ScholarMate: A Mixed-Initiative Tool for Qualitative Knowledge Work and Information Sensemaking

Runlong Ye
harryye@cs.toronto.ca
Computer Science,
University of Toronto
Toronto, Ontario, Canada

Patrick Yung Kang Lee
patricklee@toronto.edu
Computer Science,
University of Toronto
Toronto, Ontario, Canada

Matthew Varona
varona@cs.toronto.edu
Computer Science,
University of Toronto
Toronto, Ontario, Canada

Oliver Huang
oliver@cs.toronto.edu
Computer Science,
University of Toronto
Toronto, Ontario, Canada

Carolina Nobre
cnobre@cs.toronto.edu
Computer Science,
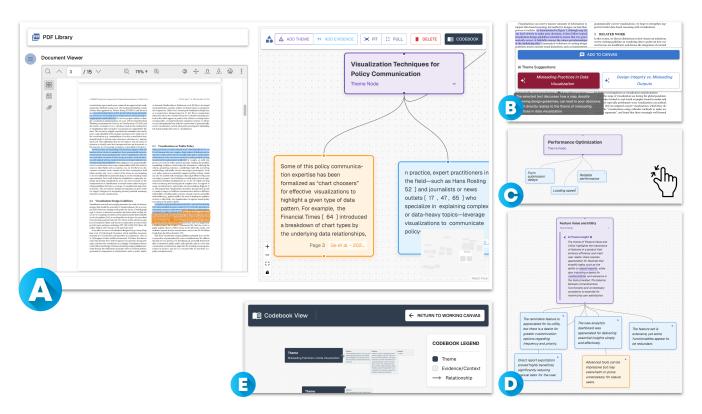University of Toronto
Toronto, Ontario, Canada

Figure 1: Interface and Key Features of ScholarMate. (A) Researchers view source PDFs (left) alongside an interactive free-form canvas (right), snippets extracted from the source document are visually organized on the canvas while preserving source links; (B) Mixed-initiative workflow with AI suggestions helps structure textual snippets into themes; (C) Semantic zoom dynamically adjusts the level of detail shown within evidence nodes (from full text to summaries) as the researcher zooms; (D) Theme insight help researcher stay informed during theme development, hoverable insight keyword highlights corresponding evidence(s); (E) Dedicated codebook view summarizes the thematic coding progress. ScholarMate balances direct manipulation with meaningful AI assistance.

Abstract

Synthesizing knowledge from large document collections is a critical yet increasingly complex aspect of qualitative research and knowledge work. While AI offers automation potential, effectively integrating it into human-centric sensemaking workflows remains challenging. We present ScholarMate, an interactive system designed to augment qualitative analysis by unifying AI assistance with human oversight. ScholarMate enables researchers to dynamically arrange and interact with text snippets on a non-linear canvas, leveraging AI for theme suggestions, multi-level summarization, and evidence-based theme naming, while ensuring transparency through traceability to source documents. Initial pilot studies indicated that users value this mixed-initiative approach, finding the balance between AI suggestions and direct manipulation crucial for maintaining interpretability and trust. We further demonstrate the system's capability through a case study analyzing 24 papers. By balancing automation with human control, ScholarMate enhances efficiency and supports interpretability, offering a valuable approach for productive human-AI collaboration in demanding sensemaking tasks common in knowledge work.

CCS Concepts

• **Human-centered computing → Interactive systems and tools**; **Interaction techniques**; **Empirical studies in HCI**.

Keywords

information seeking; multilevel exploration; sensemaking; levels of abstraction; abstraction hierarchy; large language models; systems thinking; human-AI interaction

1 Introduction

Researchers deal with large amounts of text-based data in their daily workflows, such as reviewing academic publications and conducting qualitative data analysis on interview transcripts. Making sense of these large corpora presents escalating challenges due to the sheer volume of text data being collected and produced. Traditional approaches, such as manual open coding and thematic labeling, require extensive human effort and time, making rapid adjustments to emerging patterns that are challenging when faced with large corpora [4]. This demanding manual process relies heavily on the

researcher's interpretive skill to discern patterns and relationships [32], limiting the scale and efficiency of analysis possible [23].

Recent advances in large language models (LLMs) have demonstrated their capacity to summarize, cluster, and infer patterns from large textual datasets [5]. Despite these capabilities, purely automated solutions can struggle with context awareness and interpretive flexibility, qualities that are critical in qualitative research domains [14, 15]. Moreover, integrating AI raises concerns about automation bias, the potential for diminished researcher cognitive engagement [30], and the risk of over-reliance on outputs that may lack nuance or contain errors [13, 14]. Three core challenges persist in large-scale AI-assisted analysis: (1) Maintaining transparency and traceability when automated summarization systems are introduced [18], (2) ensuring researchers remain actively engaged [13] rather than passively accepting outputs, and (3) supporting creative insight through flexible exploration of emergent connections during the research process [2].

These challenges lead us to our primary research question: ***How can an interactive system for qualitative knowledge work effectively weave AI-generated insights with human exploration and critical control, supporting purposeful sensemaking?*** To address this, we follow the human-centered approach to AI proposed by Shneiderman [29], where machines provide computational power while humans guide, interpret, and refine outputs.

We implement this in **ScholarMate**, an LLM-assisted system designed to support researchers conducting qualitative research work and facilitating the text-based sensemaking process. ScholarMate allows researchers to import a text corpus and manipulate snippets on a non-linear canvas interface, enabling them to build working theories of related documents. This visual re-arrangement process specifically supports human-AI collaborative sensemaking. Furthermore, ScholarMate explores how human-computer interaction (HCI) can shape future work practices involving AI partners, aiming for meaningful and productive outcomes in complex knowledge work.

2 Related Work

Our work builds upon prior research exploring computational assistance for qualitative research and text sensemaking. We situate ScholarMate relative to approaches focusing on enhanced manual interaction, AI assistance for specific analytic tasks, and emerging mixed-initiative sensemaking systems.

2.1 Enhanced Manual Interaction and Visualization

Prior research on manual interaction and visualization emphasizes fluid user-driven interactions but typically lacks AI capabilities for handling large-scale qualitative analysis. For instance, LiquidText [32][1] enables visually connecting document excerpts for active reading and sensemaking. Similarly, visual analytics research has demonstrated meaningful semantic relationships within user-generated spatial layouts [6]. In practice, qualitative researchers frequently use commercial tools such as Miro[2], facilitating manual affinity diagramming through sticky notes to organize textual data

---

[1] https://www.liquidtext.net/
[2] https://miro.com/

into themes [35]. While valuable, these manual systems are limited in efficiency and scalability. ScholarMate directly addresses these gaps by integrating meaningful AI suggestions alongside intuitive manual interaction.

2.2 AI Assistance for Specific Analytic Tasks

Recent systems leveraging AI, particularly large language models (LLMs), often target specific analytic subtasks. IdeaSynth [27], PersonaFlow [24], and Sensecape [31] exemplify AI's role in idea generation, iterative exploration, and multilevel sensemaking. Similarly, DiscipLink [36] and Synergi [17] facilitate structured scholarly synthesis but impose constraints on flexible, iterative exploration. These solutions typically lack extensive direct manipulation of textual data, crucial for qualitative analysis. Moreover, Anderson et al. [2] highlights LLMs' homogenizing effects on ideation, emphasizing the need for diverse interpretations in analytic processes.

ScholarMate significantly extends this prior work by combining AI-driven thematic suggestions and direct manipulation in a unified, flexible interface, uniquely supporting deeper qualitative exploration without constraining analytic processes.

2.3 Towards Mixed-Initiative Qualitative Sensemaking

ScholarMate builds upon research advocating mixed-initiative systems that synergize human expertise and AI. Scholastic [11] uses structured graph-based interactions for collaborative analysis but offers limited spatial and intuitive flexibility. Similarly, He et al. [9] explore collaborative AI-supported ideation within virtual canvases but primarily in synchronous group contexts, presenting distinct challenges compared to individual or asynchronous qualitative analysis.

ScholarMate innovates by embedding AI within a flexible, canvas-based interface, enhancing interpretability with semantic zoom and dynamically adjusting detail granularity. It explicitly addresses transparency and traceability, linking AI suggestions directly to source documents – areas identified as critical yet underexplored [1, 34]. While other recent systems like LLooM also leverage LLMs for mixed-initiative sensemaking by generating explicit 'concepts' [20], ScholarMate uniquely bridges direct spatial manipulation valued in manual interaction tools [6, 32] and sophisticated AI assistance essential for large-scale qualitative analysis.

3 Design Goals

Building on our prior design guideline in AI-assisted research [34], we outline three primary design goals for ScholarMate. These goals address central challenges by carefully balancing AI assistance with human judgment, providing transparent reasoning for AI-generated suggestions, and adhering to ethical standards to ensure trustworthiness. Below, we summarize each design goal's rationale briefly.

- **DG1: Support Mixed-Initiative Human-AI Collaboration.** While AI can speed up qualitative analysis, over-automation risks undermining critical thinking and nuanced interpretation, potentially causing passive acceptance of flawed outcomes [21, 26], thus, ScholarMate is designed as a mixed-initiative system [12], aiming to incorporate AI assistance

with human expertise. It enables both the user and AI to propose insights while preserving ultimate human control. This design ensures efficiency without compromising essential human judgment and agency [11] in the knowledge co-creation process.

- **DG2: Ensure Interpretability of AI Reasoning.** AI models often act as opaque "black boxes," which can hinder researchers' ability to critically assess their outputs and reasoning processes [25, 34]. To illuminate the AI's reasoning process, ScholarMate provides transparency into the basis of each suggestion. It explicitly links AI-generated suggestions to the specific evidence snippets that prompted them, offering a direct rationale. A visual preview further clarifies the immediate impact of accepting a suggestion. Making this rationale transparent allows researchers to critically evaluate the logic behind a suggestion, facilitating informed interaction: acceptance, revision, or rejection, based on their own understanding [19]. This aligns with human-AI interaction guidelines promoting system intelligibility [1].

- **DG3: Promote Output Validation and Ethical Use.** AI tools can introduce errors through biases or hallucinations, posing ethical challenges and threatening research validity [10, 34]. Therefore, ScholarMate incorporates features focused on validating AI-generated content and ensuring ethical integration. Direct traceability allows researchers to easily verify AI claims or summaries against original source documents, confirming accuracy. Transparently communicating AI limitations and mandating human oversight are crucial mechanisms to mitigate risks associated with potential errors or biases. This emphasis on verification and human control helps maintain ethical standards and fosters justified trust in the final research outcomes.

4 Probe Design

ScholarMate's design was informed by observing qualitative researchers' existing practices. We analyzed how they used informal spatial arrangements and annotations (often sketches) to cluster document excerpts and build understanding during their real-world workflows. Insights from these observations, combined with iterative internal design reviews focused on addressing identified researcher needs, guided the development of successive low- and high-fidelity prototypes.

This process led to the current web-based technology probe [16], to demonstrate the initial functionalities of ScholarMate. The ScholarMate probe can be accessed online[3]. ScholarMate is built using NextJS, with `react-pdf-viewer` and `ReactFlow` components, and designed to support interactive, mixed-initiative qualitative analysis through a flexible, non-linear canvas-based interface. It addresses challenges researchers face in systematically analyzing data while balancing automated assistance with human analytic control. Grounded in mixed-initiative interaction principles [12] and established qualitative research methodologies [4], ScholarMate ensures automated suggestions remain transparent and subject to human validation, recognizing the critical role of human judgment in resolving ambiguities and refining interpretations.

---

[3]https://scholar-mate-eta.vercel.app/

Researchers begin by uploading their corpus into SCHOLARMATE, enabling them to extract relevant text snippets via a built-in document viewer directly onto an interactive canvas. The choice of this canvas-based model aligns with literature findings on spatial document clusterization [6], indicating spatial organization of textual data can meaningfully support researchers in discerning thematic structures in their corpus. This approach proves particularly valuable during exploratory analysis, where ambiguity about themes is most prevalent.

4.1 Interactive Canvas

Within this canvas, extracted snippets become *manipulable nodes*. SCHOLARMATE distinguishes between *evidence nodes*, containing direct excerpts or annotations, and *theme nodes*, used to visually cluster related evidence. Researchers maintain full agency to create, rearrange, connect, edit, and delete nodes, facilitating an iterative process of structuring and refining their analysis. This direct manipulation ensures researchers retain control over data organization and interpretation.

To navigate potentially complex canvas layouts while preserving relative spatial positioning of data, SCHOLARMATE incorporates *Semantic Zoom*. Grounded in visualization principles like Shneiderman's mantra – *"overview first, zoom and filter, then details on demand"* [28] – for exploring large datasets, this allows fluid transitions between high-level thematic overviews and detailed evidence views. As a researcher zooms out, evidence text transforms into progressively shorter, LLM-generated summaries of original content. Conversely, zooming in reveals longer summaries or original text excerpts. While evidence nodes adapt their content detail, theme nodes primarily scale visually with the zoom level, retaining their fixed text to ensure stable thematic anchors.

This combination of direct manipulation and AI-driven semantic zoom primarily supports *Mixed-Initiative Collaboration (DG1)* by blending user control with AI assistance. Furthermore, the crucial ability to trace any evidence snippet on the canvas directly back to its context in the original source document promotes *Output Validation (DG3)*, allowing researchers to easily verify information and trust the evidence underpinning their analysis.

4.2 Working Canvas and Codebook View

SCHOLARMATE provides two distinct but interconnected interfaces: the Working Canvas and the Codebook View. Both interfaces utilize `ReactFlow` component, semantic zoom, node-based interactions, and maintain direct evidence-to-PDF highlighting capabilities.

The *Working Canvas* is designed to support active data analysis, enabling users to freely organize, edit, and refine evidence and theme nodes. It supports full interactivity, allowing detailed modifications, annotations, and iterative analysis, empowering users to co-create with AI, and dynamically explore emerging thematic insights. Conversely, the *Codebook View* offers a stable, read-only visualization of current themes and evidence. To facilitate comprehension without overwhelming users, the Codebook View limits the display to two pieces of evidence per theme. This interface emphasizes high-level theme exploration, providing a quick reference for users to check their understanding or synthesize existing themes.

This dual-interface design directly enhances *Mixed-Initiative Collaboration (DG1)* by providing distinct spaces for dynamic, AI-supported exploration versus human-led refinement and finalization. Separating exploratory work (Working Canvas) from the stable, read-only synthesis (Codebook View) also promotes *Ethical Use (DG3)* by improving the accountability and trustworthiness of the final thematic structure.

4.3 AI-Assisted Features

SCHOLARMATE integrates AI seamlessly throughout the analytical process, enhancing researchers' efficiency and depth of insight. Key AI-assisted functionalities include:

Theme Suggestions: After a user selects a text excerpt, SCHOLARMATE analyzes it and the existing workspace themes. It then suggests either assigning the excerpt to a relevant existing theme or using it to establish a new one. Alternatively, users can choose to manually incorporate the selected excerpt directly onto the workspace, allowing them full control over its placement and connections within their analysis structure. AI suggestions are visibly distinguished by special UI indicators and transparently explained, supporting informed choices and ensuring researchers maintain overall control over their thematic decisions, whether leveraging AI assistance or proceeding manually.

Multi-level Semantic Summarization: SCHOLARMATE employs multi-tiered summarization (medium, short, and tiny) on top of the original passage from context nodes and dynamically adapts to zoom levels on the canvas and functionally supports the semantic zoom feature. This ensures optimal readability and contextual relevance at varying scales.

Contextual Theme Naming: Researchers can utilize AI-generated theme naming functionality that suggests theme names based on the aggregated context of connected evidence nodes. This feature supports the iterative nature of qualitative research by allowing researchers to continually refine theme names as their understanding of the data evolves. It helps researchers quickly capture and communicate the essence of thematic clusters, enhancing conceptual coherence and facilitating iterative sense-making.

Evidence-Grounded Theme Descriptions: Leveraging AI, SCHOLARMATE generates theme descriptions that automatically incorporate salient, evidence-derived keywords. These keywords are interactively linked to the source evidence nodes. Hover interactions provide immediate visual feedback by highlighting these connections, directly grounding the abstract theme description in concrete data and enhancing clarity through traceable AI insights.

These AI-assisted features embody *Mixed-Initiative Collaboration (DG1)* by offering suggestions while preserving user autonomy. Features like Theme Suggestions and Evidence-Grounded Theme Descriptions directly address *Interpretability of AI Reasoning (DG2)* by linking suggestions and keywords to the underlying evidence, explaining why they were proposed. Concurrently, the reliance on source evidence for suggestions and summaries necessitates and supports *Output Validation (DG3)*, encouraging researchers to verify AI contributions against original documents.

5 Evaluation

We evaluate ScholarMate through a small pilot study as well as a use case with a real-world corpus of papers.

5.1 Pilot Study

We conducted a preliminary evaluation of ScholarMate through a small pilot study with two graduate students (P1, P2), both with backgrounds in human-computer interaction and data visualization. We mainly aim to understand: (1) participants' strategies for using ScholarMate in a qualitative research task and (2) their perceptions of its usefulness. Participants were asked to complete a thematic analysis task on two familiar research papers [7, 33], pre-loaded into ScholarMate with researcher-provided highlights matching two predefined themes. Participants, however, were not informed of these pre-identified themes and were asked to independently identify themes using ScholarMate. Afterward, we administered NASA-TLX [8] to measure perceived cognitive load and UMUX-LITE [22] to assess system usability, followed by brief semi-structured interviews for qualitative feedback.

5.2 Quantitative Results

While our sample size was small, participants rated ScholarMate highly, with an average satisfaction score of 4.5 out of 5. The NASA-TLX scores indicated low-moderate cognitive demands (aggregated load: 2.41/7), primarily driven by low frustration (average 1.5/7), physical demand (average 1/7), and high perceived task performance (average 6/7). UMUX-LITE results showed strong usability ratings, indicating high participant agreement that the system met their requirements and was easy to use.

5.3 Qualitative Results

Participants identified both strengths and areas for improvement:

5.3.1 Beneficial AI Integration. Participants found the AI suggestions for themes and insights helpful, particularly for managing the coding process. P2 noted the utility of AI suggestions, especially when initial ideas weren't clear: *"I used the AI suggestion when it felt like none of the existing ones fit particularly well" (P2)*. The AI's ability to suggest existing themes was seen as a significant time-saver: *"If this thing can consistently assign the right [theme] that saves me a ton of time" (P2)*. New AI-generated themes were considered valuable starting points, provided they could be modified: *"I think it's still useful to have that option [to use an AI-suggested theme and edit it later]" (P2)*. P1 also acknowledged the AI's capability at the initial stage: *"For each of the highlighted texts... they can do separate things very well" (P1)*.

5.3.2 Freeform Interaction. The canvas-based, flexible nature of organizing themes was appreciated. P2 liked the different ways to categorize highlights: *"There's like all of the modes of categorizing stuff because you have the you can just add to the canvas and leave the categorization free later or make a manual theme node" (P2)*. This flexibility contributed to a positive overall impression: *"it seems like a way I would actually want to do qualitative research" (P2)*. Participants also took vastly different approaches to organizing their evidence and theme nodes: P1 generated a multi-layered tree structure with sub-themes, while P2 preferred a flat structure where multiple themes were combined into larger ones.

5.3.3 Limitation and Future Improvments.

- **Usability:** P2 desired better node management: *"[It would be useful] maybe being able to like drag multiple nodes at once... it's useful to not have to drag every single one" (P2)*. Scaling was also a concern for real-world use: *"the scale, I think, is super different for an actual project... whether it's easy to manage like multiple nodes" (P2)*.
- **AI Limitations:** P1 observed that the insight generation might not properly account for sub-theme structures: *"if I click 'generate insight', they still [don't seem to prioritize the hierarchy]... they don't take too much attention to the sub [themes], like the hierarchical [themes] underneath"* (P1).
- **Feature Requests:** P1 wished for an undo or state history feature: *"sometimes I want to perhaps go back to my last stage... I don't know how to go back to the original one after I combined it" (P1)*. P1 suggested a chat interface: *"perhaps there could be a like a little chat bot... So sometimes I want to combine certain things together. I can ask [the] AI for suggestions before I actually combine those" (P1)*. Both participants discussed the potential for a "top-down" approach, allowing users to guide the AI: *"maybe before I even reading this paper... I can input a general requirement I'm looking for in those papers" (P1)*, or *"give a prompt saying, Hey, this is generally what I'm interested in" (P2)*.

5.4 Use Case: Analyzing Vector Database Papers with ScholarMate

We also evaluate ScholarMate's capabilities through a use case analyzing 24 state-of-the-art vector database papers. We reference the thematic structure established in a graduate seminar course. We utilized the system to automatically generate themes and classify papers using abstracts. ScholarMate showed moderate alignment (7/16 correctly classified) after the first iteration. Interactively refining the system-generated themes substantially improved categorization accuracy (13/16 correctly classified), with higher accuracy achievable by incorporating additional excerpts.

Notably, ScholarMate is not limited to matching predefined categories. Instead, its primary value is enabling iterative sensemaking and thematic exploration. The interactive workspace and semantic zoom capabilities empowered researchers with fluid transitions, moving smoothly from macro-level thematic overviews to micro-level analysis of individual excerpts, highlighting ScholarMate's strength in supporting sophisticated analysis.

6 Limitations and Future Work

As ScholarMate incorporates interactions with LLMs, it faces inherent risks of AI error. To reduce over-reliance on AI, we designed ScholarMate to foster cognitive engagement by providing direct links to source excerpts, explanations for suggested themes, and easy-to-use editing capabilities. Future iterations could enhance the analytical workflow by supporting more complex, realistic hierarchical structures (e.g., themes, sub-themes, nested concepts) to better capture nuanced relationships within the data. Building

upon existing interpretability features, we plan to investigate various interaction designs for human-AI collaboration, evaluating how they influence appropriate trust and reliance based on recent insights into user interactions with LLMs [3]. Additionally, while initial feedback is promising, larger-scale user studies involving diverse researchers and analysts are essential to rigorously evaluate SCHOLARMATE's interactions, scalability, and impacts on real-world practices. Finally, to further streamline analysis, future work could integrate more automated tagging features guided by user-defined codebooks and implement enhanced privacy controls for sensitive data.

7 Conclusion

We introduced SCHOLARMATE, an LLM-assisted visual interactive tool enhancing qualitative analysis through a mixed initiative canvas. Guided by design goals of supporting mixed-initiative human-AI collaboration (DG1), ensuring interpretability of AI reasoning (DG2), and promoting output validation and ethical use (DG3), SCHOLARMATE aims to foster effective human-AI partnerships that augment, rather than automate, the nuanced process of qualitative research. Our initial case study and pilot evaluations suggest that SCHOLARMATE can serve as a useful tool for textual data exploration and organization, successfully supporting sensemaking while maintaining the researcher's critical cognitive engagement.

Beyond its primary function in thematic analysis, the flexible architecture of SCHOLARMATE by combining spatial organization, semantic zoom, node-based interactions, and integrated AI assistance, points towards compelling possibilities for broader scholarly applications. We envision researchers leveraging its interactive canvas for tasks such as *visualizing the conceptual structure of literature reviews*, spatially mapping relationships between key publications or concepts, identifying thematic connections across disparate sources, and pinpointing research gaps more systematically. Similarly, SCHOLARMATE could be adapted for *analyzing argumentation structures*, allowing users to map claims, counterclaims, and supporting evidence extracted from scholarly texts. Furthermore, it holds potential as a *dynamic scholarly notebook or personal knowledge management tool*, enabling researchers to visually organize and interconnect notes, emergent ideas, and textual evidence across projects, thereby allowing for conceptual synthesis and supporting the potential for emergent insights.

SCHOLARMATE offers insights for designing future human-AI collaborative work environments, where technology not only enhances productivity and expedites analytical tasks, but also deepens researchers' analytical engagement with their data and ideas. Ultimately, we aim to facilitate purposeful, transparent, fluid, and trustworthy human-AI partnerships in the pursuit of knowledge.

Acknowledgments

References

[1] 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

[2] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization Effects of Large Language Models on Human Creative Ideation. In *Proceedings of the 16th Conference on Creativity & Cognition* (Chicago, IL, USA) *(C&C '24)*. Association for Computing Machinery, New York, NY, USA, 413–425. doi:10.1145/3635636.3656204

[3] Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2024. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. *arXiv preprint arXiv:2412.15584* (2024).

[4] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

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

[6] Alex Endert, Seth Fox, Dipayan Maiti, Scotland Leman, and Chris North. 2012. The semantics of clustering: analysis of user-generated spatializations of text documents. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. 555–562.

[7] Joshua Gorniak, Yoon Kim, Donglai Wei, and Nam Wook Kim. 2024. VizAbility: Enhancing Chart Accessibility with LLM-based Conversational Interaction. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 89, 19 pages. doi:10.1145/3654777.3676414

[8] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. doi:10.1016/S0166-4115(08)62386-9

[9] Jessica He, Stephanie Houde, Gabriel E. Gonzalez, Darío Andrés Silva Moran, Steven I. Ross, Michael Muller, and Justin D. Weisz. 2024. AI and the Future of Collaborative Work: Group Ideation with an LLM in a Virtual Canvas. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work* (Newcastle upon Tyne, United Kingdom) *(CHIWORK '24)*. Association for Computing Machinery, New York, NY, USA, Article 9, 14 pages. doi:10.1145/3663384.3663398

[10] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. 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, 1–16. doi:10.1145/3290605.3300830

[11] Matt-Heun Hong, Lauren A. Marsh, Jessica L. Feuston, Janet Ruppert, Jed R. Brubaker, and Danielle Albers Szafir. 2022. Scholastic: Graphical Human-AI Collaboration for Inductive and Interpretive Text Analysis. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (UIST '22). Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3526113.3545681

[12] 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

[13] Yoyo Tsung-Yu Hou, Wen-Ying Lee, and Malte Jung. 2023. "Should I Follow the Human, or Follow the Robot?" — Robots in Power Can Have More Influence Than Humans on Decision-Making. 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 114, 13 pages. doi:10.1145/3544548.3581066

[14] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. doi:10.1145/3703155

[15] Chia-Chien Hung, Wiem Ben Rim, Lindsay Frost, Lars Bruckner, and Carolin Lawrence. 2023. Walking a Tightrope – Evaluating Large Language Models in High-Risk Domains. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, Dieuwke Hupkes, Verna Dankers, Khuyagbaatar Batsuren, Koustuv Sinha, Amirhossein Kazemnejad, Christos Christodoulopoulos, Ryan Cotterell, and Elia Bruni (Eds.). Association for Computational Linguistics, Singapore, 99–111. doi:10.18653/v1/2023.genbench-1.8

[16] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy,

Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 17–24. doi:10.1145/642611.642616

[17] Hyeonsu B Kang, Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. 2023. Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 43, 19 pages. doi:10.1145/3586183.3606759

[18] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 9332–9346. doi:10.18653/v1/2020.emnlp-main.750

[19] 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

[20] Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLooM. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 766, 28 pages. doi:10.1145/3613904.3642830

[21] Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. ACM. https://www.microsoft.com/en-us/research/publication/the-impact-of-generative-ai-on-critical-thinking-self-reported-reductions-in-cognitive-effort-and-confidence-effects-from-a-survey-of-knowledge-workers/

[22] James R. Lewis, Brian S. Utesch, and Deborah E. Maher. 2013. UMUX-LITE: when there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2099–2102. doi:10.1145/2470654.2481287

[23] Weng Marc Lim. 0. What Is Qualitative Research? An Overview and Guidelines. *Australasian Marketing Journal* 0, 0 (0), 14413582241264619. doi:10.1177/14413582241264619 arXiv:https://doi.org/10.1177/14413582241264619

[24] Yiren Liu, Pranav Sharma, Mehul Jitendra Oswal, Haijun Xia, and Yun Huang. 2024. PersonaFlow: Boosting Research Ideation with LLM-Simulated Expert Personas. arXiv:2409.12538 [cs.HC] https://arxiv.org/abs/2409.12538

[25] Tim Miller. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences. arXiv:1706.07269 [cs.AI] https://arxiv.org/abs/1706.07269

[26] Samir Passi and Mihaela Vorvoreanu. 2022. Overreliance on AI literature review. *Microsoft Research* (2022).

[27] Kevin Pu, K. J. Kevin Feng, Tovi Grossman, Tom Hope, Bhavana Dalvi Mishra, Matt Latzke, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. IdeaSynth: Iterative Research Idea Development Through Evolving and Composing Idea Facets with Literature-Grounded Feedback. arXiv:2410.04025 [cs.HC] https://arxiv.org/abs/2410.04025

[28] Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*. Elsevier, 364–371.

[29] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[30] LINDA J. SKITKA, KATHLEEN L. MOSIER, and MARK BURDICK. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51, 5 (1999), 991–1006. doi:10.1006/ijhc.1999.0252

[31] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 1, 18 pages. doi:10.1145/3586183.3606756

[32] Craig S. Tashman and W. Keith Edwards. 2011. LiquidText: a flexible, multitouch environment to support active reading. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 3285–3294. doi:10.1145/1978942.1979430

[33] Huichen Will Wang, Mitchell Gordon, Leilani Battle, and Jeffrey Heer. 2025. DracoGPT: Extracting Visualization Design Preferences from Large Language Models. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (Jan. 2025), 710–720. doi:10.1109/TVCG.2024.3456350

[34] Runlong Ye, Matthew Varona, Oliver Huang, Patrick Yung Kang Lee, Michael Liut, and Carolina Nobre. 2025. The Design Space of Recent AI-assisted Research Tools for Ideation, Sensemaking, and Scientific Creativity. *arXiv preprint arXiv:2502.16291* (2025).

[35] Shutong Zhang, Tianyu Zhang, Jinghui Cheng, and Shurui Zhou. 2025. Who is to Blame: A Comprehensive Review of Challenges and Opportunities in Designer-Developer Collaboration. *arXiv preprint arXiv:2501.11748* (2025).

[36] Chengbo Zheng, Yuanhao Zhang, Zeyu Huang, Chuhan Shi, Minrui Xu, and Xiaojuan Ma. 2024. DiscipLink: Unfolding Interdisciplinary Information Seeking Process via Human-AI Co-Exploration. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3654777.3676366