Get on the Train or be Left on the Station: Using LLMs for Software Engineering Research

Bianca Trinkenreich
bianca.trinkenreich@colostate.edu
Colorado State University
Fort Collins, USA

Fabio Calefato
fabio.calefato@uniba.it
University of Bari
Bari, Italy

Geir Hanssen
ghanssen@sintef.no
SINTEF
Trondheim, Norway

Kelly Blincoe
k.blincoe@auckland.ac.nz
University of Auckland
Auckland, New Zealand

Marcos Kalinowski
kalinowski@inf.puc-rio.br
PUC-Rio
Rio de Janeiro, Brazil

Mauro Pezzè
mauro.pezze@usi.ch
USI Università della Svizzera Italiana
Lugano, Italy

Paolo Tell
pate@itu.dk
IT University of Copenhagen
Copenhagen, Denmark

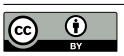
Margaret-Anne Storey
mstorey@uvic.ca
University of Victoria
Victoria, Canada

Abstract

The rapid adoption of Large Language Models (LLMs) is not only transforming software engineering (SE) practice but is also poised to fundamentally disrupt how research is conducted in the field. While perspectives on this transformation range from viewing LLMs as mere productivity tools to considering them revolutionary forces, we argue that the SE research community must proactively engage with and shape the integration of LLMs into research practices, emphasizing human agency in this transformation. As LLMs rapidly become integral to SE research—both as tools that support investigations and as subjects of study—a human-centric perspective is essential. Ensuring human oversight and interpretability is necessary for upholding scientific rigor, fostering ethical responsibility, and driving meaningful advancements in the field. Drawing from discussions at the 2nd Copenhagen Symposium on Human-Centered AI in SE, this position paper employs Marshall McLuhan's Tetrad of Media Laws to analyze the impact of LLMs on SE research. Through this theoretical lens, we examine how LLMs enhance research capabilities through accelerated ideation and automated processes, make some traditional research practices obsolete, retrieve valuable aspects of historical research approaches, and risk reversal effects when taken to extremes. Our analysis reveals opportunities for innovation and potential pitfalls that require careful consideration. We conclude with a call to action for the SE research community to proactively harness the benefits of LLMs while developing frameworks and guidelines to mitigate their risks, to ensure continued rigor and impact of research in an AI-augmented future.

Keywords

Generative AI, LLM, AI4SE, McLuhan's Tetrad



FSE Companion '25, Trondheim, Norway

ACM ISBN 979-8-4007-1276-0/2025/06
https://doi.org/10.1145/3696630.3731666

ACM Reference Format:
Bianca Trinkenreich, Fabio Calefato, Geir Hanssen, Kelly Blincoe, Marcos Kalinowski, Mauro Pezzè, Paolo Tell, and Margaret-Anne Storey. 2025. Get on the Train or be Left on the Station: Using LLMs for Software Engineering Research. In *33rd ACM International Conference on the Foundations of Software Engineering (FSE Companion '25), June 23–28, 2025, Trondheim, Norway.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3696630.3731666

1 Introduction

Integrating Large Language Models (LLMs) into Software Engineering (SE) research reflects a broader transformation across scientific disciplines. Generative AI technologies are fundamentally changing how research is conducted, from accelerating hypothesis generation to enhancing data analysis and interpretation [26]. This transformation is particularly relevant for SE research, where LLMs are becoming integral both as subjects of our investigations and as tools we use to conduct research. These models have demonstrated their potential to revolutionize research in our field by supporting various tasks, such as enhancing brainstorming processes [23], generating representative data [7, 20], aiding in data analysis and qualitative research [3], and automating repetitive or tedious tasks.

As SE researchers increasingly incorporate LLMs into their workflows, it becomes crucial to maintain a human-centric perspective, particularly when studying human aspects of SE [18]. The transformative potential of LLMs extends beyond mere automation as these tools can augment our ability to understand developer experiences, team dynamics, and socio-technical interactions in software development. However, this potential must be balanced against the need to preserve human agency and ensure that our research methods remain rigorous, transparent, and ethically sound. This is particularly important as we study how software developers adapt to and integrate LLMs into their work practices, requiring us to critically examine our own use of these tools in researching such phenomena. Therefore, understanding the broad impact of LLMs requires a comprehensive framework that evaluates their benefits and potential unintended consequences.

Marshall McLuhan's Tetrad of Media Effects [14] provides a compelling lens through which to examine the different ways a new

medium may have on augmenting human abilities across several dimensions. The Tetrad prompts us to critically assess technologies by addressing four key questions: What does the technology enhance? What does it make obsolete? What does it retrieve? And what does it reverse into when taken to extremes? As McLuhan observed, recognizing what a technology retrieves can be challenging, as it demands a deep historical understanding of its predecessors [15]. Speculating what a technology may reverse into can also be difficult, especially when a technology is novel and disruptive.

Inspired by previous applications of McLuhan's Tetrad to disruptive technologies [23], we applied the framework to speculate about the effects when we use LLMs in SE research. This speculation was conducted collaboratively by a team of 10 researchers during the 2nd Symposium on Human-Centered AI in SE. We note that other researchers may have different ideas about the effects LLMs will have on SE research and that our speculation would also change as LLMs evolve.

This position paper explores the multifaceted role of LLMs in SE research and discusses potential risks, including biases, creativity echo chambers, and a decline in essential research skills. By applying the Tetrad across the Research Pipeline Stages, this paper examines LLMs from the perspectives of enhancement, obsolescence, retrieval, and reversal for each stage of the pipeline, offering a structured and reflective analysis of their impact on SE research.

2 McLuhan's Tetrad for SE Research

We discuss the impact of LLMs on SE research through the lens of Marshall McLuhan's Tetrad of Media Laws (Enhance, Obsolesce, Retrieve, and Reverse) [14] for each stage of a generic research pipeline, summarized in Table 1. The structure of Table 1, including the research pipeline phases, was generated using a GPT associated with the Disruptive Playbook for studying the effects of technologies on SE [23], then refined and filled in by the authors. In this section, we expand on two stages of the research pipeline pertinent to Human-Centric aspects of SE research: Research Goals and Questions Formulation and Analysis and Interpretation.

2.1 Research Goals and Questions Formulation

The formulation of research goals and questions represents a foundational yet often rushed phase of the research process, where initial decisions are frequently determined prematurely without sufficient exploration of the problem space [23]. While traditionally reliant on manual literature review and brainstorming, this phase is being transformed by the capabilities of LLMs.

Enhance: LLMs can amplify rapid idea generation and accelerate the initial phase of research, as researchers can quickly explore a broad range of ideas, generate novel research hypotheses, a detailed research overview, and experimental protocols [8], facilitating more robust ideation [23]. As demonstrated in a recent AI-augmented Brainwriting study [19], LLMs contribute to the divergence stage of ideation by introducing novel perspectives and generating diverse research angles that researchers might not have considered independently. This fosters creativity and mitigates common brainstorming barriers such as fixation and cognitive inertia.

Recent LLM-based contributions, such as the Disruptive Research Playbook [23], which helps to formulate socially relevant research questions to challenge assumptions and refine focus, and the AI Co-Scientist [8], which supports generating, debating, and refining hypotheses through a multi-agent system that allows iteratively improving research directions, point toward a future where LLM-based solutions evolve research goal and question formulation to accelerate impactful scientific discovery.

In a broader perspective, LLMs can also enable researchers to define more ambitious research goals that call for mixed-method research [22]. LLMs can assist in parts of the research process where the researcher lacks experience or resources or when the total research design becomes complex. E.g., LLMs may advise the researcher in avoiding known anti-patterns in mixed method research such as 'Sample contamination' and 'Integration failure' [22].

Obsolece: LLMs can reduce the manual effort in systematic literature reviews (SLRs), automating several time-consuming tasks. Recent research found that ChatGPT can support study selection and also help improve search string formulation by suggesting synonyms and relevant terms for more effective Boolean queries [6]. LLMs are already being used in other domains, such as medicine, to help extract [10] and synthesize data [11], helping researchers identify key concepts and themes in large volumes of literature. LLMs also support inclusion/exclusion decisions, reducing manual effort while maintaining consistency. Human oversight remains critical despite these advantages, as LLMs can sometimes provide persuasive but inaccurate information. By using LLMs for preliminary automation and human validation of critical decisions, researchers can streamline the SLR process, reduce cognitive load, and focus on higher-level analysis while ensuring accuracy and reliability.

Retrieve: LLMs can revive the culture of informal ideation and intellectual discourse in research, previously known as "coffee house" research [5]. Through natural language interaction with LLMs and rapid information synthesis, these tools facilitate the rapid exploration and prototyping of research ideas before formalizing them into papers, which may take time to generate feedback. Furthermore, by reducing much of the tedious work involved in research, LLMs can allow researchers to reallocate that time to engage in more intellectual discourse, fostering deeper engagement in small discussion-based workshops. Potentially enhanced by AI-driven insights, these conversations can lead to a more thorough exploration of fundamental research questions, helping researchers refine their goals, develop innovative ideas, and uncover novel interdisciplinary connections. [1]

LLMs can also switch trending research topics. For example, since LLMs are used to automate more aspects of code generation and validation, we may see a resurgence of research interest in formal specification and verification research [16] to ensure that the outputs of LLMs are correct. We expect to also see an increased research focus on parts of the software development cycle where humans must remain in the loop. For example, requirements engineering and research to improve practices around understanding user needs may see a resurgence as code generation becomes more fully automated. Research can investigate how LLMs can help to manage natural language subjectivity to transform unstructured data into structured requirements, supporting automated elicitation, inconsistency detection, and traceability.

[1] It should be noted that historical coffee house research was not always inclusive. Women, for example, often did not attend [5]. We envision a modern and more inclusive version of coffee house research being retrieved.

Research Pipeline Stage	Enhance What does it amplify?	Obsolesce What does it push aside?	Retrieve What does it bring back?	Reverse What happens when pushed to extremes?
Research Goals & Questions Formulation (*)	Rapid idea generation (*), auto-suggested hypotheses, literature summarization automation	Manual literature review (*), brainstorming without AI assistance	"Coffee house research" (*), switch trending research topics (*), sketchbook of ideas	Creativity echo chamber (*), homogenized research questions, potential loss of novelty from AI ideas
Experimental Design & Methodology	Automated experiment setup, code synthesis for study prototypes, reproducibility improvements	Tedious manual setup, reliance on domain experts for experiment structuring	Human "intractable" models of research field, modular and reusable experimental designs	Over-reliance on AI-generated methodologies may lead to reduced critical evaluation
Data Collection	Faster extraction from repositories (GitHub, Stack Overflow), automated data cleaning	Human-driven data curation, traditional data wrangling techniques	Historical datasets revisited for new insights	Bias amplification in datasets, lack of transparency in synthetic data creation
Data Processing	Improved statistical modeling via AI, anomaly identification	Pushing aside the risk of human errors	Large-scale or longitudinal ethnographic studies	Errors and biases that humans cannot easily detect
Analysis & Interpretation (*)	Qualitative, quantitative, and mixed-methods analysis (*), diverse viewpoints (*)	Manual coding (qualitative and quantitative) (*), manual selection and execution of statistical techniques (*)	Finding related theories in other domains (*), holistic and interdisciplinary analysis (*)	AI hallucinations (*), loss of human's role in theories' construction, misleading interpretations if results are blindly trusted
Writing & Dissemination	Automated paper drafting, AI-assisted summaries, multilingual dissemination	Manual academic writing, sole reliance on human synthesis	Collaborative, rapid prototyping of research papers	Proliferation of low-quality or AI-generated papers, diminishing originality and rigor
Cross-cutting Impacts (*)	Research speed and creativity (*)	Manual/tedious research tasks (*)	Impactful research (*)	Lower skills of researchers (*)

Table 1: Applying McLuhan's Tetrad to LLMs across the Software Engineering Research Pipeline. (*) are discussed in this paper.

Reverse: While LLMs can accelerate various research tasks, their use in formulating research questions, defining study goals, and structuring investigations raises concerns about creativity stagnation. Ideally, research questions should be driven by intellectual curiosity, domain expertise, and an ability to challenge conventional wisdom—qualities that LLMs, by design, lack. Since LLMs today tend to generate content based on existing knowledge, relying too heavily on them for research ideation risks creating a CREATIVITY ECHO CHAMBER, where generated questions and study designs reflect common patterns rather than genuinely novel insights (that is the opposite of the benefit of using LLMs for creative research directions). This effect is particularly concerning given that modern academic incentives often emphasize rapid publication over deeply impactful contributions. If LLMs reinforce established knowledge structures, researchers may unknowingly converge on "safe" and predictable topics, leading to a homogenization of research rather than groundbreaking discoveries.

2.2 Analysis & Interpretation

The analysis and interpretation of SE data present unique challenges when studying human and social aspects, requiring researchers to make sense of complex, qualitative, and often interrelated findings from multiple sources [2]. LLMs are now transforming how researchers approach this intricate analytical process.

Enhance: By processing vast amounts of data, LLMs can enhance QUANTITATIVE, QUALITATIVE, AND MIXED-METHODS ANALYSIS by identifying trends, anomalies, correlations, and patterns that might be missed in manual analysis. In SE research, this applies to quantitative data (*e.g.*, controlled study measurements, test results, defect reports, commit logs) and qualitative data (*e.g.*, interviews, meeting transcripts). LLMs can assist in applying quantitative and qualitative research methods to analyze such data. In quantitative research, they can be used to analyze structured datasets to extract key patterns and relationships. In qualitative research, they have been applied to identify sentiment [27], themes[4, 13], and to support the application of a variety of qualitative research methods [3]. By enhancing efficiency and scalability, LLMs can reduce the time required for data analysis, alleviating manual effort in both quantitative and qualitative research. They can also improve consistency in calculations and coding while enhancing generalizability by enabling pattern identification across larger datasets and broader contexts. Furthermore, LLMs can facilitate mixed-methods research by integrating the visualization of qualitative and quantitative findings—a traditionally challenging task [9].

LLMs can integrate DIVERSE VIEWPOINTS from a wide range of stakeholders –often overlooked in traditional research– by processing large volumes of data, including user feedback, developer discussions, and practitioner reports from forums, social media, and documentation repositories. This capability enables researchers to surface varied perspectives, identify emerging trends, and capture insights that span technical, social, and organizational contexts. By synthesizing this diverse input, LLMs can highlight patterns and conflicting opinions, and find emerging trends that enrich qualitative analysis with voices from end-users and practitioners.

Obsolesce: LLMs are increasingly pushing MANUAL CODING aside in both qualitative and quantitative analyses. LLMs significantly reduce the time and effort required for manual annotation [1], which, when done with appropriate care to check semantic aspects and consistency, can improve efficiency in SE research.

In qualitative studies, LLMs are already being used to code interview transcripts, documents, surveys, interviews, issue-tracker comments, and software reviews [2]. In quantitative studies, LLMs can process large amounts of repository data to automatically classify code changes (e.g., bug fixes, refactorings, feature additions), extract and structure performance metrics, API usage patterns, and technical debt indicators from repositories, reducing the need for manual intervention. Additionally, by offering guidance on test selection, assumption checking, and result interpretation, LLMs

can aid in MANUALLY SELECTING AND EXECUTING STATISTICAL TECHNIQUES—tasks that traditionally demanded significant expertise.

Retrieve: LLMs can revive interest in FINDING RELATED THEORIES IN OTHER DOMAINS to interpret SE findings (something we have not been doing enough of in recent years [12]). By processing vast interdisciplinary literature, LLMs can foster information-seeking practices and facilitate analysis for interdisciplinary research [28], connecting SE challenges to established frameworks in fields such as cognitive psychology and organizational science, which could otherwise remain unexplored [26]. This retrieval of cross-domain knowledge extends beyond simple analogies, enabling researchers to recontextualize technical findings within broader theories, identify disciplinary intersections, and explore new research directions.

LLMs can bring back more HOLISTIC, INTERDISCIPLINARY STUDIES by uncovering historical patterns and forgotten theories from different domains to help SE researchers contextualize new qualitative and quantitative insights. By analyzing past literature, LLMs can potentially trace how SE theories and ideas developed over time and why some approaches became more popular [21]. Tshitoyan et al. [24] showed we can use AI to extract hidden patterns from scientific publications that predict future discoveries. Through this historical lens, researchers can gain a deeper understanding of ongoing challenges in the SE field and avoid reinventing the wheel.

Reverse: Interdisciplinary research requires caution, as overreliance on LLMs without domain expertise can lead to misinterpretation and reduced research rigor. Excessive reliance on LLMs can also cause AI HALLUCINATIONS —fabricated or inaccurate outputs— leading to misleading interpretations if results are blindly trusted. Driven by probabilistic patterns, LLMs may generate plausible but incorrect conclusions, misattribute sources, or identify false patterns from noisy data, distorting findings, and compromising research validity. One example is about using LLMs as annotators. While the model-to-model agreement can predict when LLMs could safely replace human annotators, it also highlights the potential for systemic bias, where LLMs reinforce each other's errors, creating a false perception of reliability [1]. Hence, while LLMs can support analysis and interpretation, human oversight is required. It is important to keep in mind that LLMs currently cannot independently assess the validity of an argument, and critical thinking remains a human responsibility.

3 Call to Action

Using LLMs for SE research presents a significant opportunity to speed up discovery and unlock new avenues for impactful research. Although caution is necessary to ensure that model biases, reliability, and ethical considerations, among other risks, are addressed, leaning into LLMs can streamline the research pipeline. Across all research stages (see cross-cutting impacts in Table 1), we see that using LLMs can ENHANCE RESEARCH SPEED AND CREATIVITY and REDUCE MANUAL AND TEDIOUS RESEARCH TASKS. Thus, by strategically integrating LLMs, researchers can conduct more efficient, data-driven investigations, enabling faster insights into complex SE phenomena. However, OVERRELIANCE ON LLMS CAN RESULT IN LOWER SKILLS OF RESEARCHERS. We must ensure that LLMs are used for research in a human-centric perspective to augment, and not replace, researchers. Balancing the accelerated pace enabled by LLMs with methodological rigor involving human oversight and critical thinking will ensure the research remains valid and transformative.

We also believe that the accelerated pace of research enabled by LLMs presents a pivotal decision for our research community. Although LLMs can be used to accelerate paper production, their true transformative potential lies in enabling researchers to redirect their time toward deeper intellectual engagement and FOCUS ON DOING IMPACTFUL RESEARCH. By freeing researchers from time-consuming tasks such as searching for the literature and initial data analysis, LLMs create space for more meaningful collaborative discussions about the impact of research and its social implications.

We urge our fellow members of the software engineering research community to use any time saved by LLMs to invest in thoughtful dialogue with colleagues, stakeholders, and potential beneficiaries of their work. Such conversations can help identify pressing research questions and ensure that research addresses genuine societal needs rather than merely contributing to academic metrics. This shift from quantity to quality of research output could lead to more impactful and purposeful scientific contributions that better serve both the academic community and society at large.

The SE research community must take proactive steps to harness the LLMs' benefits while mitigating their risks. As these models become increasingly integrated into research workflows, structured guidelines, evaluation processes, and educational initiatives are essential to maintaining scientific rigor and reproducibility. Without clear guidance, there is a danger of overreliance, hidden biases, loss of the human's role in constructing theory, and compromised methodological integrity. To address these concerns, we propose four key actions: (1) experimenting with LLMs in SE research to build concrete experiences, (2) developing guidelines for transparent reporting and reviewing [17],(3) establishing benchmarks for evaluating LLM-generated research artifacts, and (4) creating educational resources to train researchers in responsible LLM usage.

(1) We encourage curiosity and wide experimentation of LLMs in SE research, working with real cases and real data. We may see various effects on SE research as a practice across the research pipeline, but some of these are so far based on (qualified) assumptions, grounded on early and fragmented experience. We need to gain and share more experience to contrast and detail the issues we have started identifying and - most likely - identify new ones.

(2) We advocate for clear and transparent reporting on using LLMs in research, in combination with clear review guidelines about the use of LLMs in SE research. Every study incorporating LLMs should explicitly document how these models influence study design, data collection, and analysis. This includes specifying the LLM model and version used, the exact prompting strategies employed (including any sensitivity analyses), and the mechanisms for human oversight. Without standardized reporting, it becomes hard to assess the validity of findings, compare results across studies, or detect systematic biases introduced by AI-generated outputs. Establishing clear disclosure standards will enhance the interpretability, reproducibility, and credibility of research involving LLMs [25].

(3) To address the high variability in LLM performance, we propose establishing and maturing benchmarks to evaluate the quality and reliability of LLM-generated research artifacts. This initiative should include a publicly available dataset covering a range of SE research tasks, such as annotation, summarization, and causal

inference, to systematically assess where and how LLMs can be reliably used. Additionally, validation protocols should be established, leveraging metrics such as inter-rater agreement scores (e.g., Krippendorff's α) and consistency checks across different prompts and models [1]. By creating a standardized evaluation process, the research community can ensure that LLMs are deployed only in contexts where their reliability has been empirically demonstrated, reducing the risk of misleading or flawed research conclusions.

(4) We highlight the urgent need for educational resources to guide researchers in the responsible use of LLMs. Without proper training, early-career researchers risk losing foundational skills like critical thinking and hands-on analysis. To prevent this, we recommend instruction in prompt engineering, bias mitigation, and hybrid human-LLM workflows, supported by workshops and case studies. As the community begins to understand the challenges of integrating LLMs into SE research, efforts are underway to develop best practices and guidelines[2].

By equipping researchers with LLM literacy, strong methodological foundations, structured reporting, robust validation frameworks, and targeted education, the SE research community can responsibly integrate LLMs as an aid rather than a substitute for critical inquiry, preserving essential research skills while safeguarding the rigor and reliability of empirical studies. While this is a position paper, future work could build on our discussion by grounding it in specific SE areas—such as program analysis or repair and software traceability—and how LLMs can change the review process, a concern across the entire community.

4 Acknowledgments

We thank Thomas Zimmermann for his essential contributions to this study. Our sincere appreciation also goes to the Alfred P. Sloan Foundation and the Carlsberg Foundation for the 2nd Copenhagen Symposium on Human-Centered SE and AI (G-2024-22586 and CF24-0693 to Daniel Russo), held at Aalborg University on Nov2024, the National Science and Research Council of Canada (NSERC) RGPIN-2025-6813, the Rutherford Discovery Fellowship administered by the Royal Society Te Apārangi, the DARE project (PNC0000002, CUP B53C22006420001), and the QualAI project (2022B3BP5S, CUP H53D23003510006).

References

[1] Toufique Ahmed, Premkumar Devanbu, Christoph Treude, and Michael Pradel. 2025. Can LLMs Replace Manual Annotation of Software Engineering Artifacts?. In *2025 IEEE/ACM 22nd Int'l. Conf. on Mining Software Repositories (MSR)*.

[2] Muneera Bano, Rashina Hoda, Didar Zowghi, and Christoph Treude. 2024. Large language models for qualitative research in software engineering: exploring opportunities and challenges. *Automated Software Engineering* 31, 1 (2024), 8.

[3] Cauã Barros, Bruna Azevedo, Valdemar Neto, Mohamad Kassab, Marcos Kalinowski, Hugo Nascimento, and Michelle Bandeira. 2025. Large Language Model for Qualitative Research - A Systematic Mapping Study. In *Workshop on Methodological Issues with Empirical Studies in Software Engineering (WSESE@ICSE'25)*.

[4] Stefano De Paoli. 2024. Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Soc Sci Comput Rev* 42, 4 (2024), 997–1019.

[5] Markman Ellis. 2008. An introduction to the coffee-house: A discursive model. *Language & Communication* 28, 2 (2008), 156–164.

[6] Katia Romero Felizardo, Márcia Sampaio Lima, Anderson Deizepe, Tayana Uchôa Conte, and Igor Steinmacher. 2024. ChatGPT application in Systematic Literature Reviews in Software Engineering: an evaluation of its accuracy to support the selection activity. In *Empirical Software Engineering and Measurement*. 25–36.

[2]Guidelines for using LLMs in SE Research - https://llm-guidelines.org/

[7] Marco Gerosa, Bianca Trinkenreich, Igor Steinmacher, and Anita Sarma. 2024. Can AI serve as a substitute for human subjects in software engineering research? *Automated Software Engineering* 31, 1 (2024), 13.

[8] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. 2025. Towards an AI co-scientist. https://storage.googleapis.com/coscientist_paper/ai_coscientist.pdf.

[9] Timothy C Guetterman, Michael D Fetters, and John W Creswell. 2015. Integrating quantitative and qualitative results in health science mixed methods research through joint displays. *The Annals of Family Medicine* 13, 6 (2015), 554–561.

[10] Qusai Khraisha, Sophie Put, Johanna Kappenberg, Azza Warraitch, and Kristin Hadfield. 2024. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods* (2024).

[11] Xuân-Lan Lam Hoai and Thierry Simonart. 2023. Comparing meta-analyses with ChatGPT in the evaluation of the effectiveness and tolerance of systemic therapies in moderate-to-severe plaque psoriasis. *J Clin Med* 12, 16 (2023), 5410.

[12] Tobias Lorey, Paul Ralph, and Michael Felderer. 2022. Social science theories in software engineering research. In *Int'l Conference on Software Engineering*. 1994–2005.

[13] Walter S Mathis, Sophia Zhao, Nicholas Pratt, Jeremy Weleff, and Stefano De Paoli. 2024. Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? *Computer Methods and Programs in Biomedicine* 255 (2024), 108356.

[14] Marshall McLuhan. 1977. Laws of the Media. *ETC: A Review of General Semantics* (1977), 173–179.

[15] Marshall McLuhan. 2017. The medium is the message. In *Commun Theory*. Routledge, 390–402.

[16] Bertrand Meyer. 2023. AI Does Not Help Programmers. https://cacm.acm.org/blogcacm/ai-does-not-help-programmers/

[17] Paul Ralph, Rashina Hoda, and Christoph Treude. 2020. ACM SIGSOFT empirical standards. (2020).

[18] Daniel Russo, Sebastian Baltes, Niels van Berkel, Paris Avgeriou, Fabio Calefato, Beatriz Cabrero-Daniel, Gemma Catolino, Jürgen Cito, Neil Ernst, Thomas Fritz, et al. 2024. Generative ai in software engineering must be human-centered: The copenhagen manifesto. *J. Syst. Softw.* 216 (2024), 112115.

[19] Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew Kun, and Hagit Ben Shoshan. 2024. AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. In *CHI Conf. on Human Factors in Computing Systems*. 1–17.

[20] Igor Steinmacher, Jacob Mcauley Penney, Katia Romero Felizardo, Alessandro F Garcia, and Marco A Gerosa. 2024. Can ChatGPT emulate humans in software engineering surveys?. In *Proc. of the 18th ACM/IEEE Int'l. Symposium on Empirical Software Engineering and Measurement*. 414–419.

[21] Klaas-Jan Stol. 2024. Teaching Theorizing in Software Engineering Research. arXiv:2406.17174 [cs.SE] https://arxiv.org/abs/2406.17174

[22] Margaret-Anne Storey, Rashina Hoda, Alessandra Maciel Paz Milani, and Maria Teresa Baldassarre. 2025. Guiding Principles for Using Mixed Methods Research in Software Engineering. http://arxiv.org/abs/2404.06011

[23] Margaret-Anne Storey, Daniel Russo, Nicole Novielli, Takashi Kobayashi, and Dong Wang. 2024. A disruptive research playbook for studying disruptive innovations. *ACM TOSEM* 33, 8 (2024), 1–29.

[24] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nat.* 571, 7763 (2019), 95–98. doi:10.1038/S41586-019-1335-8

[25] Stefan Wagner, Marvin Muñoz Barón, Davide Falessi, and Sebastian Baltes. 2025. Towards Evaluation Guidelines for Empirical Studies involving LLMs. arXiv:2411.07668 [cs.SE] https://arxiv.org/abs/2411.07668

[26] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Linfeng Zhang, Connor Coley, Yoshua Bengio, and Marinka Zitnik. 2023. Scientific discovery in the age of artificial intelligence. *Nature* 620, 7972 (Aug. 2023), 47–60. doi:10.1038/s41586-023-06221-2

[27] Ting Zhang, Ivana Clairine Irsan, Ferdian Thung, and David Lo. 2024. Revisiting Sentiment Analysis for Software Engineering in the Era of Large Language Models. *ACM Trans. Softw. Eng. Methodol.* (Sept. 2024). doi:10.1145/3697009

[28] 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 *ACM Symposium on User Interface Software and Technology*. 1–20.