"So what if I used GenAI?" - Legal Implications of Using Cloud-based GenAI in Software Engineering Research

Gouri Ginde
Department of Electrical and Software Engineering
University of Calgary
Email: gouri.ginde@ucalgary.ca

***Abstract*—Generative Artificial Intelligence (GenAI) advances have led to new technologies capable of generating high-quality code, textual content, and images. The next step is to integrate GenAI technology into various aspects while conducting research or other related areas, a task typically conducted by researchers. Such research outcomes always come with a certain risk of liability. This vision paper sheds light on the various research aspects in which GenAI is used, thus raising awareness of its legal implications to novice and budding researchers. In particular, there are two risks: data protection and copyright. Both aspects are crucial for GenAI. We summarize key aspects regarding our current knowledge that every software researcher involved in using GenAI should be aware of to avoid critical mistakes that may expose them to liability claims and propose a checklist to guide such awareness.**

I. Introduction

Generative AI (GenAI) is revolutionizing software development and software engineering (SE) in gereral more profoundly than any other recent technology [1]. It has proliferated into research-oriented aspects as well [2]. At the core of GenAI are large language models (LLMs), vast neural networks trained on massive text datasets. Approximately 60% of GenAI applications are utilized during the software development phase [1]. However, it is important to note that any input provided—whether a simple prompt or content-specific details—can contribute to the ongoing evolution of these typically cloud-based LLMs. Conversely, the outputs generated by LLMs are not guaranteed to be free from intellectual property rights (IPR) considerations, posing potential legal and ethical challenges. Thus, while using cloud-based GenAI services, software researchers must be careful of its implications on their research outcomes.

GenAI can be used in various aspects of SE research such as scholarly paper reviewing, brainstorming and ideating using GenAI, writing manuscripts, and programming and developing (source) code. A few of the dilemmas faced by the researchers are shown in Figure 1. The ownership of the content generated by the model and the use of third-party content within the generated elements (reviews/code/manuscript) are also significant ethical concerns. For example, while GenAI can assist editors and peer reviewers in completing repetitive or tedious tasks, there is a risk that it may not mitigate existing biases and that human judgment calls are still necessary.

The academic community has been rattled by the free access to GenAI services and their usage in research. Well aclaimed conference ICML wrote the following as part of thier policies: "The Large Language Model (LLM) policy for ICML 2023 prohibits text produced entirely by LLMs (i.e., "generated"). Similarly, ICSE and other conferences in SE have similar guidelines for LLM usage policy. Although this does not prohibit authors from using LLMs for editing or polishing author-written text, it is evident that the GenAI tools raise concerns about research transparency, reproducibility, and potential biases that threaten scientific integrity and equity in research outcomes.

Our objective in this study is to understand the risks and implications of using cloud-based GenAI in SE research and propose a checklist to overcome or mitigate the possible risks associated with GenAI's utilization in research.

GenAI is a broader category of AI tools designed to generate new content such as images, text, video, code and audio. In this paper, we will focus on LLMs, which are specifically designed to process and generate human language. LLMs can be grouped into three types: open-sourced, enterprise and free-tier access [3], [4], [5]. Table I provides detailed information on these types and their differences.

Motivation: Our high-level analysis on the academic and law communities on Stack Exchange[1], a widely used question-answering platform in SE [6], showed some intriguing findings where mostly free-tier (cloud-based) GenAI tools are discussed. Of the 45,000 questions posted on the academic Stack Exchange, over 2,000 are on ethics, 960 on plagiarism, and 180 on research misconduct. However, only 45 questions with the Generative-AI tag are posted. The law Stack Exchange with an artificial intelligence tag had 81 questions. A closer look at these questions (a few listed below) and their views (V) raise serious concerns about research's legal and ethical aspects.

- Should I report a review I suspect to be AI-generated? (5k views)

[1]https://stackexchange.com/about - a network of question-and-answer (Q&A) websites on topics in diverse fields, each community site covering a specific topic, where questions, answers, and users are subject to a reputation award process. Comprises over 173 Q&A communities

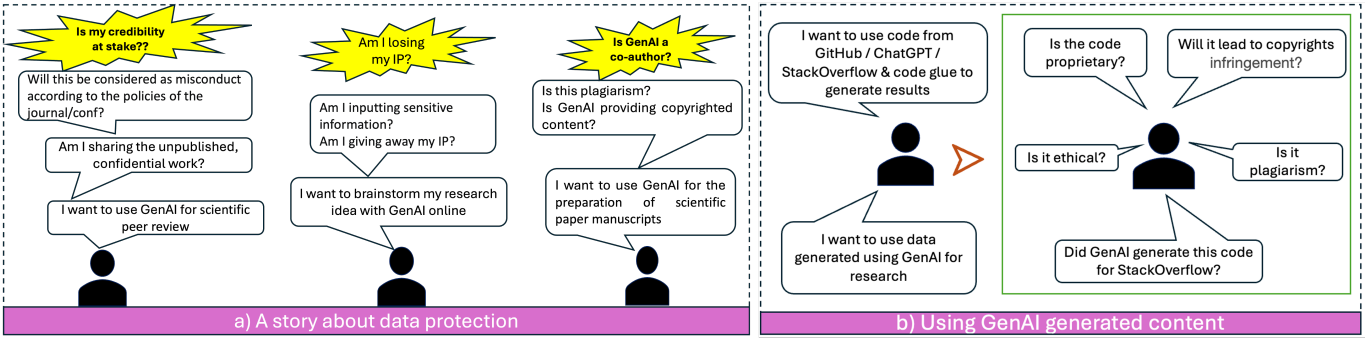

Fig. 1. Conceptual scenarios when legal implications of GenAI should be considered while doing research

TABLE I
SUMMARY OF COMPARISON OF VARIOUS TYPES OF LLMS [3], [4], [5]

	Open-Source	**Free-Tier (cloud-based)**	**Enterprise (license-based)**
Accessibility	Publicly available with open licensing	Limited free access via APIs/interfaces	Subscription-based access
Customization	Fully customizable; fine-tuning supported.	Limited to prompt engineering; no fine-tuning.	Domain-specific tuning.
Data Privacy	Full control; user responsibility for security	Data may be retained for model improvements	Guaranteed privacy
Cost	Free to use; hosting and training incur costs	Free but with usage limits	Subscription pricing
Ease of Use	Requires ML expertise to deploy and manage	Beginner-friendly; no setup needed	Seamless integration options.
Performance	Varies by model; requires optimization.	Optimized for general-purpose tasks.	High performance, tailored for business needs.
Examples	LLaMA, Falcon, GPT-J, GPT-NeoX, Mistral.	ChatGPT Free, Google Bard, Claude Instant.	ChatGPT Enterprise, Azure OpenAI

- What should I do if I suspect one of the journal reviews I got is AI-generated? (27K views)
- Co-author uses ChatGPT for academic writing - is it ethical? (18K views)
- Should I preemptively confess after submitting work that was partially generated by ChatGPT? (9k views)
- Are there examples of journals with an explicit policy on GPT-3 and equivalent language models? (2K views)
- Is it OK to generate parts of a research paper using a large language model such as ChatGPT? (23K views)
- Is Utilizing AI Tools for Conducting Literature Reviews in Academic Research Advisable? (9K views)
- How can a programming beginner effectively utilize ChatGPT as a tool for programming? (338 views)
- Is there any automated system available that validate the accuracy of the data generated by GenAI? (397 views)
- Is the generated code by Code-Llama and LLama-2 models licensed somehow or it has copyright issues? (287 views)
- Copyright risks for code contributed by generative AI (699 views)
- Do Llama-2 and Code-Llama models collects my code? (422 views)

These questions show evidence for a widespread curiosity regarding the use of GenAI in academic writing, reviewing, and research processes in general; however, there is little interest regarding the copyright, licensing, and ethical issues about GenAI's utilization in research (last four questions) although its repercussions can be detrimental. This motivated us to explore the ethical implications of using GenAI in research, especially in SE, as data, code, and evaluation are the core of it.

II. WHAT ARE THE RISKS AND IMPLICATIONS OF USING GENAI IN SE RESEARCH?

Recently, legal implications of using GenAI have been explored in domains such as media [7] and medical education [8] [9]. GenAI has exponentially triggered unethical news, gossip outlets, and disinformation networks in the media domain, making it impossible for the general public to differentiate weed from chaff. The medical education study emphasizes that the rapid development, adoption and use of AI technologies in healthcare requires healthcare professionals to master experimental techniques, even if they are not yet recognized as standards. Similar implications and risks exists for SE domain and we list them as follows.

Data Privacy and Security: Researchers have been increasingly using GenAI such as ChatGPT to assist in ideation [10] due to its ability to act as a conversatioal agent. LLM models such as GALACTICA [11] and MINERVA [12] can store, combine and reason about scientific language have shown remarkable results. However, researchers might overlook the terms of services (TOS) which clearly mention that the unless opted out, they may use content to train their future models to provide, maintain, develop, and improve their services [13]. Thus, one might give away their ideas and highly private and sensitive information to a unknown entity. Additionally, in a recent study conducted at 2024 Neural Information Processing Systems (NeurIPS) conference on the use of LLMs in the scientific peer review process [14] demonstrated that over 70% authors found it useful and were willing to revise their papers based on the feedback. However, it was cautioned that this approach has serious data privacy and security concerns.

While there is a debate regarding if it is acceptable to using

an AI to do the bulk of writing Latex [15] as it would be considered cheating in the same way that the use of a calculator would be looked upon as cheating if you were being tested on your ability with mental arithmetic. Additionally, the data privacy and security are still at stake as the work is still unpublished.

Licensing Issues: Noticing the gravity of the issue, Stackoverflow (SO) made a strict policy against use of GenAI use while posting answers [16] as they observed that there are SO users active for years that previously produced only few answers now posting over 50 in less than a day. The amount of AI generated answers could suffocate SO if everyone starts doing it without giving proper credit to the AI. Contributors tend to present the content as their own, thus misrepresenting someone else's work. GenAI models are trained on massive scale datasets available online. However, not all reveal the source of their training data which has raised wide scaled uproar against copyright violations. Recent innovation of GitHub copilot came under hammer due to such an allegation [17] which clearly explains that the Training AI systems on public GitHub repositories, and potentially additional sources, has led to the violation of legal rights of many creators who posted code or other work under specific open-source licenses on GitHub. These licenses include 11 popular open-source licenses, such as the MIT, GPL, and Apache, all requiring proper attribution of the author's name and copyright.

Academic integrity: It is debatable whether LLMs like ChatGPT are suitable for editing and polishing text. For example, interviewed by the Verge [18], Deb Raji (AI research fellow, Mozilla Foundation) highlightes that LLM differ from tools like Grammarly, as they are not solely designed for text refinement but also generate novel content, including potentially problematic outputs like spam. This makes them more complex and distinct from simpler corrective tools. Similarly, COPE (Committee on Publication Ethics) [19] is international body which is committed to educating and supporting editors, publishers, universities, research institutes, and all those involved in publication ethics warns that critics have recently suggested that using AI for review puts confidential information back into the public domain.

Copyright and Intellectual Property: SO has put two year ban on using GenAI for coding in SO answers, however, they have been struggling to identify and put heavy moderating efforts towards it [16]. Moderator's post regarding this which is viewed 1.4M times emphasizes that there are SO users active for years that previously produced only few answers now posting over 50 in less than a day. Also, moderators cautioned that amount of AI generated answers could suffocate SO if everyone starts doing it.

Evolving AI regulations: OpenAI TOS policies on their website [13] says, "The content co-authored with the OpenAI API policy, creators who wish to publish their first-party written content (e.g., a book, compendium of short stories) created in part with the OpenAI API are permitted to do so under the condition that the published content must be attributed to the author's name or company, with a clear disclosure of the AI's role in generating the content, ensuring that readers can easily understand the involvement of AI". For instance, according to this statement, one must detail in a Foreword or Introduction (or some place similar) the relative roles of drafting, editing, etc. People should not represent API-generated content as being wholly generated by a human or wholly generated by an AI. It is a human who must take ultimate responsibility for the content being published [20][2].

III. Proposing: Generative AI Transparency & Accountability Evaluation (GATE) Checklist

Using a checklist to document any process is a well established concept [21]. For example, since the 1930s, checklists have been a standard operating procedure for pilots and other aviators in the aviation industry. In medicine, checklists are used as a decision aid to identify a medical condition and decide on an appropriate course of treatment. In comparison, surgical checklists are recommended as a safety measure to reduce the margin of human error and any adverse effects during surgery [21]. In SE, Wieringa et al. [22] developed a checklist as a guide to performing empirical research effectively. Belli et al. [23] developed a checklist to streamline code reviewing. Recently, Patel et al. [24] proposed a comprehensive release-readiness checklist for GenAI-based Software Products. This was designed to guide practitioners in evaluating release readiness aspects such as performance, monitoring, and deployment strategies, aiming to enhance the reliability and effectiveness of LLM-based applications in real-world settings. Taking inspiration from these works, in this vision paper, we propose a two pronged checklist to guide researchers when using cloud-based LLM tools in various academic tasks that need understanding regarding data protection and awareness regarding legal implications while using GenAI generated content. Table II is the checklist we propose. Please refer to additional notes in the last column of the table for more details.

IV. Related work

GenAI in SE: Ebert et al. [2] explored the utility of GenAI for improving software development and software productivity through code generation, test case generation from requirements, re-establishing traceability, explaining code, refactoring of legacy code, and software maintenance with augmented guidance. However, Ebert et al. caution that while generative AI can help in all these tasks, several risks need to be considered and mitigated. For example, AI tools can hallucinate, causing privacy and security implications as the code shared with the tool is not open-sourced or, worse, if proprietary, might be used for training leading to grave consequences. Sauvola et al. [25] analyzed the potential of generative AI and LLM technologies for future software development paths and highlighted the need for new tools to understand the potential, limitations, and risks of generative AI, and provided

[2]Policies emerging in European Union, the People's Republic of China, and the United States, as well as the governance efforts in multilateral settings (e.g., G7) are trying to design safeguards into the processes and procedures around using GenAI

TABLE II
GENERATIVE AI ACCOUNTABILITY & TRANSPARENCY EVALUATION (GATE) CHECKLIST FOR RISK ASSESSMENT AND RESEARCH DISCLOSURE STANDARDS

Transparency assessment			
Data legality	Is the data source legally compliant?	[ ] Yes [ ] No	- Verify the dataset's source (publicly available, licensed, proprietary). - Confirm compliance with data usage rights. - Ensure no copyrighted or sensitive data is used. - Check compliance with regulations (e.g., GDPR, HIPAA).
Output ownership	Is output ownership clear?	[ ] Yes [ ] No	- Check the GenAI tool's terms of service. - Evaluate output's licensing implications. - Ensure proper attribution for outputs to avoid IP claims.
Regulatory compliance	Does the research comply with AI regulations?	[ ] Yes [ ] No	- Assess compliance with local AI regulations (e.g., EU AI Act). - Ensure adherence to ethical AI guidelines. - Check institutional review board (IRB) requirements.
Licensing compatibility	Is the licensing of GenAI outputs compatible?	[ ] Yes [ ] No	- Verify compatibility with open-source licenses. - Avoid combining incompatible licenses (e.g., proprietary and open-source).
Accountability (disclosure standards) assessment			
GenAI usage declaration	Is GenAI usage disclosed?	[ ] Yes [ ] No	- Clearly state which parts of the research utilized GenAI. - Provide the name and version of the GenAI tool used.
Output attribution	Is GenAI output clearly attributed?	[ ] Yes [ ] No	- Distinguish between human-authored and GenAI-generated content. - Attribute outputs (e.g., "Generated using [Tool Name] Accessed on [Date]").
Compliance statement	Is there an ethical/legal compliance statement?	[ ] Yes [ ] No	- Include a compliance statement for ethical guidelines and legal requirements. - Disclose and reference data sources.
GenAI contribution	Is GenAI's contribution documented?	[ ] Yes [ ] No	- Describe GenAI's specific role in the methodology. - Disclose limitations and threats to validity and how they were addressed.
Authorship	Are researchers credited, not GenAI?	[ ] Yes [ ] No	- Ensure researchers are credited as primary contributors. - Credit GenAI as a tool or assistant..
Open Science	Is source code made public referencing other code repositories used for development?	[ ] Yes [ ] No	- Acknowledge other source repositories such as GitHub and Stackoverflow which were used in the code development.

guidelines for using it. Carteton et al. [26] highlighted that GenAI brings new ethical dilemmas and intellectual property (IP) challenges. The ownership of AI-generated code remains ambiguous, questions such as, "Who is responsible for the generated end-product?", demand legally sound guidelines for AI-assisted creation to ensure accountability. From a regulatory perspective, the responsibility of automatically generated codes and content bypassing ethical considerations must be addressed.

Legal dimensions of GenAI usage: Weis et al. [27] highlight that the GenAI may have been trained on data-protected by regulations such as the General Data Protection Regulation (GDPR)[3], which prohibits the re-use of data beyond the purposes for which it was collected. LLMs, often called "stochastic parrots," can reproduce or remix training data, potentially violating copyrights or imposing restrictive licenses on outputs. For instance, Codex may generate copyrighted code or code under non-commercial Creative Commons licenses. A lawsuit against GitHub, Microsoft, and OpenAI highlights these concerns [17]. Fransces et al. [28] explained that the GenAI tools use copyrighted works for training and store copies of protected works for training purposes. Also, the nature of outputs generated by genAI—unlike traditional AI, which follows explicit rules provided by programmers—relies on techniques and acquired knowledge without direct human intervention. This complexity makes determining who holds copyright ownership for such outputs challenging. It was also highlighted that the legal status of using copyrighted works for non-market-encroaching purposes (such as research) remains unclear and may depend on specific circumstances. Researchers have given special guidelines to clarify their position through TOS if other users could use their generative model, and to keep updated about the evolution of the legislative frameworks at the national and international level [29].

[3]A legal framework that governs the collection and processing of personal data for individuals in the European Union (EU) and the European Economic Area (EEA)

V. Conclusion

In this vision paper, we investigates the need for a checklist while using cloud-based free-tier GenAI that entails 1) Transparency aspects such as data legality, ethical considerations, licensing and regulatory compliance and 2) Accountability (GenAI generated content usage) related aspects such as authorship and GenAI usage declaration in software engineering research. Through this work, we envision to spark discussion and awareness about the legal risks of GenAI in software engineering research and offer a forward-looking vision and actionable steps for researchers to address these challenges. We believe that our proposed checklist can guide researchers in evaluating legal and ethical implications of using GenAI products in research. In the future, we will work on demonstrating the utility of this checklist through case studies and empirical validation.

References

[1] C. Ebert, J. P. Arockiasamy, L. Hettich, and M. Weyrich, "Hints for generative ai software development," *IEEE Software*, vol. 41, no. 5, pp. 24–33, 2024.

[2] C. Ebert and P. Louridas, "Generative ai for software practitioners," *IEEE Software*, vol. 40, no. 4, pp. 30–38, 2023.

[3] "Open-source llms vs closed: Unbiased guide for innovative companies [2025]." https://hatchworks.com/blog/gen-ai/open-source-vs-closed-llms-guide/. (Accessed on 12/06/2024).

[4] "Open-source llm vs closed source llm for enterprises." https://datasciencedojo.com/blog/open-source-llm/. (Accessed on 12/06/2024).

[5] H. Alipour, N. Pendar, and K. Roy, "Chatgpt alternative solutions: Large language models survey," *arXiv preprint arXiv:2403.14469*, 2024.

[6] S. Wang, T.-H. Chen, and A. E. Hassan, "Understanding the factors for fast answers in technical q&a websites: An empirical study of four stack exchange websites," *Empirical Software Engineering*, vol. 23, pp. 1552–1593, 2018.

[7] J. Bayer, "Legal implications of using generative ai in the media," *Information & Communications Technology Law*, vol. 33, no. 3, pp. 310–329, 2024.

[8] M. M¨arz, M. Himmelbauer, K. Boldt, and A. Oksche, "Legal aspects of generative artificial intelligence and large language models in examinations and theses," *GMS Journal for Medical Education*, vol. 41, no. 4, p. Doc47, 2024.

[9] Z. N. Khlaif, A. Mousa, M. K. Hattab, J. Itmazi, A. A. Hassan, M. Sanmugam, and A. Ayyoub, "The potential and concerns of using ai in scientific research: Chatgpt performance evaluation," *JMIR Medical Education*, vol. 9, p. e47049, 2023.

[10] R. Gozalo-Brizuela, "A survey of generative ai applications," *Cornell University*, 2023.

[11] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic, "Galactica: A large language model for science," *arXiv preprint arXiv:2211.09085*, 2022.

[12] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, *et al.*, "Solving quantitative reasoning problems with language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3843–3857, 2022.

[13] "Terms of use — openai." https://openai.com/policies/row-terms-of-use/. (Accessed on 12/05/2024).

[14] A. Goldberg, I. Ullah, T. G. H. Khuong, B. K. Rachmat, Z. Xu, I. Guyon, and N. B. Shah, "Usefulness of llms as an author checklist assistant for scientific papers: Neurips'24 experiment," *arXiv preprint arXiv:2411.03417*, 2024.

[15] "mathematics - is it plagiarism using an ai to do the bulk of my latex? - academia stack exchange." https://academia.stackexchange.com/questions/206563/is-it-plagiarism-using-an-ai-to-do-the-bulk-of-my-latex. (Accessed on 12/06/2024).

[16] "Policy: Generative ai (e.g., chatgpt) is banned - meta stack overflow." https://meta.stackoverflow.com/questions/421831/policy-generative-ai-e-g-chatgpt-is-banned. (Accessed on 12/06/2024).

[17] "Github copilot litigation · joseph saveri law firm & matthew butterick." https://githubcopilotlitigation.com/. (Accessed on 12/06/2024).

[18] "Chatgpt and ai language tools banned by ai conference for writing papers - the verge." https://www.theverge.com/2023/1/5/23540291/chatgpt-ai-writing-tool-banned-writing-academic-icml-paper. (Accessed on 12/06/2024).

[19] "About cope — cope: Committee on publication ethics." https://publicationethics.org/about/our-organisation. (Accessed on 12/06/2024).

[20] M. Christodorescu, R. Craven, S. Feizi, N. Gong, M. Hoffmann, S. Jha, Z. Jiang, M. S. Kamarposhti, J. Mitchell, J. Newman, *et al.*, "Securing the future of genai: Policy and technology," *Cryptology ePrint Archive*, 2024.

[21] N. M. Minhas, J. Börstler, and K. Petersen, "Checklists to support decision-making in regression testing," *Journal of Systems and Software*, vol. 202, p. 111697, 2023.

[22] R. Wieringa, "Towards a unified checklist for empirical research in software engineering: first proposal," in *16th International Conference on Evaluation & Assessment in Software Engineering (EASE 2012)*, pp. 161–165, IET, 2012.

[23] F. Belli and R. Crisan, "Towards automation of checklist-based code-reviews," in *Proceedings of ISSRE'96: 7th International Symposium on Software Reliability Engineering*, pp. 24–33, IEEE, 1996.

[24] H. Patel, D. Boucher, E. Fallahzadeh, A. E. Hassan, and B. Adams, "A state-of-the-practice release-readiness checklist for generative ai-based software products," *arXiv preprint arXiv:2403.18958*, 2024.

[25] J. Sauvola, S. Tarkoma, M. Klemettinen, J. Riekki, and D. Doermann, "Future of software development with generative ai," *Automated Software Engineering*, vol. 31, no. 1, p. 26, 2024.

[26] A. Carleton, D. Falessi, H. Zhang, and X. Xia, "Generative ai: Redefining the future of software engineering," *IEEE Software*, vol. 41, no. 6, pp. 34–37, 2024.

[27] J. D. Weisz, M. Muller, J. He, and S. Houde, "Toward general design principles for generative ai applications," *arXiv preprint arXiv:2301.05578*, 2023.

[28] G. Franceschelli and M. Musolesi, "Copyright in generative deep learning," *Data & Policy*, vol. 4, p. e17, 2022.

[29] J. Ren, H. Xu, P. He, Y. Cui, S. Zeng, J. Zhang, H. Wen, J. Ding, P. Huang, L. Lyu, *et al.*, "Copyright protection in generative ai: A technical perspective," *arXiv preprint arXiv:2402.02333*, 2024.