Moving Forward with Generative AI in Software Engineering Research

Marina Condé Araújo
Colorado State University
Fort Collins, CO, USA
marina.condearaujo.colostate.edu

Júlia Condé Araújo
Colorado State University
Fort Collins, CO, USA
julia.condearaujo.colostate.edu

Fabio Marcos de Abreu Santos
Colorado State University
Fort Collins, CO, USA
fabio.deabreusantos@colostate.edu

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

Abstract

Generative AI is rapidly reshaping software engineering (SE) research. While widely seen as a source of renewed momentum, enabling new tools, revitalizing established problems, and increasing research visibility, it also disrupts long-standing norms around rigor, evaluation, authorship, and responsibility. In this position paper, we analyze open-ended responses from a community survey conducted ahead of ICSE 2026 FOSE to capture how SE researchers perceive which aspects of GenAI adoption are working well and bringing joy, and which are creating stress or friction. Our findings show that researchers do not call for restricting GenAI use, but for clearer norms, SE-specific evaluation standards, stronger expectations around rigor and reproducibility, and greater attention to fairness in access.

CCS Concepts

• **Software and its engineering**;

Keywords

Generative AI, GenAI, Software Engineering, Research

ACM Reference Format:
Marina Condé Araújo, Júlia Condé Araújo, Fabio Marcos de Abreu Santos, and Bianca Trinkenreich. 2026. Moving Forward with Generative AI in Software Engineering Research. In *2026 IEEE/ACM 48th International Conference on Software Engineering: Future of Software Engineering (ICSE-FoSE), April 12–18, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3793657.3793879

1 Introduction

The software engineering (SE) research community is undergoing a profound transition driven by the rapid adoption of Generative AI (GenAI), particularly large language models (LLMs). These models have demonstrated strong potential across the research pipeline, supporting tasks such as code completion, bug repair, data analysis, and research writing [4]. Evidence from a large cross-disciplinary survey of more than 1,600 researchers suggests that many scientists expect AI tools to become central to research practice in the coming decade, while simultaneously expressing concern about how these tools may reshape standards of rigor, proof, and trust in scientific work [7]. This combination of optimism and unease provides an important backdrop for understanding how GenAI is being experienced within SE research.

Within SE, these tensions are amplified by the generative nature of LLMs. While their fluency and accessibility enable rapid experimentation and lower barriers to entry, they also introduce new risks, including hallucinated content, questionable reproducibility, and ambiguity around responsibility for errors or methodological choices [4, 6]. As a result, GenAI acts not only as a technical enabler, but also as a catalyst for renewed scrutiny of how SE research is conducted, evaluated, and governed.

At the same time, many SE researchers have welcomed GenAI as a source of renewed intellectual momentum. Its rapid evolution has opened or reinvigorated problem spaces, increased the visibility of SE research beyond traditional boundaries, and prompted active experimentation rather than outright resistance. Together, these contrasting experiences position GenAI as both a source of promise and disruption, raising fundamental questions about how it is reshaping everyday research practices and the health of the SE research ecosystem.

In this paper, we contribute an empirical, community-grounded perspective on this moment of transition by analyzing open-ended responses from a survey [5] conducted with SE researchers in late 2025, in preparation for the ICSE 2026 Future of Software Engineering (FOSE) event. The survey invited researchers to reflect on what is working well in the SE research community, what is not, and what changes they believe are needed. We focus specifically on responses that reference GenAI or LLMs in order to examine how researchers experience the benefits and tensions associated with GenAI adoption and how they envision improving its use in SE research. The following research questions guide our study:

RQ1 What aspects of Generative AI are working well and bringing joy in SE research?
RQ2 What aspects of Generative AI are not working well and bringing stress in SE research?
RQ3 What changes do researchers believe are necessary to improve how Generative AI is used in SE research?

ICSE-FoSE, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2479-4/2026/04
https://doi.org/10.1145/3793657.3793879

The remainder of this paper is structured as follows. Section 2 describes the survey design and qualitative analysis approach. Section 3 presents the results organized around the three research questions. Section 4 discusses threats to validity. Finally, Section 5 synthesizes the findings and articulates what a healthier GenAI-enabled SE research community might look like in light of these results.

2 Research Design

In this section we lay out our research design.

2.1 Data Collection and Preparation

The data analyzed in this paper originate from a community-facing survey designed by the ICSE 2026 FOSE organizers to elicit reflections on the state, practices, and future directions of SE research [5]. The questionnaire consisted of 12 optional, open-ended questions, allowing respondents to engage selectively with topics most relevant to their experiences.

The instrument included open-ended questions about what works well (and not well) in the SE research community, what brings them joy (and stress) and what change respondents would prioritize. In this paper, we focus specifically on the subset of responses that referenced Generative AI (GenAI) or LLMs, using these mentions to address our research questions: (RQ1) which GenAI-related aspects are perceived as working well and bringing joy in SE research, (RQ2) which GenAI-related challenges are perceived and how they are experienced as stress, and (RQ3) what changes respondents believe are needed to improve how GenAI is used in SE research.

Participation was anonymous, and as stated in the consent form, respondents were informed that their answers would be publicly shared via the FOSE Discord channel to support transparency and collective reflection. The survey was broadly disseminated by the organizers and shared within professional networks, resulting in a convenience sample capturing perspectives across career stages, roles, and research backgrounds.

In this paper, we analyze the anonymized responses as provided by the organizers. Our goal is not statistical generalization, but to surface recurring themes, tensions, and aspirations related to the adoption of Generative AI in SE research, in order to inform discussion at FOSE and beyond.

2.2 Data Analysis

The authors collaboratively conducted qualitative analysis of the open-ended responses using an inductive, data-driven approach inspired by open coding [2]. Rather than applying a predefined coding scheme, the analysis focused on closely reading the responses and iteratively identifying recurring issues, perceptions, experiences, and suggested changes related to the use of GenAI in SE research.

We focused our analysis on open-ended survey responses that explicitly referenced AI, GenAI, Generative AI, LLM, Large Language Model, or ChatGPT. In total, 289 respondents completed the survey. Across the relevant open-ended questions, the number of non-blank responses was 225 (what works well), 224 (sources of joy), 233 (what does not work well), 218 (sources of stress), and 215 (suggested changes).

Emerging themes were progressively refined through comparison across responses, with attention to internal consistency and variation in how participants articulated similar concerns or positive experiences. Throughout this process, the emphasis was on surfacing patterns grounded in participants' own language rather than producing an exhaustive or mutually exclusive code system.

The developing interpretations were then discussed among the authors. Through iterative discussions, we examined alternative readings of the data and resolved disagreements through negotiated agreement.

3 Results

We analyzed survey responses related to the adoption of Generative AI in SE research using a two-level thematic coding approach. Responses were first categorized based on whether participants described GenAI-related aspects as working well or not working well. Within each category, sub-themes were then identified to capture more specific aspects of participants' experiences. Representative excerpts are reported throughout this section, with numbers in parentheses indicating anonymized Response IDs from the ICSE 2026 FOSE pre-survey data.

3.1 What is working well and the joys of GenAI in SE Research (RQ1)

Regarding the GenAI aspects that are working well, respondents emphasized that the community does not only *“accept new techniques"* (195844264), but is able to react and has *“eagerness to adopt new tools (e.g., LLMs)"* (196381072).

When speaking of joys, respondents considered GenAI as a source of intellectual stimulation and renewed relevance. Respondents pointed to new or reinvigorated research problems emerging from rapid advances in LLMs, such as *“tasks like program repair, code generation because of the rapid development of LLMs"* (196381072). Others linked LLM-related work to a sense that SE research is being heard and valued, noting that *“LLMs applied to SE are well accepted”* (196071861) and that intelligent software engineering involving LLMs *“has helped improve developer efficiency”* (196061195).

Finally, some participants expressed satisfaction in being able to transfer established SE approaches to new contexts, such as *“applying the same method to LLM-based repositories"* (192593186), as shown in Figure 1.

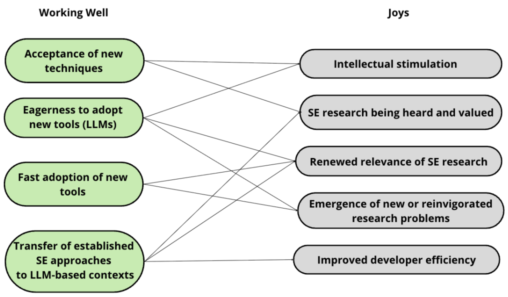
Working Well
Joys
Acceptance of new techniques
Eagerness to adopt new tools (LLMs)
Fast adoption of new tools
Transfer of established SE approaches to LLM-based contexts
Intellectual stimulation
SE research being heard and valued
Renewed relevance of SE research
Emergence of new or reinvigorated research problems
Improved developer efficiency

Figure 1: Relations between what is working well and the joys of GenAI

GenAI in SE research was reported as a catalyst for renewal, supporting fast adoption of new tools, reactivating dormant or overlooked problem spaces, and reinforcing the relevance of SE research in a rapidly evolving computing landscape.

3.2 What is NOT working well and the stressors of GenAI in SE Research (RQ2)

Regarding the GenAI aspects that are not working well, although most respondents did not reject AI-based research outright, they raised concerns about inconsistent norms and double standards, particularly around disclosure, evaluation practices, and methodological expectations. One participant pointed to asymmetries in review practices, noting that *“reviewers use ChatGPT but won’t disclose it, yet reject LLM-as-a-judge approaches”* (192311611). Others similarly described a lack of shared standards, observing that *“there are many AI-generated papers (and reviews), and often poor data and questionable reproducibility”* (195394675).

Beyond disclosure, participants expressed concern about the growing number of LLM-centric studies perceived as opportunistic or weakly grounded in SE problems. As one respondent put it, *“we have accepted too many papers of the form ‘we tried something with an LLM, here it is”’* (195811076), reflecting anxieties about trend-driven publication practices and an emerging “AI BUBBLE” in which alignment with fashionable techniques may outweigh rigor and insight. Relatedly, some participants highlighted conceptual uncertainty about contribution, questioning *“whether automating something by designing the prompt is valuable”* (192249436).

Participants also emphasized downstream consequences of LLM adoption for the peer-review system itself. Several noted that reviewing LLM-based work increases cognitive burden and erodes trust, with one respondent describing how *“double-checking the LLM as a reviewer is extremely stressful”* and shifts the burden of proof from authors to reviewers (195591458). Finally, respondents linked the LLM shift to fairness and resource inequality, emphasizing that access to computational infrastructure increasingly shapes what research can be conducted and published. One participant noted that *“the high computational cost [of LLMs] makes it difficult for researchers without industry collaboration to run large-scale experiments”* (195775809).

The rapid adoption of GenAI introduced new forms of stress linked to cognitive overload and uncertainty. Participants described difficulty keeping up with the pace of change, with one respondent stating that *“the rapid growth of AI models makes it stressful to constantly keep up”* (195775809). Others emphasized uncertainty around evaluation standards, explaining that *“I wish there was a more standard expectation for what makes a strong evaluation of software engineering-based LLM work”* (192249436).

Several participants also highlighted the additional burden placed on reviewers when assessing LLM-assisted work. One respondent described this as both a workload and accountability issue, noting that *“double-checking the LLM as a reviewer is extremely stressful”* and that the burden of proof increasingly shifts from authors to reviewers (192591458). This stress was further compounded by concerns about hidden errors introduced by LLM use, with participants reporting *“numerous errors with data extracted from papers via LLM”* (192591458). Taken together, these accounts suggest that GenAI-related stress arises not only from pace and novelty, but also from uncertainty about responsibility, trust, and the reliability of AI-assisted research outputs. As presented in Figure 2.

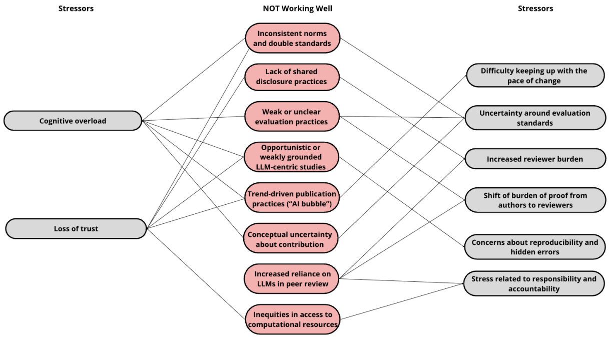
Stressors
NOT Working Well
Stressors
Cognitive overload
Loss of trust
Inconsistent norms and double standards
Lack of shared disclosure practices
Weak or unclear evaluation practices
Opportunistic or weakly grounded LLM-centric studies
Trend-driven publication practices (“AI bubble”)
Conceptual uncertainty about contribution
Increased reliance on LLMs in peer review
Inequities in access to computational resources
Difficulty keeping up with the pace of change
Uncertainty around evaluation standards
Increased reviewer burden
Shift of burden of proof from authors to reviewers
Concerns about reproducibility and hidden errors
Stress related to responsibility and accountability

Figure 2: Relations between what is not working well and the stressors of GenAI

On the concerning side, GenAI in SE research was reported to disrupt established norms and the stress this disruption generates. Researchers describe inconsistent disclosure and evaluation practices, hype-driven LLM studies, and growing inequities in access to computational resources as undermining rigor and fairness. These structural issues translate into cognitive overload, loss of trust, and increased reviewing burden, as researchers struggle to keep pace with rapid change while bearing greater responsibility for validating GenAI-assisted work.

3.3 What needs to be changed? (RQ3)

Participants proposed changes aimed at improving how GenAI is used and evaluated within SE research. Respondents emphasized the need for clearer norms, stronger evaluation practices, and safeguards against unintended negative consequences.

3.3.1 Establish clearer norms and disclosure practices. A recurring suggestion was the need for explicit and consistent norms governing GenAI use, particularly regarding disclosure. Participants expressed concern about ambiguous or uneven expectations, calling for clearer guidance on acceptable practices. One respondent noted the inconsistency in current norms, observing that *“reviewers use ChatGPT but won’t disclose it”* (192311611), while others argued that disclosure should be treated as a baseline requirement rather than an exception.

3.3.2 Define SE-specific evaluation standards for GenAI-based work. Participants also emphasized the need for clearer evaluation criteria tailored to SE research. Rather than relying on generic AI benchmarks, respondents called for expectations that reflect SE values such as rigor, relevance, and empirical grounding. One participant explicitly stated, *“I wish there was a more standard expectation for what makes a strong evaluation of software engineering-based LLM work”* (192249436). Others stressed that AI-based studies should be assessed using established SE principles, arguing that *“AI such as*

LLMs should be dealt with by using traditional software analysis and testing" (195844080).

3.3.3 *Reduce hype-driven and low-effort LLM studies.*

Participants suggested that the community should become more selective in what it accepts as meaningful GenAI research. Respondents criticized trend-driven submissions that lack substantive contribution, calling for higher bars for novelty and insight. As one participant put it, *"we have accepted too many papers of the form 'we tried something with an LLM, here it is"'* (195811076), arguing that novelty should be demonstrated beyond merely applying an LLM.

3.3.4 *Improve transparency and reproducibility.*

Another suggested change concerned transparency in data, methods, and tool use. Participants expressed concern that LLM-based studies often obscure key details, making results difficult to assess or reproduce. One respondent highlighted the prevalence of *"poor data and questionable reproducibility"* in AI-generated papers and reviews (195394675), suggesting that clearer reporting standards are needed to maintain trust and research integrity.

3.3.5 *Address inequities in access to GenAI resources.*

Finally, participants called attention to structural inequalities introduced or amplified by GenAI adoption. Respondents argued that access to computational resources increasingly shapes who can conduct large-scale LLM research, and suggested that this imbalance should be acknowledged and mitigated. One participant noted that *"the high computational cost [of LLMs] makes it difficult for researchers without industry collaboration to run large-scale experiments"* (195775809), raising concerns about fairness and inclusivity in future SE research.

The Figure 3 shows that not all items categorized as Not Working Well and Stressors are connected to Changes. While inconsistent norms and double standards, weak or unclear evaluation practices, trend-driven publication practices, concerns about reproducibility and hidden errors, and inequities in access to computational resources are explicitly linked to proposed changes, other elements such as cognitive overload, increased reviewer burden, difficulty keeping up with the pace of change, shift of burden of proof from authors to reviewers, and stress related to responsibility and accountability were not associated with any proposed change. This pattern suggests that while participants were able to articulate actionable interventions for some of the issues, others may be experienced as persistent or structural for which no clear remedies were proposed. The absence of associated changes does not imply that these issues are less salient; rather, it points to areas where the community may lack shared solutions, clear ownership, ideas about how to solve, or established mechanisms for intervention.

The SE community proposed changes that focus on clearer disclosure norms, SE-specific evaluation standards, stronger expectations around rigor and reproducibility, and higher selectivity against hype-driven LLM studies. Addressing inequities in access to computational resources is also seen as essential for ensuring that GenAI adoption strengthens, rather than fragments, the SE research community.

4 Threats to Validity

This study reflects a partial and situated view of how GenAI is experienced in SE research. The data capture self-reported reflections from researchers who chose to respond to a community-facing FOSE survey. They emphasize articulated concerns and perceived changes rather than directly observed practices or outcomes.

The sample is subject to self-selection and survivorship bias. Researchers who are disengaged, overwhelmed, or have already withdrawn from the broader community are likely underrepresented. As a result, some forms of frustration, exclusion, or disengagement may be muted rather than amplified in our findings.

Our qualitative analysis is interpretive and exploratory by design. While themes were iteratively discussed among the authors, other readings of the data are possible. Finally, consistent with a position paper, the goal is not generalization or causal inference, but to surface shared tensions and blind spots that can inform collective reflection and discussion within the SE community.

5 What a Healthier GenAI-Enabled SE Research Community Might Look Like?

Our results suggest that a healthier future for SE research with GenAI will not emerge from restricting GenAI use, but from realigning norms, responsibilities, and evaluation practices with how research is now conducted. Respondents consistently described stress arising not from GenAI itself but from ambiguity, including unclear expectations around disclosure, inconsistent evaluation criteria, and uncertainty about who bears responsibility when AI-assisted work fails. In a healthier future, GenAI use would be normalized and made explicit rather than exceptional. Disclosure practices would be consistent and expected for both authors and reviewers, reducing the current asymmetries reported by participants and alleviating the burden on reviewers to infer or police AI use. Such transparency aligns with the human-centered and accountable GenAI adoption in SE, which emphasize clarity, responsibility, and human agency as foundational values rather than optional add-ons [3].

Our participants showed concerns about hype-driven and weakly grounded LLM studies, pointing to the need for SE-specific evaluation standards that go beyond generic AI benchmarks. A healthier research ecosystem would assess GenAI-based work based on its contribution to SE knowledge, relevance to real software engineering problems, and empirical rigor—rather than novelty derived solely from applying an LLM. This ensures that LLMs augment, rather than replace, critical thinking, theory building, and methodological judgment in SE research [6].

Our findings also highlight that sustainability and well-being must be treated as first-class design goals of the research system. Participants reported a growing cognitive overload and a shifting burden of proof needed for papers, and pointed to inequities in access to computational resources that shape who can meaningfully participate in GenAI-driven research. A healthier future would explicitly recognize these pressures, for example by strengthening reporting and reproducibility expectations, discouraging low-effort LLM studies, and acknowledging resource constraints when evaluating large-scale GenAI experiments.

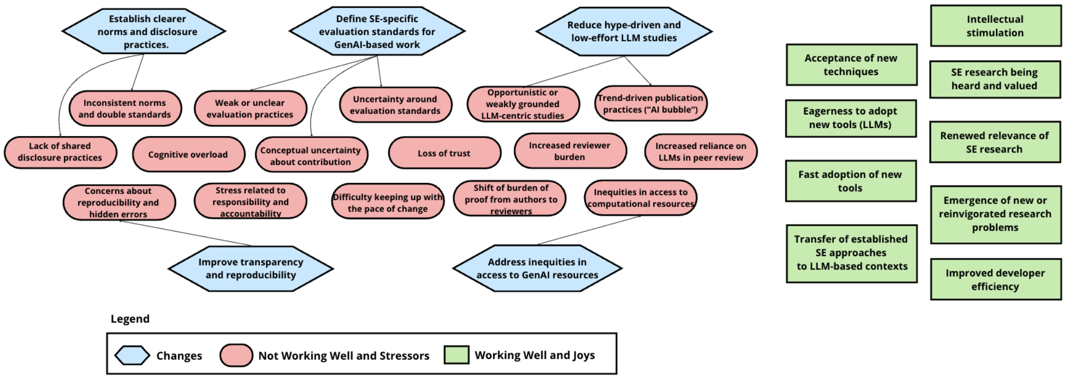
Establish clearer norms and disclosure practices.
Define SE-specific evaluation standards for GenAI-based work
Reduce hype-driven and low-effort LLM studies
Inconsistent norms and double standards
Weak or unclear evaluation practices
Uncertainty around evaluation standards
Opportunistic or weakly grounded LLM-centric studies
Trend-driven publication practices ("AI bubble")
Lack of shared disclosure practices
Cognitive overload
Conceptual uncertainty about contribution
Loss of trust
Increased reviewer burden
Increased reliance on LLMs in peer review
Concerns about reproducibility and hidden errors
Stress related to responsibility and accountability
Difficulty keeping up with the pace of change
Shift of burden of proof from authors to reviewers
Inequities in access to computational resources
Improve transparency and reproducibility
Address inequities in access to GenAI resources
Intellectual stimulation
Acceptance of new techniques
SE research being heard and valued
Eagerness to adopt new tools (LLMs)
Renewed relevance of SE research
Fast adoption of new tools
Emergence of new or reinvigorated research problems
Transfer of established SE approaches to LLM-based contexts
Improved developer efficiency
Legend
Changes
Not Working Well and Stressors
Working Well and Joys

Figure 3: Overview of what is working well and not working well in GenAI-based software engineering research. Aspects reported as not working well and causing stress are shown in the center (red ovals), while proposed changes intended to address these issues are shown above and below (blue hexagons). Arrows between hexagons and ovals indicate participant-reported relationships; participants did not propose changes for all reported issues. Aspects reported as working well and bringing joy are shown on the right (green squares).

Taken together, this vision for future GenAI reframes GenAI not as a disruptive force to be contained, but as a catalyst that makes long-standing fragilities in SE research governance visible. By acting on the changes articulated by the community—clearer norms, fairer evaluation, stronger commitments to rigor, and attention to equity, the SE research community can ensure that GenAI adoption contributes to renewal rather than burnout, and to collective progress rather than fragmentation.

Among the changes articulated by participants, having clear disclosure norms and shared evaluation criteria emerge as immediate leverage points, as they directly affect trust, the review burden, and perceptions of rigor across venues.

6 Conclusion

Taken together, these findings suggest that the challenges raised by GenAI adoption are not merely technical issues to be solved individually, but collective coordination problems that require shared decisions about norms, evaluation, and responsibility. Importantly, respondents' suggestions point to tensions that cannot be resolved unilaterally: clearer disclosure norms may increase transparency but also raise concerns about policing; higher evaluation standards may strengthen rigor but risk slowing innovation; addressing inequities in access may require trade-offs between ambition and inclusivity. These tensions underscore the need for intentional, community-level deliberation, rather than ad hoc responses by individual authors or reviewers.

In this sense, our results can serve as concrete inputs to the FOSE process, helping structure discussion around what a healthier SE research community should prioritize in the face of AI-accelerated change. The stressors identified in RQ2 highlight where current practices are breaking down, while the community-proposed changes in RQ3 delineate plausible directions for reform that merit collective evaluation. FOSE provides a timely forum to debate which expectations should be standardized across venues, how responsibility for AI-assisted work should be distributed among authors and reviewers, and what institutional mechanisms, such as guidelines (e.g. [1]) and review norms could support sustainability and fairness. By grounding these discussions in lived experiences rather than abstract principles, the community can move from diagnosing problems toward coordinated action aimed at restoring trust, reducing burnout, and sustaining meaningful participation in SE research.

References

[1] Sebastian Baltes, Florian Angermeir, Chetan Arora, Marvin Muñoz Barón, Chunyang Chen, Lukas Böhme, Fabio Calefato, Neil Ernst, Davide Falessi, Brian Fitzgerald, et al. 2025. Guidelines for empirical studies in software engineering involving large language models. *arXiv preprint arXiv:2508.15503* (2025).

[2] Matthew B Miles and A Michael Huberman. 1994. *Qualitative data analysis: An expanded sourcebook.* sage.

[3] D Russo, S Baltes, and Christoph TREUDE. 2024. Generative AI in software engineering must be human-centered: The Copenhagen Manifesto. *Journal of Systems and Software* 216 (2024), 1.

[4] June Sallou, Thomas Durieux, and Annibale Panichella. 2024. Breaking the silence: the threats of using llms in software engineering. In *Proceedings of the 2024 ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results.* 102–106.

[5] Margaret Storey and Andre van der Hoek. 2025. Community Survey for ICSE 2026 Future of Software Engineering: Toward a Healthy Software Engineering Community. doi:10.5281/zenodo.18217798 Zenodo repository record.

[6] 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 *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering.* 1503–1507.

[7] Richard Van Noorden and Jeffrey M Perkel. 2023. AI and science: what 1,600 researchers think. *Nature* 621, 7980 (2023), 672–675.