Making Sense Together: Human-AI Collaboration in Literature Review

Neel Sanjaybhai Faganiya
University of Waterloo
Waterloo, Ontario, Canada
nsfagani@uwaterloo.ca

Abstract

Literature review is not a simple search task; it is an iterative, exploratory process through which researchers identify, interpret, and synthesize prior work. The papers selected and how they are interpreted shape the direction of your entire research. However, many AI-assisted tools treat literature review primarily as a retrieval problem, presenting ranked lists of papers and positioning users as passive consumers rather than active sensemakers. Using a research-through-design approach, we developed a conceptual interactive prototype featuring scope steering, AI-generated clustering, on-demand explanations, and explicit feedback mechanisms [1]. We conducted think-aloud sessions with participants of varying technical backgrounds, followed by a post-study Likert questionnaire assessing perceived agency, understanding, and trust. Findings indicate that agency emerges when users can visibly encode intent and observe predictable changes in system behavior. Exposed intermediate structuring, such as clustering, supported interpretive sensemaking by enabling users to reason about thematic organization. Trust was calibrated rather than absolute, depending on alignment between user actions and system responses. Differences between technical and non-technical users further highlight the need for mixed-expertise design. These results frame AI-assisted literature review as a collaborative sensemaking and suggest design implications emphasizing steerability, structural transparency, adaptive explanations, and interaction-level coherence to support calibrated trust. [2]

1 Introduction

Conducting a literature review is a cognitively demanding task that requires iterative sensemaking, comparison of perspectives across papers, and continual refinement of research questions [19]. While recent artificial intelligence (AI) tools can help with paper discovery and summarizing it, they often treat users as passive consumers of information rather than active collaborators [16]. Such designs can reduce user agency by limiting user control [2], obscure how conclusions are formed by hiding system reasoning [12], and increase the risk of over-trust in AI-generated outputs [3]. Additionally, many non-technical users rely on vague prompts and expect AI systems to infer their intent in a human-like way, revealing a mismatch between how users express goals and how current AI tools require precise prompting [23].

Despite growing interest in AI-assisted literature review, most tools emphasize improving retrieval or summarization quality rather than supporting the interactive sensemaking processes through which users develop understanding. As a result, users are often left to adapt their goals to the system through repeated prompting, rather than being supported in expressing their intent, inspecting the system's reasoning, or correcting misunderstandings. To address this gap, we investigate how a human-centered, interactive AI interface can support literature review as a sensemaking activity by shifting the focus from prompt-based interaction towards user-driven steering, feedback, and transparency. To guide this exploration, the study is structured around the following research questions:

RQ1: How do interaction design choices in AI-assisted literature review tools influence users' sense of agency and control?

RQ2: How do on-demand explanations and feedback mechanisms affect users' understanding of why papers are suggested, grouped, or ranked in a particular order?

RQ3: How do these interaction mechanisms shape users' trust and reliance on AI-generated suggestions, particularly for non-technical users?

To answer these questions, this work adopts a research-through-design approach, using a conceptual interactive AI prototype as a design probe rather than building a fully functional AI system. The objective is to explore how interaction design choices, such as user-driven steering, feedback, and on-demand explanation, shape user agency, understanding, and trust during literature review, and to derive actionable design principles for human-Centered AI tools.

2 Related Work

2.1 Design Considerations for Human-AI Systems

2.1.1 ***Sensemaking and External Representations:*** Sensemaking has been widely characterized as an iterative process through which users explore, organize, and refine their understanding of complex information spaces rather than retrieve single answers. Pirolli and Card's sensemaking loop describes a cycle of information foraging, schema construction, and hypothesis refinement, emphasizing the cognitively demanding nature of analytic work [17]. Rather than progressing linearly from query to answer, users repeatedly gather evidence, construct intermediate representations, and restructure their understanding as new information emerges.

Russell et al. further analyze the "cost structure" of sensemaking, showing how different operations, such as searching, encoding, and re-encoding, redistribute cognitive and interaction costs [18]. External representations, including visualizations and structured workspaces, play a central role in reducing cognitive burden and supporting iterative reorganization. These models suggest that

[1] Working Prototype is available at https://cosense.vercel.app/
[2] Code is available at https://github.com/NeelFaganiya12/CoSense

effective tools for literature review should not merely retrieve relevant documents but should also externalize intermediate structures that users can inspect and manipulate.

Recent work extends sensemaking into AI-assisted contexts, where AI systems surface patterns, cluster information, or generate summaries to assist exploration [10]. However, many such systems focus on improving retrieval quality or summarization accuracy rather than supporting the iterative restructuring that characterizes deeper understanding. This work builds on classical sensemaking theory by treating AI-generated clusters and rankings as provisional structures that users can question, reorganize, and refine.

2.1.2 ***Human-AI Collaboration and Mixed-Initiative Interaction:*** Prior work in human-AI interaction emphasizes the importance of preserving user agency and designing systems that support effective human–AI collaboration [2, 3]. Mixed-initiative interaction frameworks argue that intelligent systems should allow users to guide, correct, and override automated behavior rather than passively accept system outputs [2, 9]. Such approaches position AI as a collaborator rather than an autonomous decision-maker.

Decision guidelines for human-AI interaction further recommend that systems make their capabilities and limitations visible, support user control, and enable meaningful feedback [2]. These principles are particularly important in high-stakes or interpretive tasks, where users must maintain ownership of decisions. Recent research also demonstrates that optimizing AI systems solely for accuracy may not produce the best outcomes in collaborative settings; instead, models should be evaluated in terms of how they support overall human-AI performance [3]. Together, this body of work suggests that collaboration requires transparency, controllability, and opportunities for user intervention.

While prior research has established general principles for mixed-initiative interaction, fewer studies examine how these principles apply specifically to scholarly workflows such as literature review. Our work operationalizes mixed-initiative collaboration in the context of AI-assisted sensemaking by introducing mechanisms for scope steering, structural feedback, and explanation inspection.

2.1.3 ***Transparency, Explanation, and Trust Calibration:*** Transparency and explanation play a central role in shaping users' mental models of AI systems. Work on explanatory debugging demonstrates that explanations are most effective when they are interactive and allow users to question and correct system reasoning rather than merely consume static justifications [1]. Such interaction helps users build more accurate mental models and supports calibrated trust.

Research on trust in automation further shows that highly accurate AI systems can inadvertently encourage over-reliance, particularly when users lack insight into how outputs are produced [3]. Over-trust may occur when users perceive systems as authoritative without understanding their limitations. These findings suggest that explanations should be designed not simply to justify system outputs but to support trust calibration and critical engagement.

In the context of literature reviews, where users must synthesize diverse perspectives and detect subtle differences across papers, an inappropriate reliance on AI-generated clusters or summaries could prematurely shape interpretations. By enabling users to inspect grouping rationales and provide corrective feedback, this work investigates how transparency mechanisms influence trust and reliance in AI-assisted scholarly workflows.

2.1.4 ***Prompt-Based Interaction and Non-Expert Users:*** Large language models (LLMs) have popularized prompt-based interaction as a primary mechanism for controlling AI systems. However, recent studies highlight challenges faced by non-technical users when interacting with prompt-driven interfaces. Zamfirescu-Pereira et al. show that non-expert users often rely on vague prompts and expect AI systems to infer intent in a human-like manner, leading to frustration, self-blame, or misplaced trust when outputs do not align with expectations [23]. Prompt formulation thus becomes a cognitive burden rather than a seamless interaction.

Complementary work on prompting in creative applications argues that effective use of generative models requires dedicated interface support rather than relying solely on trial-and-error prompting [6]. These findings suggest that prompt-based interaction alone may be insufficient for complex, evolving tasks such as literature review, where goals shift dynamically and require iterative restructuring.

Related research on intelligent systems further indicates that users prefer mechanisms that allow them to inspect, question, and refine system behavior rather than repeatedly reformulate prompts [13]. Together, this body of work motivates interaction designs that reduce reliance on prompt engineering and instead support intent expression through feedback, structural manipulation, and transparency.

2.2 AI-Assisted Literature Review Systems

AI has increasingly been applied to support literature review, evidence synthesis, and scholarly discovery workflows. Prior work in this space broadly falls into three categories: **(1)** Systematic review automation, **(2)** Exploratory scholarly discovery tools, and **(3)** large language model (LLM)-based literature assistants. While these approaches improve efficiency and access to information, they differ in how they conceptualize user involvement and interaction.

2.2.1 ***Systematic Review Automation:*** A substantial body of research focuses on reducing manual workload in systematic reviews through machine-learning-based screening and prioritization. Active learning and relevance classification techniques have been used to assist abstract screening and study selection, to reduce manual workload and improve efficiency [11, 20, 21]. These systems are typically embedded within structured review protocols and conceptualize AI primarily as a filtering mechanism that accelerates predefined tasks.

Although effective in improving efficiency, interaction is often limited to accepting or rejecting AI suggestions, with limited support for exploratory restructuring or interpretive sensemaking. As a result, these systems primarily support efficiency within well-defined review pipelines rather than open-ended scholarly exploration (e.g., AI tools assist screening but do not replace human judgment, and their usability and adaptability remain limited across diverse contexts) [4, 14].

2.2.2 ***Scholarly Discovery Tools:*** Beyond systematic review contexts, AI-powered tools for scholarly discovery platforms such as Semantic Scholar, Connected Paper, and ResearchRabbit provide semantic search, citation network visualizations, clustering, and

automated summarization to help users navigate large research landscapes by surfacing related papers, influential works, or thematic grouping [7].

These systems improve discoverability and overview generation; however, their interaction models often rely on ranked lists, static clusters, or network graphs presented for inspection. Opportunities for users to meaningfully restructure AI-generated organizations, modify grouping rationales, or provide structural feedback are typically limited. AI-generated clusters are frequently treated as informative recommendations rather than provisional structures open to revision [5].

2.2.3 ***LLM-Based Literature Assistants:*** More recently, LLM-based systems have been introduced and are trying to enhance conversational interfaces for querying and synthesizing research literature [22]. These tools enable cross-document summarization and natural language question answering, lowering barriers to access and accelerating overview generation [8].

However, prompt-based interaction can obscure intermediate reasoning and make it difficult for users to inspect relationships among source documents [8]. By centring on answer generation, such systems may shift workflows toward information consumption rather than iterative comparison and schema construction.

Across these categories, AI-assisted literature review systems predominantly prioritize efficiency, retrieval relevance, or summarization quality. Fewer systems explicitly frame literature review as an ongoing sensemaking process in which AI-generated structures are provisional and subject to user-driven refinement. In contrast, this work conceptualizes AI outputs as intermediate representations within a collaborative sensemaking loop, foregrounding scope steering, structural feedback, and transparency mechanisms to support user-agency and calibrated trust.

3 Methodology

I li

3.1 Research Approach

This study adopts a research-through-design approach to explore how interaction design choices shape user agency, sensemaking, and trust in AI-assisted literature review. Research-through-design treats the creation of artifacts not merely as implementation work, but as a method for inquiry. Rather than testing a predefined hypothesis about system preference or performance, this approach uses the design and reflective evaluation of a prototype to generate insights about user interaction, interpretive processes, and emerging design tensions within a particular design space [17].

We developed a conceptual interactive prototype as a design probe to investigate how different interaction mechanisms, such as scope steering, AI-proposed organization, and on-demand explanations, affect users' understanding and control during literature review. The goal is not to build a fully functional AI backend or evaluate model accuracy, but to examine how users engage with and interpret AI-generated structures within a structured workspace. By abstracting away algorithmic complexity and simulating AI behaviors, the study isolates design as the primary variable of interest.

This approach is particularly appropriate for the research questions posed in this work, which focus on how design choices influence perceptions of agency, transparency, and trust. These phenomena are experiential and interaction-dependent, making exploratory, design-centered inquiry more suitable than controlled performance-based evaluation. Insights generated through iterative interaction with the prototype inform the derivation of design principles for human-centered AI tools that support literature review as an ongoing sensemaking activity.

3.2 Prototype Design

The prototype is designed as an interactive literature review workspace that simulates AI-assisted organization and synthesis while foregrounding user control. Rather than implementing a fully functional retrieval or generative system, the prototype models key interaction behaviors commonly found in AI-assisted tools. This design choice allows the study to isolate and examine interaction mechanisms without conflating them with algorithmic performance.

As shown in Figure 1, the interface consists of three primary regions: (1) a steering panel for defining scope and constraints, (2) an AI-proposed organization panel that presents clustered or ranked papers, and (3) an inspection and feedback panel that enables explanation and corrective interaction.

3.2.1 ***Steering and Scope Definition.*** The left panel provides users with mechanisms to define and refine the scope of their literature review. Users can specify a research topic, adjust the search scope using a scope slider (narrow to broad), and explicitly include or exclude keywords. This design models *scope steering* as a first-class interaction mechanism rather than a single static query. By allowing users to adjust constraints iteratively, the system frames literature exploration as a dynamic process rather than a one-shot retrieval task.

An *explainability* setting further allows users to control the verbosity of AI explanations. This reflects the design hypothesis that explanation depth should be adjustable rather than universally imposed, enabling exploration of how explanation granularity influences perceived understanding and cognitive load.

3.2.2 ***AI-Proposed Organization.*** The central panel presents the AI-proposed cluster of papers. Rather than displaying a flat ranked list, the system groups papers into thematic categories (e.g., "Attack mechanisms", "Evaluation methodology", "HCI sensemaking"). Each cluster includes representative cues such as top keywords and brief metadata summaries.

Clusters are treated as provisional structures rather than authoritative categorization. Users can inspect why a paper appears within a cluster, request explanations of the grouping rationale, and observe simplified reasoning signals (e.g., shared keywords or thematic similarity). This design externalizes intermediate AI structure to support interpretive inspection.

3.2.3 ***Paper Queue and Inspection.*** A secondary review queue displays individual papers with tags and metadata, allowing users to inspect items outside the cluster view. Selecting a paper reveals a detailed panel with title, authors, venue, citation count, tags, and a summary. The interface explicitly displays an "AI confidence" indicator to make uncertainty visible rather than implicit.

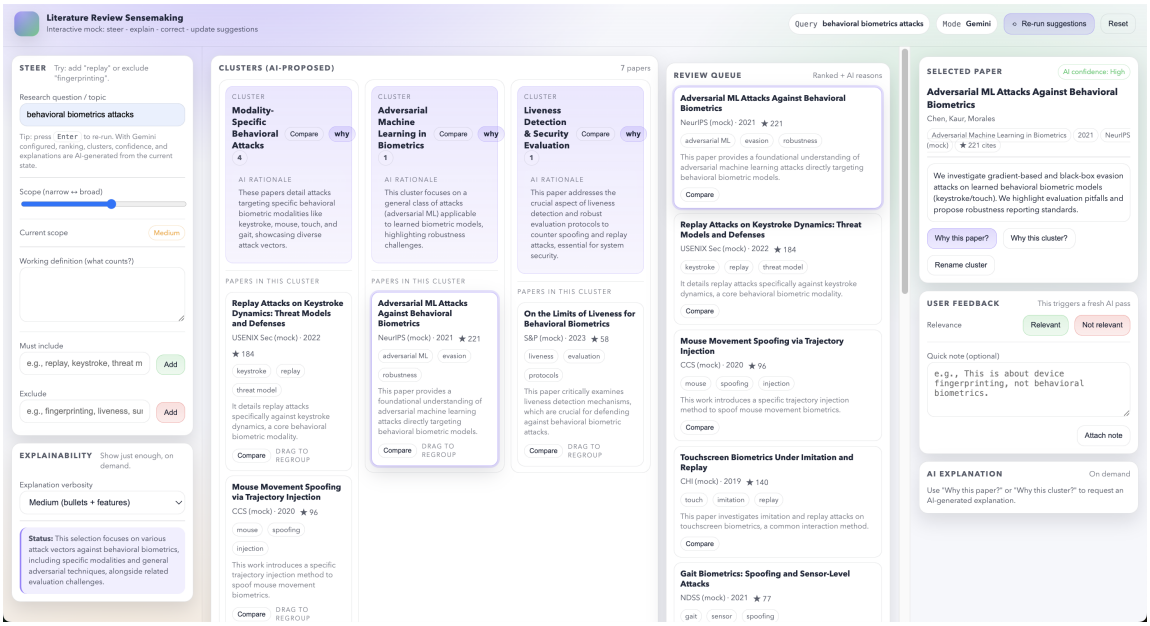

Figure 1: Overview of the interactive prototype interface. The layout consists of (A) scope steering controls, (B) AI-proposed clusters, and (C) inspection and feedback panel.

Explanation buttons (e.g., "Why this paper?" or "Why this cluster?") allow users to request on-demand rationales. Explanations are intentionally simplified and transparent, prioritizing interpretability over technical detail.

3.2.4 ***Feedback and Structural Correction.*** The prototype includes explicit mechanisms for user feedback. Users can mark papers as *relevant* or *not relevant*, add notes, and rename clusters. These actions update the AI-proposed organization, simulating how corrective feedback influences subsequent recommendations. Rather than restricting feedback to binary relevance judgments, the design allows structural intervention (e.g., questioning groupings), reinforcing the framing of AI as a collaborator rather than an oracle.

3.2.5 ***Simulation Approach.*** Our prototype operates over a corpus of candidate papers from OpenAlex, while ranking, clustering, and explanations are generated dynamically from the participant's current query, constraints, scope settings, and feedback using a Gemini model; a deterministic heuristic fallback is also used only when the model is unavailable or returns invalid output, which lets the user see and operate over a fixed local corpus of candidate papers. This design kept the document space constant across participants while still studying adaptive AI-supported interaction, enabling a focused examination of how scope steering, clustering, explanations, and feedback influenced sensemaking, perceived agency, and trust.

3.3 Participants

The study will involve a small group of participants recruited through convenience sampling from the academic and professional networks. We include 10 participants, balanced across two categories: (1) technically experienced users with prior exposure to AI tools or research workflows, and (2) non-technical users with limited experience using AI-assisted systems for academic tasks. This distinction was intentional, as prior research suggests that expectations, trust calibration, and prompting behaviors differ significantly between expert and non-expert users [23].

Participants included individuals who have prior experience conducting literature reviews, such as graduate and senior undergraduate students. While familiarity with AI tools was not required, participants must have had basic experience reading and organizing scholarly materials to ensure meaningful engagement with the prototype.

Given the exploratory and design-oriented nature of this study, the sample size is appropriate for identifying recurring interaction patterns and usability tensions rather than for statistical generalization. The goal was to surface qualitative insights into how different interaction mechanisms influence users' sense of agency, understanding, and trust.

Sessions were conducted in an informal research setting, i.e., think-aloud walkthroughs. No sensitive personal data was collected beyond basic background information related to technical experience and familiarity with literature review tasks.

3.4 Procedure

Each session lasts approximately 30-45 minutes and was conducted individually. Participants first received a brief introduction to the study's purpose, framed as an exploration of interaction design in AI-assisted literature review tools. They were then given a short tutorial explaining the interface layout and available interaction mechanisms, including scope steering, cluster inspection, explanation requests, and feedback controls. They then completed the exploratory tasks designed to simulate literature review activities.

Following the introduction, participants completed a set of structured, exploratory tasks designed to simulate a literature review. For example, they were asked to: (1) narrow or broaden the scope of a topic using the steering controls, (2) inspect AI-generated clusters and identify which appear most relevant to a particular research question, (3) request explanations for specific groupings, and (4) provide feedback by marking papers as relevant or reorganizing clusters. These tasks were intentionally kept open-ended to encourage natural interaction and exploration rather than task-completion speed.

3.4.1 ***Think-Aloud Session.*** Throughout the session, participants were asked to think aloud, verbalizing their expectations, interpretations, confusion points, and reasoning processes as they interacted with the prototype. When necessary, participants were gently prompted (e.g., "What do you think the system is doing here?" or "Why did you choose to adjust the scope?"), while avoiding directing their decisions.

Sessions were documented through written observation notes and, where permitted, audio recordings for later reference. The focus of data collection was on identifying patterns in how users interpret AI-generated structures, respond to explanations, adjust scope, and provide user feedback, as well as how these interaction mechanisms influence their perceived sense of control and trust.

3.4.2 ***Post-Study Questionnaire.*** Following the think-aloud session, participants completed a brief post-study questionnaire consisting of six Likert-scale items (*1 = Strongly Disagree, 5 = Strongly Agree*) designed to capture perceived agency, understanding, and trust. The questionnaire was intended to capture participants' reflective impressions after interacting with the system. The items were:

- I felt in control while using the system
- The system responded predictably to my inputs
- I understood why papers were grouped the way they were
- The explanations helped me understand the system's reasoning
- I felt confident relying on the system's suggestions
- I trusted the system to support my literature review process

These items were analyzed descriptively to triangulate qualitative findings rather than for statistical inference.

3.5 Data Collection and Analysis

Data was collected through a combination of observational notes, participant think-aloud verbalizations, and optional audio recordings when permitted. During each session, the participant's notable behaviors, hesitations, clarification requests, scope adjustments, and interactions with explanation or feedback mechanisms were documented. Particular attention was given to how participants interpret AI-generated clusters, how they respond to explanations, and how they describe their sense of control or trust.

Following data collection, session notes and transcripts were reviewed using an inductive thematic analysis approach. Observations were grouped into recurring patterns related to agency, understanding, trust calibration, and interaction breakdowns. Rather than coding for quantitative frequency, the analysis focused on identifying meaningful design tensions, unexpected user behaviors, and differences between technically experienced and non-technical participants.

Emerging themes were iteratively refined and synthesized into actionable design insights. These insights informed the articulation of design principles of AI-assisted literature review systems that prioritize user-driven steering, structural transparency, and calibrated trust. The goal of the analysis was not statistical generalization, but the generation of interaction-level design guidance grounded in observed user behavior. Survey responses were analyzed descriptively (e.g., through summary statistics and response distribution) to triangulate qualitative findings rather than for statistical inference of the collected data.

4 Results

4.1 RQ1 - Agency and Perceived Control

4.1.1 ***Scope Steering as a Primary Mechanism of Agency.*** Across both participant groups, scope steering (narrow ↔ broad) emerged as a central mechanism for perceived control. Participants consistently articulated clear expectations about how narrowing or broadening scope would affect specificity and coverage, and most reported that the system's behavior aligned with those expectations.

Non-technical participants described scope steering as helping them *"focus on what they care about"* and reduce irrelevant information. Technical participants articulated more detailed mental models, describing scope as adjusting retrieval precision or filtering criteria. In both groups, the ability to iteratively adjust scope reduced the need to reformulate full prompts, shifting interaction from one-shot query generation to incremental refinement. This qualitative finding was reflected in post-study responses, where the mean rating for agency-related items were high for both groups (technical: $M = 4.8$; non-technical: $M = 4.3$; see Figure 2), aligning with participants' reports that scope steering and working definitions strengthened their perceived control over the process.

Interpretation: Making scope adjustable through a persistent UI element, rather than hidden in prompt text, appears to externalize control and reinforce user ownership of the retrieval process. Instead of reformulating entire queries, users could manipulate a visible parameter and observe incremental changes in the output. This shift from hidden prompt engineering to explicit interaction suggests that agency in AI-assisted systems is supported not only by outcome quality but also by the visibility and manipulability of control mechanisms. By rendering refinement as an interactive continuum rather than a discrete query submission, scope steering transforms retrieval from a transactional exchange into an ongoing, user-directed process.

4.1.2 ***Working definition.*** Participants across expertise levels emphasized the importance of the *working definition* field - a free-text area intended for specifying inclusion criteria, conceptual boundaries, or clarifying the intended framing of the research topic. Non-technical users interpreted it as a way to clarify what is *"in scope"* versus *"out of scope"*. Technical users treated it as a fine-grained control layer that meaningfully shaped system behavior. Several technical participants explicitly stated that without a working definition, the AI might misinterpret their intent. This suggests

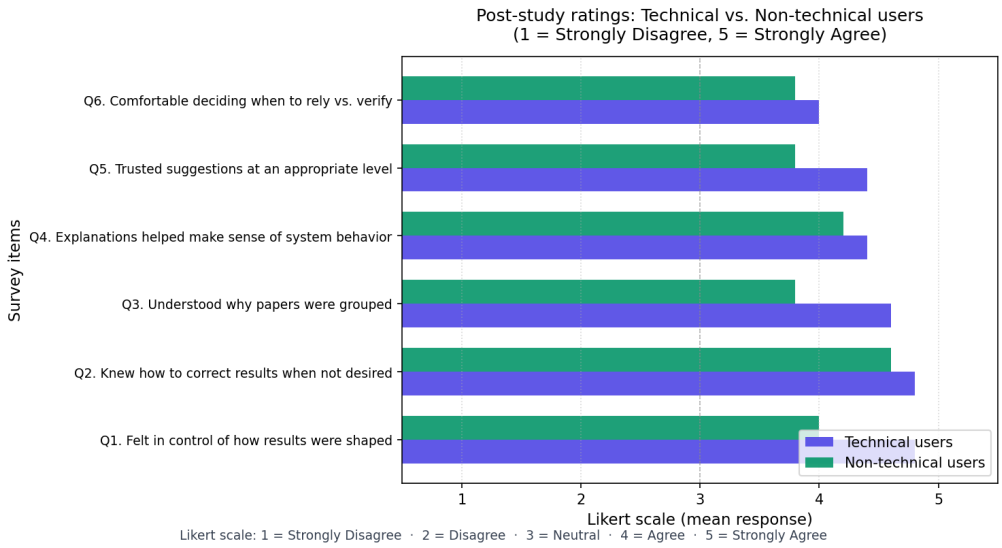

Figure 2: Mean post-study Likert ratings (1 = Strongly Disagree, 5 = Strongly Agree) comparing technical and non-technical participants across six items assessing perceived agency (Q1, Q2), understanding (Q3, Q4), and trust(Q5, Q6).

that definitional framing functions not merely as input text, but also as a psychological anchor for perceived control.

Interpretation: Users experience agency not only when they can adjust results, but also when they feel their intent has been explicitly encoded in the system. The working definition field appears to serve as more than a query refinement tool; it provides a visible space for users to articulate conceptual boundaries and expectations. This articulation process itself contributes to perceived control, as users see their own framing reflected in the system's output. Designing for agency in AI-assisted tools may therefore require not just an adaptive algorithm, but explicit mechanisms that allow users to formalize and externalize their intent.

4.1.3 ***User Feedback as Control.*** The prototype allowed participants to mark papers as *relevant* or *not relevant*, which triggered a refresh of suggested papers and, in some cases, an update to cluster compositions. This mechanism was intended to simulate adaptive refinement based on user feedback.

Marking papers as *relevant* or *not relevant* generally increases users' sense of control. Participants described feedback as "training" the system and narrowing results toward their interests. However, a subtle tension emerged. Some technical participants expressed skepticism that feedback would consistently influence AI behavior, citing prior experiences with conversational LLMs. In one case, marking a paper as *not relevant* led to the removal of another paper the participant valued, reducing perceived control. This indicates that feedback mechanisms increase agency only when their effects are predictable and interpretable.

Interpretation: Feedback strengthens perceived agency, but opaque adaptation can undermine it. While users value mechanisms that allow them to refine results iteratively [15], unexpected structural changes, such as the removal of related papers or the clusters, can reduce confidence in the system's behavior. Designing adaptive AI systems, therefore, requires not only incorporating user feedback but also communicating how feedback reshapes the underlying result space.

4.2 RQ2 - Understanding and Explanation

4.2.1 ***Clustering as Structural Sensemaking.*** Clustering emerged as one of the strongest contributors to understanding. Participants frequently described clusters as:

- Organized
- Like a mind map
- More structured than traditional AI bullet outputs

Rather than scanning a flat list of papers, users interpreted clusters as thematic groupings that supported a conceptual overview. Technical users, in particular, contrasted clustering with conversational AI interfaces, noting that structure reduced cognitive load and supported synthesis. Non-technical participants initially needed clarification about what the clusters represented, but once they understood, they described them as useful for organizing ideas and tracking subtopics. These qualitative impressions were reflected in post-study ratings (Figure 2), where understanding-related questions received high mean ratings from both the groups (technical: $M = 4.5$; non-technical: $M = 4.0$), reinforcing that clustering supported interpretive sensemaking across both the expertise levels.

Interpretation: Exposing AI's intermediate structure (groupings) supports interpretive sensemaking rather than passive consumption. Instead of presenting results as a ranked list, clustering externalizes the system's organization of knowledge, allowing users to see the relationships between subtopics. This structural visibility enabled users to compare thematic areas, identify gaps, and reason about coverage at a higher level of abstraction. In this way, clustering papers together helped users to shift their interaction from selecting individual papers to navigating conceptual terrain.

4.2.2 ***Explanation Depth Preferences Diverge by Expertise.*** Explanation verbosity *(brief/medium/deep)* revealed expertise-based differences:

- Technical participants preferred deep explanations, requesting metrics, technical details, and counterfactual reasoning
- Non-technical participants often preferred brief or medium explanations unless exploring unfamiliar topics

Despite differing preferences, both groups used *"Why this paper?"* frequently to resolve uncertainty. Explanations were particularly helpful when cluster membership was initially unclear. However, some technical users found the explanations too generic, suggesting that transparency alone is insufficient if explanations lack domain depth.

Interpretation: Explanation depth should be adaptive, and perceived usefulness depends on alignment with user expertise and task goals. For technical users, shallow explanations may feel insufficient or overly generic, reducing confidence in the system's reasoning. For non-technical users, overly detailed explanations can increase cognitive load without improving understanding. Designing explanation mechanisms as adjustable scaffolds rather than fixed outputs would allow users to calibrate the informational depth to their current needs, supporting both rapid scanning and deeper verification.

4.3 RQ3 - Trust and Calibrated Reliance

4.3.1 ***Trust as Conditional Alignment.*** Participants did not demonstrate blind trust. Instead, trust appeared contingent on observable alignment between:

- Input (scope, keywords, feedback)
- Output (cluster updates, paper changes)

Non-technical users often reported increased confidence when the system responded visibly to feedback, whereas technical users were more likely to express cautious trust, emphasizing verification and personal judgment. These qualitative patterns were reflected in post-study ratings (Figure 2), where trust-related questions received moderately high mean scores (technical: $M = 4.2$; non-technical: $M = 3.8$). While both the groups expressed overall agreement, the slightly lower mean among non-technical participants aligns with observed moments where confidence decreased when feedback led to unexpected changes in cluster composition or paper removal.

Interpretation: Trust emerges not from explanation alone, but also from consistent behavioral alignment between user action and system response. Participants dynamically calibrated their reliance, increasing trust when adjustments (e.g., scope changes and feedback) led to predictable changes in results. Conversely, unexpected updates weakened confidence, even when explanations were available. This suggests that calibrated trust in AI-assisted workflows depends not only on transparency, but also on observable responsiveness and interaction-level coherence over time.

4.3.2 ***Feedback Transparency and Trust Calibration.*** When feedback effects were visible and intuitive, trust increased. When feedback led to broader structural changes (e.g., cluster merging or paper removal), confidence sometimes decreased, especially when a needed paper is removed. This suggests that trust calibration depends on the system's ability to communicate and not just *what* has changed, but also *why* it changed.

4.4 Cross-Group Differences

Clear contrasts emerged between technical and non-technical participants:

- **Mental Models**: Technical users reasoned about underlying retrieval logic, while non-technical users focused on interface cues and visible effects
- **Explanation Use**: Non-technical users requested explanations frequently to build understanding, while technical users used explanations strategically to verify or critique the AI's suggestions
- **Trust Orientation**: Technical users were more skeptical but analytically confident, while non-technical users demonstrated experiential trust when feedback behaved predictably
- **Feedback Relevance**: Technical users relied heavily on feedback and working definition to steer the model to provide them with their expected papers, while non-technical users relied more on topic input and cluster inspection

The results, therefore, highlight that transparency and control mechanisms are not universally experienced in the same way. Instead, they function as interpretive scaffolds whose perceived value depends on users' mental models of AI behavior. Designing for mixed-expertise audiences requires not only adjustable interaction mechanisms, but also mechanisms that support alignment between system behavior and user expectations. These findings also suggest that expertise shapes not only trust levels but also how users conceptualize AI collaboration, whether as a controllable retrieval system, an adaptive assistant, or a structured exploration tool.

5 Discussion and Future Work

Across findings, AI-assisted literature review was experienced not as retrieval or answer generation process, but as an interactive sensemaking process. Participants relied heavily on scope steering, clustering, and feedback mechanisms to iteratively refine their understanding of a research area. Agency emerged when users could visibly encode intent (through scope or working definitions) and observe coherent changes in system behavior. Rather than blindly trusting the system, participants dynamically calibrated their reliance, increasing confidence when outputs aligned predictably with their inputs and recalibrating when behavior appeared inconsistent.

Clustering played a particularly important role in supporting understanding. By exposing intermediate thematic structure, the system enabled users to reason at the level of subtopics rather than isolated documents. This structural visibility reduced cognitive load and supported higher-level comparison and organization. Explanations further supported interpretation, but their usefulness depended on the alignment between depth and expertise: technical users sought detailed justification, while non-technical users prioritized clarity and brevity. These findings suggest that explanation mechanisms should function as adjustable scaffolds rather than static justifications.

Feedback mechanisms introduced a productive tension between agency and uncertainty. Marking papers as *relevant* or *not relevant* increased users' sense of control when changes were predictable

and interpretable. However, unexpected shifts in cluster composition sometimes reduced confidence, especially among technically experienced users. This suggests that calibrated trust depends not only on transparency, but also on interactional coherence. Taken together, the results frame AI-assisted literature review as a collaborative sensemaking activity in which structure, steerability, and predictable feedback loops are central to agency and calibrated trust. Designing such systems required moving beyond optimizing retrieval accuracy or summarization quality toward supporting interaction-level coherence, structural transparency, and adaptable explanation depth.

This study was exploratory and design-oriented, using a small sample of participants and a semi-simulated AI backend. Future work should examine these interaction mechanisms in a fully deployed system with a larger and more diverse participant population. Longitudinal studies could further investigate how agency and trust evolve over repeated use. Additionally, direct comparisons between structured, cluster-based interfaces and purely conversational AI tools would help clarify the role of exposed intermediate structure in supporting sensemaking.

6 Conclusion

This study examined how interaction design choices in AI-assisted literature review tools shape users' sense of agency, understanding, and trust. Our findings suggest that agency emerges when users can visibly encode intent and observe coherent changes in system behaviour through mechanisms such as scope steering, clustering, and feedback. Trust was not established solely through transparency; rather, it depended on predictable alignment between user actions and system responses. We also observed meaningful differences between technical and non-technical participants. Technical users relied more heavily on feedback mechanisms to iteratively refine results, while non-technical users emphasized definitional framing and scope clarification. These differences highlight the importance of mixed-expertise design that supports multiple interaction styles while maintaining structural transparency. Together, the results position AI-assisted literature review not as automated answer generation process, but as collaborative sensemaking process. Designing such systems requires prioritizing steerability, exposed intermediate structure, and interaction-level coherence to support calibrated trust and user ownership.

7 Ethics

This study was conducted in accordance with established ethical guidelines for human-centered research. Participation was voluntary, and all participants provided informed consent before the study. Participants were informed about the purpose of the research, the nature of the tasks, and the approximate duration of the session. No personally identifiable information was collected beyond their technical background and familiarity with the literature review. All collected data, including think-aloud audio recordings, observation notes, and questionnaire responses, were anonymized before the analysis. Audio recordings were used solely for transcription and analysis and were deleted after transcription, where applicable. The prototype was designed to support transparency and calibrated trust in AI-assisted decision-making. Ethics approval for this study was obtained in accordance with the University of Waterloo's institutional research ethics guidelines.

References

[1] Saleema Amershi, Maya Cakmak, W. Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *Ai Magazine* 35 (12 2014), 105–120. doi:10.1609/aimag.v35i4.2513

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300233

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. 2021. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. arXiv:2004.13102 [cs.AI] https://arxiv.org/abs/2004.13102

[4] Simon Baradziej. 2023. The influence of Large Language Models on systematic review and research dissemination. doi:10.7557/5.7240

[5] Louis Anthony Cox. 2025. An AI assistant for critically assessing and synthesizing clusters of journal articles. *Global Epidemiology* 10 (2025), 100207. doi:10.1016/j.gloepi.2025.100207

[6] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. arXiv:2209.01390 [cs.HC] https://arxiv.org/abs/2209.01390

[7] Mark Glickman and Yi Zhang. 2024. AI and Generative AI for Research Discovery and Summarization. arXiv:2401.06795 [cs.CL] https://arxiv.org/abs/2401.06795

[8] Aditi Godbole, Jabin Geevarghese George, and Smita Shandilya. 2024. Leveraging Long-Context Large Language Models for Multi-Document Understanding and Summarization in Enterprise Applications. arXiv:2409.18454 [cs.CL] https://arxiv.org/abs/2409.18454

[9] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. doi:10.1145/302979.303030

[10] Jad Kabbara, Thanh-Mai Phan, Marina Rakhilin, Maya E Detwiller, Dimitra Dimitrakopoulou, and Deb Roy. 2025. AI-assisted sensemaking: Human-AI collaboration for the analysis and interpretation of recorded facilitated conversations. 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 655, 8 pages. doi:10.1145/3706599.3706688

[11] Mihiretu M. Kebede, Charlotte Le Cornet, and Renée Turzanski Fortner. 2023. In-depth evaluation of machine learning methods for semi-automating article screening in a systematic review of mechanistic literature. *Research Synthesis Methods* 14, 2 (2023), 156–172. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jrsm.1589 doi:10.1002/jrsm.1589

[12] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) *(IUI '15)*. Association for Computing Machinery, New York, NY, USA, 126–137. doi:10.1145/2678025.2701399

[13] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, 1–15. doi:10.1145/3313831.3376590

[14] Maarten Moens, Guy Nagels, Nicholas Wake, and Lisa Goudman. 2025. Artificial intelligence as team member versus manual screening to conduct systematic reviews in medical sciences. *iScience* 28, 10 (2025), 113559. doi:10.1016/j.isci.2025.113559

[15] J. Nielsen. 1993. Iterative user-interface design. *Computer* 26, 11 (1993), 32–41. doi:10.1109/2.241424

[16] R. Parasuraman, T.B. Sheridan, and C.D. Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30, 3 (2000), 286–297. doi:10.1109/3468.844354

[17] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis.

[18] 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. doi:10.1145/169059.169209

[19] Mohsen Soori, Fooad Karimi Ghaleh Jough, Roza Dastres, and Behrooz Arezoo. 2024. AI-Based Decision Support Systems in Industry 4.0, A Review. *Journal of Economy and Technology* 4 (2024), 206–225. doi:10.1016/j.ject.2024.08.005

[20] Leonieke M. van den Bulk, Yamine Bouzembrak, Anand Gavai, Ningjing Liu, Lukas J. van den Heuvel, and Hans J.P. Marvin. 2022. Automatic classification of literature in systematic reviews on food safety using machine learning. *Current Research in Food Science* 5 (2022), 84–95. doi:10.1016/j.crfs.2021.12.010

[21] Gerit Wagner, Roman Lukyanenko, and Guy Paré. 2022. Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology* 37, 2 (2022), 209–226. doi:10.1177/02683962211048201

[22] Yifei Yuan, Zahra Abbasiantaeb, Yang Deng, and Mohammad Aliannejadi. 2025. Query Understanding in LLM-based Conversational Information Seeking. arXiv:2504.06356 [cs.CL] https://arxiv.org/abs/2504.06356

[23] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. doi:10.1145/3544548.3581388