Chat as Learning in Interactive Information Retrieval and Generation

Ran Yu
GESIS – Leibniz Institute for the Social Sciences
Germany
ran.yu@gesis.org

Jiqun Liu
The University of Oklahoma
OK, USA
jiqunliu@ou.edu

Abstract

Generative AI-based conversational systems are rapidly becoming the default gateway to information, yet their educational potential and risks remain under-examined. This paper introduces *Chat as Learning* (CaL), a paradigm that views dialogue not as a veneer over search but as the central mechanism for knowledge construction. We present a three-layer framework that couples fine-grained learner modeling with adaptive prompting support and a retrieval-augmented generator that blends the model's prior knowledge with verified, up-to-date sources. Each conversational turn functions as a diagnostic probe, which enables real-time refinement of intent, mastery level, and motivational state. Building on this architecture, we map a cross-disciplinary research agenda that spans user models, reinforcement-optimized prompt policies, information-quality and knowledge gap signaling, multimodal assessment instruments, and bias mitigation. By foregrounding both the promise of democratized expertise and the threat of over-reliance on fluent yet fallible outputs, CaL establishes an urgent agenda for reproducible benchmarks and responsible deployment of conversational learning systems.

1 Introduction

There has been a significant uptrend in the use of chatbots built upon generative AI (GenAI), particularly for information seeking aimed at acquiring new knowledge. Chatbots have proven to be efficient tools, as they provide direct answers and concise summaries. The conversational format helps users better articulate their information needs and avoids the overwhelming amount of results typically encountered when using search engines. However, several challenges remain in effectively supporting learning through chatbots. These include concerns about the quality of information (e.g. hallucinations), the possible hindrance of critical self-thinking, the disparities in users' ability to engage in effective conversations, the effort and time required to craft the right prompts, and the reduced transparency regarding information sources [Divekar et al., 2024].

In this paper, we focus our discussion on *informal learning* that occurs through interactions with GenAI-based chatbots. Specifically, how users seek and construct knowledge in self-directed, everyday contexts outside formal education, such as when exploring unfamiliar topics, solving real-life problems, or satisfying curiosity. Our aim is to understand the opportunities and limitations of GenAI chatbots as tools for supporting informal, self-directed learning experiences.

A range of challenges must be addressed to advance CaL across user, system, and data perspectives. From the user's perspective, it remains difficult to characterize users' learning goals and tasks during chat-based learning activities, to profile the users individual learning styles, and to assess their learning progress. From the system perspective, open questions include how to build or optimize chat systems to better support users' learning goals, whether new functions are needed specifically for learning purposes, and how to evaluate system performance beyond traditional metrics. On the data side, curating appropriate datasets and developing benchmarks to support replicable and scalable evaluation is essential. Important issues include identifying the data and signals needed to predict and assess learning progress, how these differ from those in search-as-learning (SaL), and which quality measures (e.g., conciseness, trustworthiness, bias, difficulty) should be employed to assess generated content. More broadly, there is a need to consider what empirical and methodological insights from SaL research can inform CaL, and how CaL may, in turn, challenge existing views of informal learning and online education.

Our vision is to advance the design of GenAI-based conversational tools to better support informal learning, aligned with evolving user preferences and needs. We aim to build systems that reflect current trends in learning behavior, such as the preference for concise snippets over comprehensive resources and the growing favor of multimedia formats like videos over text. Moreover, we aim to investigate into conversational systems that can adapt to users' individual learning styles, helping learners engage more effectively according to their individual needs. Our ultimate objectives are to foster higher knowledge gain and improve learning efficiency through conversational AI tools. In the long term, we see CaL as a means to reduce barriers to education, making learning more accessible and inclusive, particularly for individuals with limited resources, diverse learning backgrounds, or specific challenges. By shaping conversational tools that act as personalized and supportive learning companions, our aim is to contribute to a more equitable and empowered learning society.

2 Foundation and Research Progresses

2.1 Search as Learning

SaL has emerged as a distinct research field at the intersection of information seeking and retrieval, human-computer interaction (HCI), and learning sciences [Rieh et al., 2016]. It investigates how users engage in web search activities not just to locate information, but also to acquire new knowledge and develop understanding [Marchionini, 2006; Vakkari, 2016]. Key areas of study in SaL include user modeling, search behavior analysis, learning assessment, system support for learning, and task design [Von Hoyer et al., 2022; Urgo et al., 2025; Vakkari, 2016].

To characterize learning stages and types, researchers have adapted learning taxonomies such as Bloom's taxonomy [Bloom et al., 1956] and Anderson and Krathwohl's taxonomy [Anderson et al., 2000]. Task stages [Liu, 2019; Liu et al., 2020b], task types, and learning types have also been investigated, e.g., declarative learning [Gadiraju et al., 2018] versus procedural learning [Otto et al., 2022]. To better understand the learning process, studies have been conducted to collect data, which typically involve tracking user interactions and assessing learning outcomes. Knowledge assessment methods include quizzes (e.g. [Gadiraju et al., 2018]), self-reports (e.g. [Capra et al., 2018; Collins-Thompson et al., 2016; Ghosh et al., 2018]), essay writing (e.g. [Otto et al.,

2022]), and implicit measures such as query complexity [Chi et al., 2016]. To analyze interactions between users, systems, and resources, features from all three components are extracted. From the user perspective, behavioral features from interaction logs [Gadiraju et al., 2018] and eye gaze patterns [Bhattacharya and Gwizdka, 2019] are analyzed. Additionally, task complexity [Liu et al., 2019] and cognitive abilities [Kalyani and Gadiraju, 2019] are characterized. From the system perspective, features of SERP pages, ranking functions, and interface elements are examined [Yu et al., 2018]. From the resource perspective, characteristics such as readability, complexity, HTML structure, and multimedia content are analyzed [Yu et al., 2021; Otto et al., 2021; Eickhoff et al., 2014]. Weak to moderate correlations have been found between individual features and users' learning gains. Building on this extended understanding and the variety of extracted features, researchers have also attempted to build models for predicting users' learning outcome such as knowledge gain [Yu et al., 2018; Bhattacharya and Gwizdka, 2019; Yu et al., 2021; Otto et al., 2021]. To support learning, both direct optimizations of search systems and additional tools have been developed. Direct optimizations include re-ranking algorithms [Syed and Collins-Thompson, 2017; Rokicki et al., 2022] and different information visualization techniques [Kammerer et al., 2009; Câmara et al., 2021; Salimzadeh et al., 2021]. In terms of additional tools, note-taking, goal-setting, and self-assessment functionalities have been explored and assessed in user experiments [Roy et al., 2021; Urgo and Arguello, 2023; Syed et al., 2020].

The *International Workshop on Investigating Learning during Web Search* (IWILDS) series provides a vibrant platform to share research progresses, explore open questions, and discuss empirical methods in SaL. The workshop series consistently attracts an international and interdisciplinary audience, with participation from researchers and practitioners in computer science, information science, library studies, learning and educational science, and psychology. This diversity has enriched discussions and encouraged the examination of SaL research across different cultural and educational contexts. The IWILDS workshops not only serve as a research outlet, but also enable multidisciplinary collaboration and idea formulation by bringing together researchers and fostering interactive discussions. The workshop has been held for five years in conjunction with major IR conferences, including SIGIR, CHIIR, CIKM, and WSDM [Hoppe et al., 2020, 2021, 2022, 2025]. Interdisciplinary research topics presented at the workshop including search algorithms [Rokicki et al., 2022], search interfaces [Tolmachova et al., 2020; Qiu et al., 2020], learning assessment methods [Liu et al., 2020a; Tang et al., 2021; Gritz et al., 2021], self-regulated learning [Urgo and Arguello, 2022a], and SaL in educational environments [Landoni et al., 2022]. In the 2025 workshop held at WSDM, aligned with the conference theme and recent research trends in IR, the workshop organizers expanded the scope of IWILDS to acknowledge the growing trend of shifting learning needs from web search toward conversational systems and AI agents, and facilitated research challenges on learning in chat systems through an interactive discussion session. By responding to emerging technological shifts, such as the rise of conversational interfaces, retrieval-augmented generative responses, and agents for learning, IWILDS serves as an important platform and offers a collaboration basis for future CaL research and system evaluation.

Urgo et al. [2025] provided a comprehensive overview of the SaL research. Many findings from SaL research are directly relevant for CaL. Similar to SaL, CaL involves user-driven inquiry and requires systems to support knowledge acquisition. SaL's insights into learning intent, user characteristics (e.g., prior knowledge), and task design can inform the development of chat systems aimed at supporting learning. Similarly, SaL's evaluation methods, such as task-based user studies

and behavioral modeling, can be adapted to study learning outcomes in chat interactions. However, CaL differs in that it involves a dynamic, dialogic interaction between user and the system, rather than a static search result interface. This raises new questions about how users and chatbots interact, and how the dialog affects the learning process and experience. Unlike SaL, where users largely control the search trajectory, CaL allows systems to steer users through questions, prompts, or rephrasings, potentially affecting learning depth and direction. Issues such as trust in AI explanations, the role of uncertainty, and the learner's ability to reflect critically during the dialogue demand further study. SaL provides useful groundwork, but cannot fully address the complexity of dialogic learning without methodological adaptation and conceptual extension.

2.2 Informal Learning and AI-enabled Chat Systems

Recent research investigates how GenAI-based conversational systems, particularly ChatGPT, are reshaping information-seeking and informal learning behaviors [Lo, 2023; Zhu et al., 2023; Li et al., 2025], and also examines the potential risks, e.g. biases and unfairness, in LLM-enabled chat interactions, e.g. [Dai et al., 2024]. Divekar et al. [2024] explore how higher education students perceive learning through GenAI-based conversational systems like ChatGPT compared to traditional search engines. Using a within-subjects study design, they find that students appreciate ChatGPT's conversational interface for its ease of use, scaffolded information, and synthesis capabilities, but express concerns about hallucinated information, lack of source transparency, and challenges in crafting effective prompts. Xu et al. [2023] conduct an online experiment with a between-subjects design comparing ChatGPT and Google Search for information seeking tasks. They find that users complete tasks more quickly with ChatGPT and perceive its responses as higher in information quality and user satisfaction. However, ChatGPT struggles with fact-checking tasks and may induce overreliance or propagate misinformation in their study. In particular, ChatGPT tends to level search performance across different educational backgrounds, suggesting a potential for democratization. Karunaratne and Adesina [2023] survey students in higher education to investigate the impact of ChatGPT on their information seeking behavior. Their findings indicate widespread adoption of ChatGPT for information retrieval due to its perceived efficiency and ease of use. However, a minority of students express hesitation related to trustworthiness, relevance, and technical access issues. Opara et al. [2023] explores the opportunities and challenges associated with using ChatGPT in educational settings. The authors emphasize ChatGPT's advantages, including its ability to provide quick responses to queries, support research through natural language processing and text generation, and facilitate self-paced learning. However, they also note several key limitations, such as trustworthiness, inconsistent responses to similar questions, the potential to hinder creativity, and concerns over plagiarism. Hidayat-ur Rehman [2024] conduct a study (n=387) to examine how AI competence, chatbot usage, and perceived autonomy influence student engagement in digital informal learning. Using structural equation modeling, the study finds that chatbot use boosts engagement, and that AI competence is a key driver for both chatbot adoption and effective informal learning. Kist and Moshfeghi [2025] investigate the design of chatbots as informal learning tools for museum visitors, emphasizing information accuracy and user data privacy. Further research on the adoption of conversational systems in formal education settings (e.g., Chakraborty [2024]; Milana et al. [2024]; Gupta and Chen [2022]; Liu et al. [2024]),

including scenarios such as language learning [Wu et al., 2023] and collaborative learning [Neto and Fernandes, 2019], also offers valuable insights applicable to CaL in informal contexts.

With the rapid advancement of GenAI techniques, many issues, such as the recency of information and the lack of citations in provided knowledge, have been addressed in the latest versions of various GenAI-based conversational tools. However, challenges like understanding user intent and learning goals, as well as tailoring learning resources to these factors, remain unresolved.

3 Conceptual Framework

Build upon the research progresses on SaL, conversational agents, and GenAI, we propose a CaL framework to illustrate a novel human-AI interaction modality in supporting informal learning, and also introduce relevant research opportunities and open challenges.

3.1 CaL Framework

The *CaL framework* (illustrated in Figure 1) comprises three interconnected components, including Learner, Chat Interface, and Backend Model, each linked to both static and dynamic knowledge sources to support an adaptive, dialogic learning process. The Learner enters the system with a clearly articulated learning task (for example, mastering Bayesian inference for use in classification tasks) or ill-defined learning goal that can be clarified or defined through user-system negotiation, and an existing knowledge base comprising prior notes, solved exercises, and mental schemas. The Chat Interface serves as the conversational front end: it captures the learner's natural-language prompts and tasks, displays model-generated explanations, and offers targeted prompt suggestions, such as reframing questions or proposing follow-up inquiries, that guide reflection and deepen engagement. Behind the scenes, the Backend Model generates responses by synthesizing internalized world knowledge with additional knowledge and retrieved objects, such as recent research articles, code repositories, data visualizations, or multimedia resources, ensuring that explanations remain both conceptually coherent and anchored in up-to-date relevant material. In tandem, these elements form a continuous loop in which prompt formulation, response generation, scaffolded suggestion and prompt reformulation, and learner reflection iterate until the original learning goal is achieved or redefined in the interactive process.

A hypothetical scenario illustrates the CaL cycle. A data science student might begin by asking, "How is Bayes' theorem derived for a Gaussian likelihood with a conjugate prior?" The LLM produces a step-by-step mathematical derivation and, via the interface, suggests exploring a coding example: "Would you like to see how to implement this derivation in NumPy?" The learner integrates the theoretical explanation into their mental model, recognizes the need for practical application, and issues a refined prompt: "Show me how to write Python code that samples from this posterior using PyMC3." In response, the system retrieves a relevant GitHub notebook, annotates key lines with explanatory comments (for instance, defining the prior distribution and sampling procedure), and proposes a further scaffold: "Consider varying the prior variance and observing its impact on the posterior distribution." Through each conversational turn, the learner's knowledge base is updated (both mentally and in any external note-taking system), the learning goal shifts (from derivation to implementation to experimentation), and the system adapts its scaffolding to support conceptual clarity, practical coding skills, and experimental exploration.

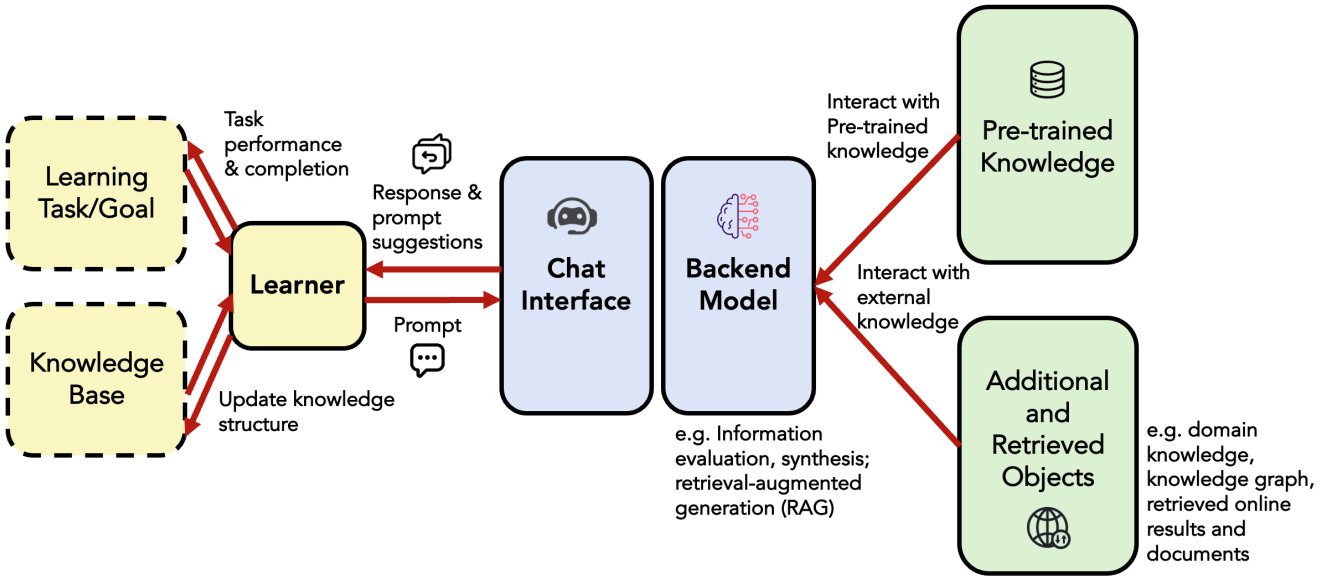

Figure 1. Chat as Learning (CaL) Conceptual Framework.

Although CaL shares affinities with SaL, where users iteratively reformulate queries motivated by multifaceted learning tasks, examine ranked search results, evaluate document relevance and usefulness, and gradually build understanding [Rieh et al., 2016; Urgo and Arguello, 2022b; Liu, 2021], CaL differs in its integration of generative scaffolding and real-time content synthesis. Traditional search engines return static lists of documents, leaving learners to navigate disjointed sources, infer connections, evaluate relevance, and manage cognitive load independently. In contrast, CaL weaves generative explanations, dynamic prompt suggestions, and curated retrieval into a unified conversational response. This approach not only reduces the overhead of source evaluation but also makes the scaffolding process explicit, providing meta-cognitive support that helps learners recognize gaps in understanding, anticipate next steps in CaL, and reflect on their own reasoning. By transforming information seeking into an adaptive dialogue, one that continuously diagnoses learner needs, delivers *just-in-time* guidance, and fosters both deep comprehension and hands-on skill development, CaL advances beyond SaL into a new paradigm of conversationally facilitated self-directed learning experience.

3.2 Open Challenges

Within the CaL framework introduced above, the first set of challenges arises in learner modeling and goal inference. Unlike traditional tutoring systems that rely on pre-defined competency models, CaL must infer a learner's evolving objectives, background knowledge, and cognitive state and loads solely from conversational exchanges. How can we design dialogue analytics that detect shifts in a learner's understanding—whether through hesitations, question reformulations, or vocabulary changes—and adapt the system's scaffolding accordingly? What features of a dialogue turn most reliably signal confusion versus deep engagement? For example, a sudden increase in follow-up questions about prior steps may indicate that the learner has reached a conceptual bottleneck, whereas an immediate request for advanced examples could signal readiness for chal-

lenge. Developing robust probabilistic or neural user models that map utterance patterns to latent knowledge states remains an open research frontier, and doing so will be critical to ensuring that CaL can provide personalized support at the granularity of individual learners.

A second dimension of inquiry concerns the design of adaptive scaffolding and prompt suggestion strategies. Although prompt suggestions have proven effective in preliminary studies, it is not yet clear which types of scaffolds, such as Socratic questioning, scaffolding hints, worked examples, or reflective prompts [Chang, 2023; Kumar et al., 2024], yield the greatest gains for different learning goals and content domains. How should the system balance between offering too much guidance, which risks diminishing learner autonomy, and too little, which may lead to frustration? Moreover, can we devise algorithms that dynamically adjust the frequency and complexity of prompt suggestions based on online assessments of cognitive load, perhaps measured through response latency or self-report? Exploring reinforcement learning approaches, where the system receives feedback on which scaffolds lead to measurable learning improvements, represents a promising but under-explored avenue.

The integration of static pre-trained knowledge and dynamic internet retrieval presents a third set of challenges. Pre-trained LLMs excel at generating coherent narratives but may reflect outdated information or undesired biases, while live retrieval of external documents introduces concerns about source credibility, recency, and consistency. How can CaL architectures seamlessly reconcile these two knowledge streams to produce explanations that are both accurate and trustworthy? Techniques such as query rewriting, citation augmentation, and provenance tracking warrant further investigation to ensure transparency and user trust. In addition, system designers must determine when to prioritize generative explanations versus direct retrieval—should the model summarize a newly published research paper or link the learner directly to the source for deeper exploration? Answering these questions will require novel retrieval-generation hybrid models and evaluation methodologies that extend beyond standard information retrieval metrics.

Closely tied to these technical challenges is the need for rigorous evaluation frameworks that capture learning outcomes rather than mere dialogue quality. Traditional IR measures such as precision, recall, and nDCG do not fully account for knowledge acquisition, conceptual understanding, or the ability to transfer skills to novel contexts. We must therefore develop new assessment instruments, such as pre- and post-conversation concept inventories, task-based performance metrics, and long-term retention tests, that can be embedded within CaL interactions. Furthermore, longitudinal studies are necessary to examine whether gains observed in a single session persist over time and across different topics and application contexts, and translate into genuine competence. Building publicly available benchmark datasets that simulate authentic learning scenarios, complete with human-annotated proficiency labels, will be instrumental in driving empirical progress and replicable evaluations on LLM learning systems.

Finally, CaL must address questions of equity, accessibility, and user diversity [Olteanu et al., 2021]. Learners vary widely in language proficiency, digital literacy, cultural background, and neurocognitive profiles; a one-size-fits-all conversational agent risks exacerbating existing educational disparities. What interface adaptations or content personalization strategies are required to support non-native speakers, individuals with learning differences, or those with limited prior domain knowledge? Moreover, as CaL systems become more widely deployed, how can we ensure that they do not inadvertently reinforce biases or exclude underrepresented perspectives? Investigating collaborative models where peers or instructors can intervene in the conversational loop,

and embedding mechanisms for learner feedback and error correction will be crucial to building trustworthy CaL tools and platforms. Addressing these open challenges will lay the groundwork for transforming chat-based interactions into scalable, equitable pathways for lifelong learning.

4 CaL Research Agenda

Building on the conceptual framework and components outlined above, we now turn to the open research challenges that define the future agenda for CaL research and applications.

4.1 Learner Modeling and Goal Inference

Understanding and supporting learners' goals and actions is the foundation of CaL systems. Yet building a conversational agent that truly advances learning is far more demanding than simply delivering correct answers. The agent must infer what the learner is trying to accomplish, recognize how that intention evolves over time, and track incremental changes in knowledge, confidence, and motivation and interest development. This inference process often occurs from fragmentary dialogue and subtle behavioral cues. Without such rich, moment-to-moment insight, even a technically capable chatbot risks offering generic help, mis-timed interventions, or feedback that ignores a learner's preferred level of autonomy. Recent work in learning analytics, intelligent tutoring, and human–AI interaction highlights three intertwined challenges [Wang et al., 2023; Yan et al., 2024]. Learners' utterances rarely reveal their full intent: a single question may signal curiosity or emerging interest, confusion, or an attempt to test mastery. In addition, knowledge is dynamic; misconceptions can fade or resurface within minutes, and progress is uneven across topics. Besides, learners differ markedly in how much guidance they want, how they engage with feedback, and how they balance effort against convenience. Addressing these challenges requires a principled framework for modeling intentions, estimating real-time knowledge states, and adapting support to diverse motivational stances.

Framing the problem in this way foregrounds the critical research questions below. Each question targets a gap that must be bridged before CaL systems can move from reactive answer engines to proactive learning partners capable of sustaining growth over extended interactions:

- How to characterize users' intents and tasks in CaL activities?
- How to model users' real-time knowledge status and evaluate their learning progress over time (i.e. changes in knowledge state)? Real-time understanding and adaptation.
- How to account for learning styles and user expectations (e.g. desired level of effort)? How do users expect a chatbot to do to support learning? How should systems balance autonomy versus guidance in informal learning activities?

4.2 Adaptive Scaffolding and Prompt Suggestion in CaL

Knowing what a learner understands is only the starting point; a CaL agent must also decide how to intervene or whether to intervene at all at each conversational turn. We treat this choice as a sequential decision problem: given a state that combines mastery estimates, moment-to-moment cognitive-load signals (e.g., typing latency, pupil dilation, self-reported effort), and the learner's

expressed role preferences, the agent selects an instructional action defined by three interlocking parameters. A CaL system can start with adopting a persona (teacher, peer, reference librarian, subject-matter expert). Then it may chooses a scaffold genre (worked example, Socratic probe, reflective paraphrase, micro-hint). After that, the system can set a delivery cadence (granularity, modality, frequency). Because short-term comprehension, long-term retention, motivation, and workload rarely align, any effective policy must optimize multiple objectives and remain sensitive to rapid changes in the learner's readiness. Novices may benefit from densely annotated code snippets delivered every few exchanges, whereas advanced learners profit from occasional meta-cognitive nudges that invite them to generate analogies or critique solutions. Each prompt therefore carries an explicit "why this prompt?" rationale that users can inspect or override, reinforcing transparency and shared agency. Framing adaptive scaffolding in this way foregrounds a set of research questions that need to be addressed before AI agents can move beyond reactive Q&A toward genuine learning partners:

- LLMs can adopt various roles: teacher, tutor, peer, or information retrieval assistant. The most effective role may depend on the learning objective and user characteristics. How can a CaL system adapt towards users' needs?
- How does learner performance on distinct cognitive goals—conceptual understanding, procedural fluency, metacognitive awareness, and near-/far-transfer—differ when instruction is supported by Socratic questioning, worked examples, reflective prompts, or hint-based scaffolding within a conversational tutoring setting?
- To what extent do learner characteristics (e.g., prior knowledge and belief, self-regulation skill) and task complexity moderate the relative effectiveness of each scaffolding strategy, thereby informing an adaptive policy for selecting the most suitable scaffold type?
- How accurately do real-time cognitive-load indicators, such as response latency, pupil dilation, and self-reported effort, predict the optimal frequency and complexity of instructional prompts in a conversational learning environment?
- Can a reinforcement-learning policy that continuously adjusts prompt timing and difficulty based on these cognitive-load signals outperform static or heuristic schedules in enhancing short-term learning gains and long-term knowledge retention?
- Under which content characteristics (e.g., length, readability, topical overlap) and learner states (e.g., prior knowledge, trust calibration) does a conversational learning agent improve comprehension, retention, and source evaluation by supplying an AI-generated summary rather than linking directly to the original document, and how do query-rewriting, citation-augmentation, and transparency affordances modulate these effects?

4.3 Knowledge Integration

High-quality conversational explanations emerge from a delicate coordination between two knowledge sources: the extensive but static world knowledge embedded in an LLM's parameters and the ever-changing stream of external information that can be retrieved on demand and ranked according to queries and expressed information needs. Achieving that coordination requires mechanisms that can decide whether to rely on the model's internal memory, to inject freshly retrieved passages, or to blend the two. Each decision must balance multiple, often conflicting objectives:

factual accuracy and recency, narrative coherence and fluency, viewpoint diversity, cognitive load on the learner, and transparency about provenance and potential bias. Interface designs that surface lightweight "quality badges" (e.g., confidence, reading difficulty, bias risk) promise to give learners agency in deciding when to drill down into primary sources, when to compare perspectives, and when to flag contradictions. Designing and evaluating such pipelines thus foregrounds a new set of research questions as follows at the intersection of retrieval-augmented generation (RAG) [Lewis et al., 2020], explanation quality, and learner trust calibration in chat sessions:

- Under what topic conditions and time-sensitivity requirements does augmenting a pre-trained language model with on-the-fly retrieval improve factual accuracy and recency without sacrificing the narrative coherence that learners value in conversational explanations?
- How can token- or segment-level gating strategies dynamically allocate content between an LLM's static memory and retrieved passages to optimize the trade-off between fluency and factuality across diverse knowledge domains in CaL interactions?
- Which automatically computed information-quality signals, e.g. conciseness, trustworthiness, bias likelihood, reading difficulty, and source provenance, most strongly predict human judgments of explanation quality and correlate with measurable learning gains?
- How does explicitly surfacing quality signals within the dialogue interface influence learners' trust calibration, source-evaluation behavior, and long-term retention when compared with explanations that embed the same signals invisibly in generation-time scoring?

4.4 Evaluation

Robust evaluation of CaL systems requires moving far beyond traditional relevance-focused IR metrics to capture whether interactions truly reshape conceptual understanding, foster transfer, and sustain self-regulated learning progression over time. Yet the field still lacks empirically validated instruments that can detect short-term knowledge gains, discriminate among diverse learner trajectories, and trace consolidation months after a dialogue ends. As discussed in Section 2.1, knowledge assessment methods from SaL research offer a useful starting point; however, in CaL settings, the dialogic nature of interaction may amplify the influence of pre-assessment instruments (e.g., questionnaires or quizzes) on participant behavior. Comparative evidence is also sparse on how explicit concept inventories stack up against task-embedded performance measures, or on which behavioral and discourse cues, such as dialogue-act sequences, scaffold uptake, or source-request patterns, most reliably signal learning in naturally flowing conversations. Compounding these gaps, existing evaluation benchmarks rarely pair fine-grained interaction logs with human-labeled knowledge states [Markwald et al., 2023], limiting our ability to conduct fair, cross-system evaluations or to explore how micro-level behaviors predict long-term retention. Addressing these foundational deficiencies is essential for establishing psychometrically sound, reproducible, and generalizable assessment frameworks, which in turn motivates the research questions below.

- How reliable and sensitive are pre- and post-conversation concept inventories in detecting short-term knowledge gains across diverse subject domains when compared with task-based performance assessments in a CaL setting?
- How can user knowledge be assessed in evaluation studies without interfering with natural behavior and engagement during conversational interactions?

- To what extent do task-embedded performance measures (e.g., problem-solving accuracy, solution-path efficiency) predict both immediate comprehension and long-term retention, and how do they complement or supersede concept inventories in evaluating CaL systems?
- Which behavioural and discourse signals are uniquely diagnostic of learning progression in CaL (e.g., dialogue-act sequences, scaffold uptake, source-request patterns), and how do these differ from the signals traditionally captured in SaL benchmarks?
- What design principles and annotation protocols are required to curate replicable CaL datasets that pair rich interaction logs with human-labeled knowledge states and progression indicators, thereby enabling fair, cross-system comparisons of conversational learning efficacy over extended time horizons?

4.5 Overarching Questions

Conversational agents matter only if they reshape real-life learning ecosystems and promote fairness and accessibility in learning resources distribution and use. Researchers can stage longitudinal trials that pit chat agents against web search and static media in libraries, adult education classes, and hobbyist forums, tracking far transfer, self-regulation, and citation practices over time. Expectation surveys can be adopted to measure how persona switching, quality badges, and override controls mold trust, help seeking, and metacognitive planning in AI-assisted learning. Ethnographic studies and field observation can map how chat agents redistribute expertise and shift citation practices within peer networks. Logging tools and metrics from SaL work, including query-intent taxonomies and session markers, can anchor precise cross-system comparisons. As GenAI models evolve rapidly, research on CaL must also adapt, e.g. developing evaluation methods, benchmarks, and datasets that are reproducible, generalisable, and responsive to changing system capabilities. Sustaining meaningful insights requires approaches that generalize across models and remain robust over time. These efforts lead directly to the overarching research questions that follow.

- Are generative AI-based chat systems more effective for informal learning than traditional media or Web search?
- What expectations do users have of chat systems in learning contexts?
- How should systems be optimized to support informal, self-regulated learning?
- What insights from SaL research can inform CaL design across different learning stages, scenarios, and communities?
- How could CaL systems shape educational practices, especially in informal settings?
- How can research on CaL remain reproducible and sustainable amid the rapid evolution of GenAI models, ensuring that insights, evaluations, and datasets remain relevant and adaptable to new developments?

5 Broader Impacts and Applications

Conversational agents built on the CaL paradigm promise to democratize expertise by delivering adaptive, dialogue-driven explanations that meet learners at their current level of knowl-

edge and linguistic proficiency. By fusing rich learner and interest development modeling with provenance-aware retrieval and adjustable scaffolding, CaL systems can lower the thresholds that have historically limited self-directed study and online information evaluation, and allow patrons in small public libraries to explore quantum mechanics, children to identify and assess algorithms in maker-based learning activities [Jung et al., 2024], adult learners to reskill for data-centric jobs, and patients to convert clinical jargon into personally meaningful guidance [Liu and Jung, 2021; Tate and Warschauer, 2022]. Because the same core architecture supports multilingual paraphrases, readability controls, and modality shifts, these benefits extend to non-native speakers, neurodiverse users, and people with low literacy and limited technical support, potentially narrowing long-standing equity gaps in access to high-quality online learning resources.

The very features that make CaL attractive, however, introduce serious pedagogical risks. A persuasive agent that always supplies neat summaries may erode epistemic vigilance and diminish the productive struggle that fosters critical thinking, encouraging users to accept coherent prose at face value and defer judgment to the system. If reinforcement signals reward only short-term correctness, adaptive policies could converge on spoon-feeding explanations that inflate immediate quiz scores while undermining long-term retention and knowledge transfer [Gajos and Mamykina, 2022]. Information quality failures present a parallel threat: hallucinated facts, customized misinformation, outdated sources, and subtle biases can propagate quickly when explanations lack transparent provenance or surface only mainstream viewpoints, thereby entrenching misinformation and marginalizing less-represented perspectives. Left unchecked, these dynamics could widen achievement gaps for learners with weak media-literacy skills, compromise academic integrity and critical thinking, and erode trust in assessment mechanisms that rely on independent work.

Responsible and fair deployment therefore requires safeguards at multiple layers [Liu, 2023; Baum et al., 2022]. Interfaces should default to bias- and uncertainty-aware responses that display confidence ranges, alternative perspectives, and direct links to primary evidence, re-introducing constructive friction so users practise source evaluation rather than passive consumption. Reward functions will need to incorporate measures of effortful engagement, such as self-explanation and successful transfer challenges, along with accuracy, discouraging over-guidance. Personalization of the dialougue should consider the risk such as bias and ignorant in users' knowledge structure might caused by over-personlization towards user satisfaction. Continuous bias audits and stereotype stress tests should monitor both retrieved and generated content, with rapid rollback channels when harms surface. Finally, transparent data-governance policies are essential: interaction logs used for adaptation should be de-identified and inspectable by learners themselves. If these governance mechanisms evolve in concert with the technical advances outlined earlier, CaL can become a pillar of equitable, lifelong learning; without them, it risks becoming another opaque recommendation engine that quietly weakens the very cognitive autonomy it aims to enhance.

6 Conclusion

CaL reframes conversational interaction as the primary medium through which knowledge is elicited, negotiated, and consolidated, rather than as a mere interface overlay on search, recommendation, or tutoring functions. Building on insights from SaL, learning sciences, and conversational IR, this position paper proposed an integrated framework that couples fine-grained learner modeling with adaptive scaffolding, provenance-aware knowledge integration, and equity-centered

design. The accompanying research agenda pinpoints urgent questions-how to infer intent and knowledge mastery in real time, when to switch pedagogical roles, how to blend static and dynamic information sources, and how to embed multimodal assessment instruments—while also charting evaluation pathways that promote durable understanding over surface fluency. By operationalizing these questions through open benchmarks and transparent protocols, the IR community can unlock CaL's promise to personalize explanations, surface just-in-time prompts, and extend specialized expertise to learners historically underserved by traditional educational infrastructures.

That promise, however, is inseparable from systemic risks. Fluent yet fallible outputs can erode critical thinking and "healthy frictions" in human-AI interactions [Shah and Bender, 2024], over-guidance may blunt metacognitive skills, and biased retrieval pipelines threaten to entrench existing disparities. We therefore pair technical opportunities with governance imperatives: interfaces that expose uncertainty and alternative viewpoints, reward functions that value effortful engagement, continuous bias audits with rapid rollback, and data-handling policies that keep learners in control of their interaction traces. Realizing the CaL vision will demand sustained collaboration across IR, HCI, learning analytics, and AI ethics. By co-creating shared resources and embracing human-in-the-loop evaluation, researchers and practitioners can ensure that future conversational systems not only answer questions efficiently but also cultivate the reflective, self-directed habits of mind essential for equitable lifelong learning in an AI-mediated world.

Acknowledgment

Jiqun Liu was supported by a Faculty Investment Program (FIP) award from the University of Oklahoma Office of the Vice President for Research and Partnerships and a faculty research award from Microsoft. This work was partially funded by the DFG, German Research Foundation ("EmergentIR", 548295069).

References

Lorin W. Anderson, David R. Krathwohl, Peter W. Airasian, Kathleen A. Cruikshank, Richard E. Mayer, Paul R. Pintrich, James Raths, and Merlin C. Wittrock. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Abridged Edition*. Allyn & Bacon, 2000. ISBN 080131903X. URL http://www.amazon.com/Taxonomy-Learning-Teaching-Assessing-Educational/dp/080131903X.

Kevin Baum, Susanne Mantel, Eva Schmidt, and Timo Speith. From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology*, 35(1):12, 2022.

Nilavra Bhattacharya and Jacek Gwizdka. Measuring learning during search: Differences in interactions, eye-gaze, and semantic similarity to expert knowledge. In *Proceedings of the 2019 conference on human information interaction and retrieval*, pages 63–71, 2019.

Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, David R Krathwohl, et al. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York, 1956.

Arthur Câmara, Nirmal Roy, David Maxwell, and Claudia Hauff. Searching to learn with instructional scaffolding. In *Proceedings of the 2021 conference on human information interaction and retrieval*, pages 209–218, 2021.

Robert Capra, Jaime Arguello, Heather O'Brien, Yuan Li, and Bogeum Choi. The effects of manipulating task determinability on search behaviors and outcomes. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 445–454, 2018.

Sanjay Chakraborty. Generative AI in modern education society, 2024. URL https://arxiv.org/abs/2412.08666.

Edward Y Chang. Prompting large language models with the socratic method. In *2023 IEEE 13th annual computing and communication workshop and conference (CCWC)*, pages 0351–0360. IEEE, 2023.

Yu Chi, Shuguang Han, Daqing He, and Rui Meng. Exploring knowledge learning in collaborative information seeking process. In *CEUR workshop proceedings*, volume 1647, 2016.

Kevyn Collins-Thompson, Soo Young Rieh, Carl C Haynes, and Rohail Syed. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *Proceedings of the 2016 ACM on conference on human information interaction and retrieval*, pages 163–172, 2016.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. Bias and unfairness in information retrieval systems: New challenges in the LLM era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6437–6447, 2024.

Rahul R Divekar, Sophia Guerra, Lisette Gonzalez, and Natasha Boos. Choosing between an llm versus search for learning: A highered student perspective. *arXiv preprint arXiv:2409.13051*, 2024.

Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. Lessons from the journey: a query log analysis of within-session learning. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 223–232, 2014.

Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. Analyzing knowledge gain of users in informational search sessions on the web. In *Proceedings of the 2018 conference on human information interaction & retrieval*, pages 2–11, 2018.

Krzysztof Z Gajos and Lena Mamykina. Do people engage cognitively with ai? impact of ai assistance on incidental learning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pages 794–806, 2022.

Souvick Ghosh, Manasa Rath, and Chirag Shah. Searching as learning: Exploring search behavior and learning outcomes in learning-related tasks. In *Proceedings of the 2018 conference on human information interaction & retrieval*, pages 22–31, 2018.

Wolfgang Gritz, Anett Hoppe, and Ralph Ewerth. On the impact of features and classifiers for measuring knowledge gain during web search-a case study. In *CEUR Workshop Proceedings; 3052*, volume 3052, page 6. Aachen, Germany: RWTH Aachen, 2021.

Sambhav Gupta and Yu Chen. Supporting inclusive learning using chatbots? a chatbot-led interview study. *Journal of Information Systems Education*, 33(1):98–108, 2022.

Imdadullah Hidayat-ur Rehman. Examining ai competence, chatbot use and perceived autonomy as drivers of students' engagement in informal digital learning. *Journal of Research in Innovative Teaching & Learning*, 17(2):196–212, 2024.

Anett Hoppe, Ran Yu, Yvonne Kammerer, and Ladislao Salmerón. IWILDS'20: The 1st international workshop on investigating learning during web search. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3535–3536, 2020.

Anett Hoppe, Ran Yu, Irina Brich, and Jiqun Liu. IWILDS'21: Second international workshop on learning during web search. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4880–4881, 2021.

Anett Hoppe, Ran Yu, and Jiqun Liu. IWILDS'22–third international workshop on investigating learning during web search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3482–3484, 2022.

Anett Hoppe, Ran Yu, Jiqun Liu, and Nilavra Bhattacharya. IWILDS'25: The 5th international workshop on investigating learning during web search. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 1116–1117, 2025.

Yong Ju Jung, Jiqun Liu, Harun Karahan, and Mahdieh Nazari. Tech treasure hunt: Promoting children's learning on how algorithm works at a public library. *Proceedings of the Association for Information Science and Technology*, 61(1):962–964, 2024.

Rishita Kalyani and Ujwal Gadiraju. Understanding user search behavior across varying cognitive levels. In *Proceedings of the 30th ACM conference on hypertext and social media*, pages 123–132, 2019.

Yvonne Kammerer, Rowan Nairn, Peter Pirolli, and Ed H Chi. Signpost from the masses: learning effects in an exploratory social tag search browser. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 625–634, 2009.

Thashmee Karunaratne and Adenike Adesina. Is it the new google: Impact of chatgpt on students' information search habits. In *Proceedings of the 22nd European Conference on e-Learning ECEL*, pages 147–155, 2023.

Cassandra Kist and Yashar Moshfeghi. Designing chatbots for informal learning in museums. 2025.

Harsh Kumar, Ruiwei Xiao, Benjamin Lawson, Ilya Musabirov, Jiakai Shi, Xinyuan Wang, Huayin Luo, Joseph Jay Williams, Anna N Rafferty, John Stamper, et al. Supporting self-reflection at scale with large language models: Insights from randomized field experiments in classrooms. In *Proceedings of the eleventh ACM conference on learning@ scale*, pages 86–97, 2024.

Monica Landoni, Maria Soledad Pera, Emiliana Murgia, and Theo Huibers. Let's learn from children: Scaffolding to enable search as learning in the educational environment. *arXiv preprint arXiv:2209.02338*, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. From matching to generation: A survey on generative information retrieval. *ACM Transactions on Information Systems*, 43(3):1–62, 2025.

Chang Liu, Xiaoxuan Song, Hanrui Liu, and Nicholas J Belkin. Modeling knowledge change behaviors in learning-related tasks. In *CIKM (Workshops)*, 2020a.

Chen-Chung Liu, Chen Wei Chiu, Chia-Hui Chang, and Fang-ying Lo. Analysis of a chatbot as a dialogic reading facilitator: its influence on learning interest and learner interactions. *Educational technology research and development*, 72(4):2103–2131, 2024.

Hanrui Liu, Chang Liu, and Nicholas J Belkin. Investigation of users' knowledge change process in learning-related search tasks. *Proceedings of the Association for Information Science and Technology*, 56(1):166–175, 2019.

Jiqun Liu. Characterizing the stages of complex tasks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1445–1445, 2019.

Jiqun Liu. Deconstructing search tasks in interactive information retrieval: A systematic review of task dimensions and predictors. *Information Processing & Management*, 58(3):102522, 2021.

Jiqun Liu. Toward a two-sided fairness framework in search and recommendation. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pages 236–246, 2023.

Jiqun Liu and Yong Ju Jung. Interest development, knowledge learning, and interactive ir: toward a state-based approach to search as learning. In *Proceedings of the 2021 conference on human information interaction and retrieval*, pages 239–248, 2021.

Jiqun Liu, Shawon Sarkar, and Chirag Shah. Identifying and predicting the states of complex search tasks. In *Proceedings of the 2020 conference on human information interaction and retrieval*, pages 193–202, 2020b.

Chung Kwan Lo. What is the impact of chatgpt on education? a rapid review of the literature. *Education sciences*, 13(4):410, 2023.

Gary Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

Marco Markwald, Jiqun Liu, and Ran Yu. Constructing and meta-evaluating state-aware evaluation metrics for interactive search systems. *Information Retrieval Journal*, 26(1):10, 2023.

Marcella Milana, Ulrik Brandi, Steven Hodge, and Tetyana Hoggan-Kloubert and. Artificial intelligence (ai), conversational agents, and generative ai: implications for adult education practice and research. *International Journal of Lifelong Education*, 43(1):1–7, 2024. doi: 10.1080/02601370.2024.2310448.

Antonio Justiniano Moraes Neto and Márcia Aparecida Fernandes. Chatbot and conversational analysis to promote collaborative learning in distance education. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, volume 2161, pages 324–326. IEEE, 2019.

Alexandra Olteanu, Jean Garcia-Gathright, Maarten de Rijke, Michael D Ekstrand, Adam Roegiest, Aldo Lipani, Alex Beutel, Alexandra Olteanu, Ana Lucic, Ana-Andreea Stoica, et al. Facts-ir: fairness, accountability, confidentiality, transparency, and safety in information retrieval. In *ACM SIGIR Forum*, volume 53, pages 20–43. ACM New York, NY, USA, 2021.

Emmanuel Opara, Adalikwu Mfon-Ette Theresa, and Tolorunleke Caroline Aduke. Chatgpt for teaching, learning and research: Prospects and challenges. *Opara Emmanuel Chinonso, Adalikwu Mfon-Ette Theresa, Tolorunleke Caroline Aduke (2023). ChatGPT for Teaching, Learning and Research: Prospects and Challenges. Glob Acad J Humanit Soc Sci*, 5, 2023.

Christian Otto, Ran Yu, Georg Pardi, Johannes von Hoyer, Markus Rokicki, Anett Hoppe, Peter Holtz, Yvonne Kammerer, Stefan Dietze, and Ralph Ewerth. Predicting knowledge gain during web search based on multimedia resource consumption. In *International Conference on Artificial Intelligence in Education*, pages 318–330. Springer, 2021.

Christian Otto, Markus Rokicki, Georg Pardi, Wolfgang Gritz, Daniel Hienert, Ran Yu, Johannes von Hoyer, Anett Hoppe, Stefan Dietze, Peter Holtz, et al. Sal-lightning dataset: Search and eye gaze behavior, resource interactions and knowledge gain during web search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, pages 347–352, 2022.

Sihang Qiu, Alessandro Bozzon, and Ujwal Gadiraju. Conversational interfaces for search as learning. In *CIKM (Workshops)*, 2020.

Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science*, 42(1):19–34, 2016.

Markus Rokicki, Ran Yu, and Daniel Hienert. Learning to rank for knowledge gain. In *INRA/IWILDS@ SIGIR*, pages 60–68, 2022.

Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, and Claudia Hauff. Note the highlight: incorporating active reading tools in a search as learning environment. In *Proceedings of the 2021 conference on human information interaction and retrieval*, pages 229–238, 2021.

Sara Salimzadeh, David Maxwell, and Claudia Hauff. The impact of entity cards on learning-oriented search tasks. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 63–72, 2021.

Chirag Shah and Emily M Bender. Envisioning information access systems: What makes for good tools and a healthy web? *ACM Transactions on the Web*, 18(3):1–24, 2024.

Rohail Syed and Kevyn Collins-Thompson. Retrieval algorithms optimized for human learning. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 555–564, 2017.

Rohail Syed, Kevyn Collins-Thompson, Paul N Bennett, Mengqiu Teng, Shane Williams, Dr Wendy W Tay, and Shamsi Iqbal. Improving learning outcomes with gaze tracking and automatic question generation. In *Proceedings of the web conference 2020*, pages 1693–1703, 2020.

Rui Tang, Ran Yu, Markus Rokicki, Ralph Ewerth, and Stefan Dietze. Domain-specific modeling of user knowledge in informational search sessions. In *CEUR Workshop Proceedings; 3052*, volume 3052, page 8. Aachen, Germany: RWTH Aachen, 2021.

Tamara Tate and Mark Warschauer. Equity in online learning. *Educational Psychologist*, 57(3): 192–206, 2022.

Tetiana Tolmachova, Eleni Ilkou, and Luyan Xu. Working towards the ideal search history interface. In *CIKM (Workshops)*, 2020.

Kelsey Urgo and Jaime Arguello. Capturing self-regulated learning during search. In *Workshop on Investigating Learning During Web Search*, 2022a.

Kelsey Urgo and Jaime Arguello. Learning assessments in search-as-learning: A survey of prior work and opportunities for future research. *Information Processing & Management*, 59(2): 102821, 2022b.

Kelsey Urgo and Jaime Arguello. Goal-setting in support of learning during search: An exploration of learning outcomes and searcher perceptions. *Information Processing & Management*, 60(2): 103158, 2023.

Kelsey Urgo, Jaime Arguello, et al. Search as learning. *Foundations and Trends® in Information Retrieval*, 19(4):365–556, 2025.

Pertti Vakkari. Searching as learning: A systematization based on literature. *Journal of Information Science*, 42(1):7–18, 2016.

Johannes Von Hoyer, Anett Hoppe, Yvonne Kammerer, Christian Otto, Georg Pardi, Markus Rokicki, Ran Yu, Stefan Dietze, Ralph Ewerth, and Peter Holtz. The search as learning spaceship: Toward a comprehensive model of psychological and technological facets of search as learning. *Frontiers in Psychology*, 13:827748, 2022.

Xinghua Wang, Qian Liu, Hui Pang, Seng Chee Tan, Jun Lei, Matthew P Wallace, and Linlin Li. What matters in ai-supported learning: A study of human-ai interactions in language learning using cluster analysis and epistemic network analysis. *Computers & Education*, 194:104703, 2023.

CH Wu, HY Lam, A Kong, and WLH Wong. The design and evaluation of a digital learning-based english chatbot as an online self-learning method. *International Journal of Engineering Business Management*, 15:18479790231176372, 2023. doi: 10.1177/18479790231176372.

Ruiyun Xu, Yue Feng, and Hailiang Chen. Chatgpt vs. google: A comparative study of search performance and user experience. *arXiv preprint arXiv:2307.01135*, 2023.

Lixiang Yan, Samuel Greiff, Ziwen Teuber, and Dragan Gašević. Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour*, 8(10):1839–1850, 2024.

Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. Predicting user knowledge gain in informational search sessions. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 75–84, 2018.

Ran Yu, Rui Tang, Markus Rokicki, Ujwal Gadiraju, and Stefan Dietze. Topic-independent modeling of user knowledge in informational search sessions. *Information Retrieval Journal*, 24 (3):240–268, 2021.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023.