Sookjin**, for their unconditional love and support. Your quiet strength and belief in me have been my foundation. **My sister and best friend Seoyoung**. I'll miss countless late-night talks—about life, research, careers, dreams, and even tomorrow's lunch.