5-2024

Large language models for qualitative research in software engineering: exploring opportunities and challenges

Muneera BANO

Rashina HODA

Didar ZOWGHI

Christoph TREUDE
Singapore Management University, ctreude@smu.edu.sg

Automated Software Engineering (2024) 31:8
https://doi.org/10.1007/s10515-023-00407-8

Large language models for qualitative research in software engineering: exploring opportunities and challenges

Muneera Bano[1] · Rashina Hoda[2] · Didar Zowghi[1] · Christoph Treude[3]

Received: 15 November 2023 / Accepted: 26 November 2023 / Published online: 21 December 2023

Abstract
The recent surge in the integration of Large Language Models (LLMs) like ChatGPT into qualitative research in software engineering, much like in other professional domains, demands a closer inspection. This vision paper seeks to explore the opportunities of using LLMs in qualitative research to address many of its legacy challenges as well as potential new concerns and pitfalls arising from the use of LLMs. We share our vision for the evolving role of the qualitative researcher in the age of LLMs and contemplate how they may utilize LLMs at various stages of their research experience.

Keywords Large language models · LLMs · Qualitative research · Software engineering

1 Introduction

The advent of Large Language Models (LLMs) such as OpenAI's ChatGPT and Google's Bard has been nothing short of a paradigm shift in academia, much like in many other professions. Within a year of their inception, these models have become a focal point of academic scrutiny, with researchers exploring their potential across a plethora of domains (Bano et al. 2023). From analyzing its pivotal role in research and academia to understanding its transformative potential in educational settings, the emerging body of literature paints an intriguing picture of the far-reaching implications of LLMs.

✉ Muneera Bano
muneera.bano@csiro.au

[1] CSIRO's Data61, Clayton, Australia

[2] Monash University, Melbourne, Australia

[3] University of Melbourne, Melbourne, Australia

We found an increased interest from researchers in the use of LLMs in Software Engineering (SE), and the work is characterized by a multifaceted evaluation of potential benefits and inherent challenges. Ozkaya (2023) projects an AI-augmented software development lifecycle, with AI assistants contributing to various SE tasks like specification generation and legacy code translation. Concurrently, Jalil et al. (2023) analyze ChatGPT's role in software testing education, revealing its potential and the risks of overreliance due to varying response accuracy. Ebert and Louridas (2023) discuss how generative AI can automate SE tasks, urging a balanced integration considering ethical and privacy concerns. Scoccia (2023) gathers early adopter experiences with ChatGPT's code generation, indicating its significant impact yet mixed usage outcomes. The empirical study by Kuhail et al. (2023) presents a nuanced perspective on AI's role in SE, suggesting increased trust with frequent use but also heightened job security concerns. Arora et al. (2023) propose a SWOT analysis for LLMs in Requirements Engineering, suggesting a cautiously optimistic view towards AI's role in elicitation and validation. Nguyen-Duc et al. (2023) outline a research agenda for Generative AI, emphasizing its potential for partial automation in SE tasks. Finally, Hou et al. (2023) provide a systematic literature review of LLMs in SE offering a comprehensive roadmap for future research and practical applications in SE. Each of these contributions underscores the transformative potential of LLMs in SE, while also acknowledging the complexity of their integration and the need for ongoing research to navigate the challenges they present.

However, augmenting research processes with LLMs is yet to be fully investigated. In the evolving landscape of SE research, the intertwining of technological advancements with human-driven insights has been a constant. While no one can claim they saw the exact nature and shape of LLMs emerging, predictions have been made about "further advancements in technology and artificial intelligence" offering "unexplored potential in supplementing, augmenting, and automating parts of qualitative data analysis to ease human effort and improve both the quality and scale of theory development" (Hoda 2021).

The work of Byun et al. (2023) shows that LLMs like GPT-3 have the capacity to produce text that is comparable to that written by humans, even in qualitative analysis, which traditionally relies heavily on human insight. Their work demonstrates that AI can not only generate text but also identify themes and provide detailed analysis similar to that of human researchers. They suggest that AI could potentially match human capabilities in interpreting qualitative data. Their findings indicate a promising avenue for using AI in qualitative research, where it could serve as a tool to both augment and potentially replace human analysis, raising important questions about the future role of humans in research processes. Bano et al. (2023) challenge these claims. They acknowledge the potential of AI to align with human analysis in some cases but caution against an overreliance on AI due to significant disparities between AI and human reasoning. Their study reveals that while AI, specifically LLMs like ChatGPT 3.5 and GPT-4, can sometimes provide logical classifications, there is often a lack of consensus between AI-generated and human-generated insights, raising questions about the AI's capability to fully grasp the complexities of human language and the contextual nuances important in qualitative research.

Despite preliminary progress, there remains a significant lack of clarity on how LLMs compare to human intelligence in qualitative research (Bender et al. 2021). The role of LLMs in SE qualitative research presents both unprecedented opportunities and inherent risks, especially when viewed from the perspective of researchers at different stages of their academic journey.

2 Addressing legacy challenges of SE qualitative research with LLMs

Historically, qualitative research in SE has grappled with challenges like the time-intensive nature of the work, limitations to scalability due to its manual nature, and the inherent subjectivity that qualitative methodologies can sometimes entail.

2.1 Time-intensive work

Conducting qualitative research often requires intensive data analysis, which can be time-consuming. LLMs can help to automate or expedite parts of these processes. For example, they could help to make sense of large amounts of textual data, identify themes and patterns within data, and generate initial codes or categories. Such technical assistance could significantly speed up the data analysis process and allow researchers to handle larger datasets, thereby allowing them to scale qualitative analysis in ways hardly possible through a commensurate amount of manual effort.

2.2 Generalizability

Qualitative research is hard to generalize universally or to wider populations outside the originally studied context, which is typically a relatively narrow phenomenon. Based on the constructive worldview, it may even be undesirable. However, the use of AI-based models and advanced natural language processing, such as those offered by LLMs, can help improve the relevance and generalizability of the qualitative findings, such as descriptive findings, taxonomies, and theories by expanding the contexts studied (Hoda 2021).

2.3 Consistency

Variations in qualitative data analysis are expected to exist across different researchers, but consistency can still be an issue for individual researchers. Depending on several factors not limited to external and personal circumstances, achieving high levels of human consistency is a known challenge (Gentles et al. 2015; Watson 2006). On the other hand, LLMs, being computing entities, can process and analyze data in a consistent way, considering the consistency in prompts. Improved consistency is likely to lend itself to better repeatability of the process and higher

reproducibility of the research outcomes. This may be particularly desirable from a positive perspective.

2.4 Subjectivity

While it may be impossible or even undesirable to eliminate human subjectivity from qualitative research, LLMs could potentially add an additional layer to the analysis. For example, the use of LLMs can help a team of qualitative researchers discuss and agree on the concepts emerging from their individual analyses. Furthermore, the concepts generated by LLMs can act as a 'third party' reference to help address and reconcile differences emerging from personal beliefs, experiences, or emotions. It seems early and somewhat naive to suggest that an LLM can act as an objective baseline or a source of a deciding 'expert opinion'. LLMs, like humans, are known to harbor their own set of biases based on the training data and parameters that can influence their inference logic when it comes to qualitative research (Navigli et al. 2023). With rapid enhancements in LLM capabilities, these aspects can be reexamined in the future.

3 New frontiers, new challenges

While LLMs may seem to be the panacea for many traditional qualitative research issues, they bring with them a set of unique challenges. We summarize these below.

3.1 Ethical and privacy concerns

Incorporating LLMs into data analysis poses ethical and privacy challenges, especially with sensitive data. Ethical issues include ensuring data consent, proper anonymization to enable de-identification, and addressing biases that AI may perpetuate (Arora et al. 2023; Ebert and Louridas 2023). These concerns necessitate a responsible AI framework that respects individual privacy and data rights. For ethical usage, Nguyen-Duc et al. (2023) recommend integrating AI with an awareness of ethical implications and privacy risks, such as by using AI to enhance rather than replace human decision-making, and keeping sensitive raw data local to avoid exposure. Ozkaya (2023) further suggests robust data governance to ensure AI applications adhere to ethical standards and privacy regulations, balancing AI's potential with necessary oversight.

3.2 Model biases

Like all machine learning models, LLMs can have inherent biases based on the data they were trained on, which can be flawed or insufficient.[1] This could potentially

[1] https://www.csiro.au/en/news/all/articles/2023/june/humans-and-ai-hallucinate.

skew the analysis or conclusions drawn from their use in qualitative research. For example, in SE qualitative research, if an LLM is trained on data that predominantly consists of contributions from male developers, it may inadvertently downplay or overlook the communication styles, coding preferences, or problem-solving approaches more common among female developers or those from underrepresented groups. In such cases, researchers have the responsibility to be aware of and acknowledge the inherent biases in the underlying data on which the LLMs are trained, as part of the limitations of their research.

3.3 Lack of contextual and philosophical understanding

While LLMs can process and generate text based on patterns learned, they lack a true understanding of the context, which is crucial in qualitative research. This could lead to oversights and misinterpretations. For example, in a SE qualitative study analyzing developer communication on issue trackers, an LLM might interpret technical jargon or project-specific slang literally, missing the nuanced meaning intended by the developers. While LLMs could identify and summarize discussions on a given research topic from various sources, articles, and grey literature, but they might not fully grasp the subtleties of concerns that require a deeper philosophical understanding and contextual awareness, which human researchers provide. In such cases, the researchers should be paying special attention to any missing or misinterpreted contexts.

3.4 Dependency on technology

There is a risk of becoming overly dependent on technology for research. While LLMs can assist in data analysis, they should not replace the human element of research, which includes critical thinking, contextual understanding, and ethical judgment (Bano et al. 2023). To educate and train the next generation of qualitative researchers it is important to not overly rely on augmented research technologies such as LLMs. We elaborate further on the level of expertise of researchers later in this paper.

3.5 Quality control

Ensuring the quality and accuracy of the results generated by LLMs can be challenging. Researchers need to be vigilant and critical when interpreting the outputs of LLMs. For example, ChatGPT is known to be prone to *hallucinations*, instances where LLMs generate inaccurate or entirely fabricated information. Not checking for inaccurate and fake information generated by LLMs can land researchers in trouble.[2] To address the issue of hallucinations the involvement of human researchers is imperative. As pointed out by Rudolph et al. (2023) and Alkaissi and McFarlane

[2] https://www.cyberdaily.au/digital-transformation/9779-researchers-apologies-to-big-4-consultancy-firms-for-false-ai-based-accusations.

(2023), these hallucinations can lead to misinterpretation of research outcomes, compromise the validity of results, and introduce bias or error. To counteract this, researchers must scrutinize, verify, and interpret the outputs of LLMs meticulously, ensuring that the conclusions are aligned with the actual context and maintain the integrity of the research. This human intervention is necessary not only for validation but also to continually refine and calibrate the models, thereby improving their understanding and minimizing potential drawbacks (Watkins 2023).

3.6 Reproducibility

As LLMs are continuously updated, and old models are deprecated, the ability to reproduce an analysis with the same precision diminishes over time, a phenomenon known as *model drift*.[3] Researchers may provide exhaustive details on their methodology, including data sets, prompts, parameters, and the versions of models used, but this does not guarantee that the same analysis can be reproduced in the future by LLMs. Unlike human researchers, where insights and analytical reasoning can be revisited or clarified, LLMs do not offer the possibility to revisit the reasoning behind their outputs once the model version is no longer available.

3.7 Context of related work

Integrating an LLM's data analysis within the broader context of related work poses a significant challenge, primarily because the model cannot access the entirety of potentially relevant literature due to constraints on data availability and access rights due to paywalls. This limitation hampers the LLM's ability to draw comprehensive connections and insights that are informed by the existing research, potentially narrowing the scope and depth of its analytical outputs. In the future, if LLMs are capable of handling large quantities of raw data from literature along with the context of related work, this could lead to augmenting systematic literature reviews (Kitchenham 2004) with LLMs.

3.8 Critical thinking

Developing critical thinking in LLMs is a complex challenge, as it involves the model's exposure to a variety of data, including incorrect statements, to enhance its evaluative capabilities (Emmert-Streib 2023). To ensure LLMs are exposed to such a range of data, researchers could deliberately include datasets with known errors or contradictory information during the training phase. This method could potentially help LLMs learn to discern and evaluate the accuracy of information they analyze. However, this approach also raises concerns about how to effectively teach LLMs to recognize and appropriately handle incorrect information without perpetuating or

[3] https://c3.ai/glossary/data-science/model-drift/.

amplifying these errors in their outputs. Currently, it's unclear how critical thinking might be incorporated in LLMs when analysing qualitative data.

3.9 Intellectual property (IP)

IP concerns are another dimension to consider in the use of LLMs in research. The contribution of LLMs' responses and analyses to the creation of a research output could raise questions about authorship, such as whether ChatGPT should be credited as a co-author, reflecting the model's role in data processing and knowledge generation (Balel 2023). Another layer of complication is the copyright and IP of the data on which LLMs are trained on. Determining the extent of LLMs' contribution, and that of underlying sources, and its implications for IP rights and academic recognition is an ongoing debate in the research community (Polonsky and Rotman 2023).

4 The evolving role of human researcher

Amid the LLM revolution, the role of the human researcher is undergoing a nuanced shift.

4.1 Ensuring ethical practices

Researchers must ensure that their studies are conducted ethically. This includes obtaining informed consent from participants, ensuring privacy and confidentiality, and treating the data in a way that respects the rights and dignity of the participants/sources (Watkins 2023).

4.2 Prompt engineering

Prompt engineering is emerging as a crucial skill, underscoring the fact that the quality of LLM outputs hinges significantly on the inputs it receives. It's important to note that prompt engineering can also be a stage where researchers might unintentionally introduce bias, as the way questions are framed can influence the direction and nature of the LLM's response, potentially reinforcing certain perspectives or excluding others.

4.3 Defining research questions

Although LLMs can be used to brainstorm research topics and ideas, the researcher must define the research questions and objectives. An LLM can help process data, but LLMs do not have intellectual curiosity, intention, motivation, or enough information to set research directions, which will depend on the researcher.

4.4 Data collection

While LLMs can help process and analyze large amounts of data, and now, with web searchability can collect data as well, it is still the researcher's responsibility to collect the data in certain qualitative research contexts such as interviews or surveys. However, in some instances where it is extremely difficult to recruit real participants for research, e.g. in health domain patients with chronic ailments, LLMs can be used to simulate and role-play certain personas for data collection. The known limitations of using personas in research, as well as the lack of lived human experience in simulated data, will continue to be a challenge.

4.5 Interpreting outputs of the LLM

An LLM can proficiently identify patterns and themes within a dataset, presenting a synthesized analysis. However, it remains the domain of a human researcher to ascribe meaning to these findings, contextualizing them within the framework of the research objectives. One might wonder why the task of interpretation cannot also be delegated to an LLM. The reason lies in the nuanced understanding and subjective judgment required—qualities that are distinctly human and currently beyond the ability of LLMs. Additionally, while it is possible for one LLM to analyze another LLM's output (Jiang et al. 2023), this still does not replace the depth of insight and complex reasoning a human brings to the interpretation of research data.

4.6 Quality checking

It is important for researchers to check the quality of the work done by the LLM. For instance, they need to look for biases in the analysis and ensure that the LLM is correctly interpreting and coding the data.

4.7 Theorizing

Developing rich theories that are grounded in evidence requires a deep understanding of the data, the ability to see connections and patterns, and the creativity to formulate a theory. These are all skills that are currently beyond the reach of LLMs.

4.8 Writing and dissemination

Finally, the researcher is responsible for writing up the results of the study and disseminating them, and is generally accountable for the research and its results. For example, the Journal of Information and Software Technology allows the use of Generative AI for improving readability and language, provided that authors have

to give explicit acknowledgment statements for the accountability of their produced work.[4]

This includes presenting the findings in a way that is understandable and useful to others and publishing or sharing the results in relevant forums.

5 The promise of LLMs across varied research expertise

Qualitative research is often rooted in a constructivist paradigm emphasizing the non-replicable human capacity to understand and contextualize social phenomena (Easterbrook et al. 2008; Hoda 2021). The constructivist paradigm in SE research is concerned with socio-technical realities that are not objective but constructed through human experiences and contexts. This paradigm values the researcher's role in interpreting data, where their involvement and perspective are considered integral to the analysis, especially in methods like ethnography, participant observation, and grounded theory.

Qualitative research in SE also offers unique advantages in exploring complex socio-technical processes and aiding in theory construction. It can reveal underlying reasons behind intricate socio-technical dynamics and is often used to generate new research questions and insights. These aspects underscore the necessity of the human element in data interpretation, despite the analytical capabilities of LLMs.

The expertise of a researcher is crucial across all research modalities, including the application of LLMs, as it guides the critical interpretation of data, the strategic questioning that leads to deeper insights, and the contextual understanding that LLMs alone cannot provide.

Further to the opportunities and challenges presented by LLMs in SE qualitative research discussed above, we present our collective thoughts on how these may vary by the experience level of the researchers. Firstly, and most importantly, with the introduction of LLMs, ethical considerations come to the fore. It is crucial for researchers at all stages to understand and uphold ethical practices, especially concerning data privacy, possible plagiarism, and potential biases that the LLMs might introduce or perpetuate (Treude and Hata 2023).

For *novices* in qualitative research in SE, LLMs can be both an assistive tool and a challenge. LLMs can be used to sift through extensive datasets, identify initial patterns, and assist in some basic data coding, making the initiation phases smoother. However, novice researchers must be cautious. Relying heavily on LLMs without understanding the underlying domain of inquiry or the principles of qualitative data analysis can compromise the quality of research outputs and their own capabilities as researchers. It is essential to strike a balance to ensure data integrity and true learning of the research process.

Intermediate researchers will find LLMs useful as they dive into more complex data. LLMs can aid in identifying recurring themes and intricate patterns,

[4] https://www.sciencedirect.com/journal/information-and-software-technology/publish/guide-for-authors.

potentially elevating the quality of the analysis through its comprehensive approach. However, there is a potential risk of overreliance on the technology, leading to overconfidence in automated outputs. It is crucial for researchers to maintain a critical eye, ensuring that their growing reliance on LLMs does not overshadow the need for rigorous human oversight and contextual interpretation that their increasing experience affords them.

For *seasoned qualitative researchers*, LLMs present an opportunity to explore new breadth and depth within data analysis. For example, LLMs can be used to scale qualitative research beyond what is typically possible through human effort. Experienced qualitative researchers can boost their practice by taking on larger datasets for analysis, training bespoke LLMs where accessible, and developing descriptive findings, taxonomies, and theories that capture a wider range of contexts and are, therefore, more widely generalizable. But with this deeper dive comes a heightened responsibility for research integrity and accountability. The research outputs, while possibly enhanced by LLMs, must be thoroughly reviewed for inadvertent errors or biases. Furthermore, while LLMs can handle the heavy lifting of data analysis, experts must remain fully accountable for the interpretations and conclusions drawn.

For all levels of researchers, LLMs can expedite the data processing phase, but it is paramount that researchers do not bypass the essential learning and understanding phases of the research process. LLMs should be tools to enhance the process, not shortcuts that diminish the depth and richness of qualitative research in software engineering. The use of LLMs should not eclipse the importance of human judgment and insight.

6 Conclusion

As LLMs entrench themselves into most disciplines, SE research will not remain untouched. For qualitative SE research, LLMs offer a landscape rife with opportunities and challenges. Researchers, whether novices, intermediates, or experts, can embrace the potential of LLMs while remaining vigilant and anchored in the core tenets of qualitative inquiry.

Amidst the rising discourse on the potential threats of AI and LLMs, accentuated by media narratives, there exists a palpable concern within professional communities about AI's capability to replace human roles. Contrarily, empirical findings (Bano et al. 2023), rooted in an understanding of LLM capabilities and extant research, debunk the AI doomsday notion, particularly for qualitative researchers in software engineering. We project a harmonious future where LLMs and human researchers collaboratively further qualitative research. However, while LLMs, like GPT-4 and ChatGPT, show promise, the irreplaceable role of the human researcher in ensuring ethical conduct, well-motivated studies, the validity and reliability of research findings, and appropriate dissemination remains pivotal.

Considering the broader interaction between humans and LLMs, while the latter's adeptness in qualitative data analysis can optimize certain facets of research, it is imperative to note their limitations in capturing the intricate nuances inherent to human researchers. This sentiment is echoed in seminal anthropological and sociological works that emphasize the human touch in interpreting and understanding data. Critically, the ethical considerations surrounding LLM use, ranging from data privacy to intellectual property rights, call for rigorous scrutiny.

Author contributions MB, RH, and DZ contributed to the ideation. MB and RH wrote the main manuscript text. DZ and CT reviewed, updated, and improved the manuscript. MB and RH prepared the final version.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

Alkaissi, H., McFarlane, S.I.: Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus **15**, 192 (2023)

Arora, C., John, G., Mohamed, A.: Advancing requirements engineering through generative AI: assessing the role of LLMs. (2023) *arXiv preprint* arXiv:2310.13976.

Balel, Y.: The role of artificial intelligence in academic paper writing and its potential as a co-author', *Euro. J. Therap.*. (2023)

Bano, M., Didar Z., Jon W.: Exploring qualitative research using LLMs. (2023) *arXiv preprint* arXiv:2306.13298.

Bender, E.M., Timnit G., Angelina M.-M., Shmargaret S.: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp 610–23 (2021)

Byun, C., Piper, V., Kevin, S.: Dispensing with Humans in Human-Computer Interaction Research. In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–26 (2023)

Easterbrook, S., Singer, J., Storey, M.A., Damian, D.: Selecting empirical methods for software engineering research. Guide to Adv. Emp. Softw. Eng. **8**, 285–311 (2008)

Ebert, C., Louridas, P.: Generative AI for software practitioners. IEEE Softw. **40**, 30–38 (2023)

Emmert-Streib, F.: Importance of critical thinking to understand ChatGPT. Europ. J. Human Genet. **15**, 1–2 (2023)

Gentles, S.J., Cathy, C., Jenny, P., Ann Mckibbon, K.: Sampling in qualitative research: insights from an overview of the methods literature. Qual. Rep. **20**, 1772–1789 (2015)

Hoda, R.: Socio-technical grounded theory for software engineering. IEEE Transaction Software Engineering. **48**, 3808–3832 (2021)

Hou, X., Yanjie, Z., Yue, L., Zhou, Y., Kailong, W., Li, L., Xiapu, L., David, L., John, G., Haoyu, W.: Large language models for software engineering: a systematic literature review. *arXiv preprint* arXiv:2308.10620

Jalil, S., Suzzana, R., Thomas, D.L., Kevin, M., Wing, L.: Chatgpt and software testing education: Promises & perils. In: *2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, 4130–37. IEEE (2023)

Jiang, D., Xiang R., Bill Y-L.: LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. (2023) *arXiv preprint* arXiv:2306.02561.

Kitchenham, B.: Procedures for performing systematic reviews. Keele UK Keele Univ. **33**(2004), 1–26 (2004)

Kuhail, M.A., Sujith, S.M., Ashraf, K., Jose, B., Syed J.S.: Will I be replaced? Assessing chatgpt's effect on software development and programmer perceptions of Ai tools. *Assessing Chatgpt's Effect on Software Development and Programmer Perceptions of Ai Tools*.

Navigli, R., Simone, C., and Björn, R.: Biases in large language models: origins, inventory and discussion. *ACM J. Data Inform. Qual*. (2023)

Nguyen-Duc, A., Beatriz C.-D., Adam, P., Chetan, A., Dron, K., Tomas, H., Usman, R., Jorge, M., Eduardo, G., Kai-Kristian K.: Generative artificial intelligence for software engineering–a research Agenda, (2023) *arXiv preprint* arXiv:2310.18648.

Ozkaya, I.: Application of large language models to software engineering tasks: opportunities. Risks Implicat. IEEE Software. **40**, 4–8 (2023)

Polonsky, M.J., Jeffrey D.R.: Should artificial intelligent agents be your co-author? Arguments in favour, informed by ChatGPT. In: 91–96. SAGE Publications Sage UK: London, England (2023)

Rudolph, J., Tan, S., Tan, S.: ChatGPT: bullshit spewer or the end of traditional assessments in higher education? J. Appl. Learn. Teach. **24**, 6 (2023)

Scoccia, G.L.: Exploring Early Adopters' Perceptions of ChatGPT as a Code Generation Tool. In: *2023 38th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*, pp 88–93 (2023)

Treude, C., Hideaki H.: She Elicits Requirements and he tests: software engineering gender bias in large language models. (2023) *arXiv preprint* arXiv:2303.10131.

Watkins, R.: Guidance for researchers and peer-reviewers on the ethical use of large language models (LLMs) in scientific research workflows. AI Ethics **16**, 1–6 (2023)

Watson, C.: Unreliable narrators?'Inconsistency'(and some inconstancy) in interviews. Qual. Res. **6**, 367–384 (2006)