



Proceeding Paper

Can You Trust Your Copilot? A Privacy Scorecard for AI Coding Assistants [†]

Amir Al-Maamari

Faculty of Computer Science and Mathematics, University of Passau, 94032 Passau, Germany; almaam03@ads.uni-passau.de

[†] Presented at the 1st International Conference on Emerging Tech & Innovation (ICETI), Buraydah, Saudi Arabia, 10 February 2026.

Abstract

The rapid integration of AI-powered coding assistants into developer workflows has raised significant privacy and trust concerns. As developers entrust proprietary code to services like OpenAI's GPT, Google's Gemini, and GitHub Copilot, the unclear data handling practices of these tools create security and compliance risks. This paper addresses this challenge by introducing and applying a novel, expert-validated privacy scorecard. The methodology involves a detailed analysis of four document types—from legal policies to external audits—to score five leading assistants against 14 weighted criteria. A legal expert and a data protection officer refined these criteria and their weighting. The results reveal a distinct hierarchy of privacy protections, with a 20-point gap between the highest- and lowest-ranked tools. The analysis uncovers common industry weaknesses, including the pervasive use of opt-out consent for model training and a near-universal failure to filter secrets from user prompts proactively. The resulting scorecard provides actionable guidance for developers and organizations, enabling evidence-based tool selection. This work establishes a new benchmark for transparency and advocates for a shift towards more user-centric privacy standards in the AI industry.

Keywords: AI coding assistants; data protection; large language models; software engineering; LLM policy

1. Introduction

AI-powered coding assistants like GitHub Copilot and Amazon Q Developer have rapidly integrated into developer workflows, becoming essential programming tools [1]. A recent survey highlights their utility: 53% of developers use them for boilerplate code and 43% for debugging [2]. This reliance on remote processing, however, creates privacy risks, as proprietary code is transmitted to third-party servers. High-profile incidents, such as Samsung banning ChatGPT after confidential code leaks [3,4] and Italy's GDPR-related suspension of the service [5], underscore the conflict between utility and privacy.

1.1. Problem and Goal

The central problem is the opacity of privacy practices across competing AI coding assistants, which prevents developers from making informed trust decisions. Lengthy, jargon-laden legal documents obscure what happens to submitted code, creating compliance and intellectual property (IP) risks. This paper addresses this gap by evaluating the privacy practices of five prominent assistants—OpenAI GPT, Anthropic Claude, Google



Academic Editors: Haifa F. Alhasson
and Shuaa S. Alharbi

Published: 25 May 2026

Copyright: © 2026 by the author.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Gemini, GitHub Copilot, and Amazon Q Developer—using a novel, comparative privacy scorecard. The study is guided by four research questions (RQs) exploring what data are collected (RQ1), how it is used and what controls are offered (RQ2), how the tools compare (RQ3), and the implications of these findings (RQ4).

1.2. Contributions

This work contributes to privacy in AI-assisted software engineering by: (1) introducing a novel, expert-validated privacy scorecard framework with weighted criteria; (2) providing a comparative analysis of five leading tools, revealing key differences and undocumented risks; and (3) delivering actionable rankings and guidance for developers, organizations, and policymakers to mitigate risks and foster trustworthy AI practices.

2. Related Work

The adoption of LLM-based coding tools has sparked discussion on privacy, security, ethics, and IP implications [6–8]. Studies have examined how AI services align with legal requirements; for instance, ref. [9] analyzed LLM providers' privacy policies against GDPR, producing compliance scores and identifying legal risks. Similarly, ref. [10] evaluated the privacy compliance of ChatGPT and Claude, finding gaps between policy statements and actual data protection measures. These analyses provide valuable high-level benchmarks but focus on general-purpose models, not the specific context of coding assistants.

Another research line explores developer trust. A survey by [8] found that over 64% of practitioners distrusted AI assistants for privacy-related tasks due to a lack of transparency. This aligns with findings that emphasize reliability as key to trusting AI. The security of generated code is also under scrutiny, with studies showing GitHub Copilot can produce vulnerable code [11]. Furthermore, research on model memorization demonstrates that LLMs can regurgitate sensitive training data, underscoring the risks of using user code for retraining [12].

This paper builds on these efforts but diverges in key ways. First, it focuses specifically on AI coding assistants, where privacy lapses translate directly into IP leaks and security vulnerabilities. Second, its methodology extends beyond privacy policies to a broader evidence corpus, including technical documentation and enterprise agreements, for a more holistic assessment. Finally, it contributes an expert-validated quantitative scorecard that translates complex policies into an actionable ranking, shifting from general analysis toward a specialized, practitioner-focused comparison.

3. Methodology

This research follows a three-phase methodology: evidence collection, framework development, and quantitative scoring.

3.1. Phase 1: Evidence Corpus Collection

The analysis is grounded in a comprehensive evidence corpus snapshotted from mid-2025. This corpus spans four document categories for each provider: (1) Core Legal Documents (e.g., Privacy Policy, Terms of Service); (2) Enterprise-Tier Documents (e.g., Data Processing Addendums, Trust Center pages); (3) Technical Documentation (e.g., API guides, white papers); and (4) External Evidence (e.g., SOC 2 audits, public records). For each criterion, specific document sources and keywords were predefined to ensure every score is traceable to documented evidence.

3.2. Phase 2: Framework Development and Validation

Grounded in privacy principles like GDPR, an evaluation framework of 14 sub-criteria across three categories was developed. A legal expert and a data protection officer (DPO) validated the framework, assessing each sub-criterion for relevance and clarity and performing a 100-point constant-sum allocation to determine category weights. This produced the final scorecard (Table 1), with weights confirming Technical Safeguards (42.5%) as the most critical area, followed by Data Governance (32.5%) and Transparency & User Control (25.0%).

Table 1. Validated framework showing criteria, weights, and expert validation scores.

Criterion Category & Sub-Criteria	Relevance	Clarity	Weight	Expert Comment
A: Data Governance & Legal Compliance	5.0	5.0	32.5%	Confirmed as essential. Experts emphasized IP rights and explicit consent.
A1: Public & clear DPA (Data Processing Addendum)	5.0	5.0		
A2: Explicit consent for training	5.0	5.0		
A3: Clear data residency policy	5.0	5.0		
A4: Definition of IP rights	5.0	5.0		
A5: Straightforward DSAR (Data Subject Access Request)	5.0	5.0		
B: Technical Privacy & Security Safeguards	5.0	4.6	42.5%	Confirmed as most critical category. Minor ambiguity in terms noted.
B1: Anonymization for training data	5.0	5.0		
B2: Role-Based Access Control (RBAC)	5.0	4.0		
B3: Encryption in transit and at rest	5.0	5.0		
B4: Filtering of secrets/PII	5.0	5.0		
B5: Regular 3rd-party audits/pen-testing	5.0	4.0		
C: Transparency & User Control	5.0	5.0	25.0%	Confirmed as essential. Experts valued granular controls and disclosures.
C1: Granular opt-out controls	5.0	5.0		
C2: Transparency on telemetry	5.0	5.0		
C3: Clear data retention periods	5.0	5.0		
C4: Publicly available model card	5.0	5.0		

3.3. Phase 3: Quantitative Scoring and Ranking

The five selected assistants were scored against the 14 sub-criteria using a 3-point rubric (0 = Not Met, 1 = Partially Met, 2 = Fully Met). A final composite score (scaled

to 100) was calculated using the expert-defined weights via the formula: $\text{Score}_{\text{tool}} = 100 \cdot \sum_{i \in \{A,B,C\}} (w_i \cdot s_i)$, where w_i is the weight for category i and s_i is the tool's normalized score for that category. The complete rubric is in a public replication package.

4. Results and Discussion

The analysis reveals a clear hierarchy in privacy performance (Figure 1). Google Gemini leads with a score of 89.25, followed by Anthropic Claude (81.88) and GitHub Copilot (78.75). A significant gap separates them from Amazon Q Developer (72.38) and OpenAI's GPT models (68). This 20-point spread shows that tool choice has measurable privacy implications.

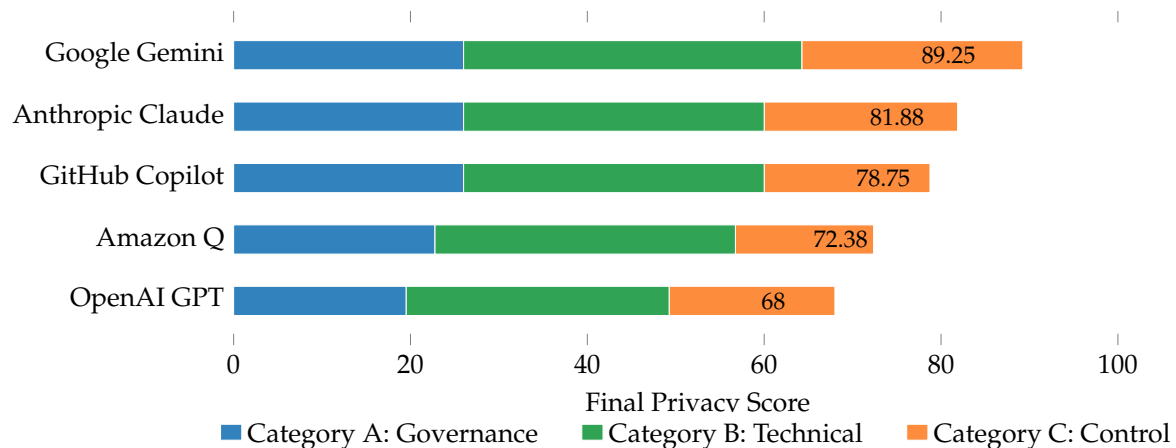


Figure 1. Final privacy scores and per-category performance of AI coding assistants. The total length of each bar corresponds to the final weighted score, labeled on the right.

4.1. Analysis of Privacy Dimensions

The overall scores reflect distinct strengths and weaknesses (RQ1-3). In Data Governance, most providers scored well, but all except Anthropic Claude (the only one with an opt-in model) use an **opt-out** model for training consent on consumer tiers. In Technical Safeguards, a critical cross-vendor weakness emerged: a near-universal failure to proactively filter secrets from prompts, placing the burden on the user. In Transparency & User Control, a clear maturity gap appeared. Leaders like Gemini and Claude provide granular controls and clear retention policies, while others offer only global opt-outs and vague statements like keeping data “for as long as is necessary.”

4.2. Implications and Recommendations (RQ4)

The findings have direct implications. For Practitioners and Organizations, selecting an AI assistant must be a security decision. We recommend to: (1) Prioritize Enterprise Tiers, which consistently offer superior safeguards like zero-data-retention and contractual guarantees; (2) Assume Zero-Filtering and enforcing strict policies against including proprietary code or secrets in prompts; and (3) Favor Opt-In by Default by selecting tools like Anthropic Claude where privacy is paramount.

For Providers and Policymakers, the industry must improve. We recommend to: (1) Make Opt-In the Standard for model training to eliminate dark patterns; (2) Build Proactive Safeguards to automate the detection and redaction of sensitive data from prompts; and (3) Embrace Radical Transparency with clear, numerical data retention timelines and comprehensive model cards.

5. Threats to Validity

Threats to validity were addressed. Construct validity was ensured by grounding criteria in privacy principles and expert validation by a legal expert and DPO. Internal validity was addressed through a detailed, evidence-based rubric and a public replication package to mitigate researcher bias. External validity is limited by the dynamic nature of the AI landscape (a mid-2025 snapshot) and the focus on five proprietary tools, which may not generalize to open-source or self-hosted alternatives.

6. Conclusions

This paper introduced and applied an expert-validated privacy scorecard to benchmark leading AI coding assistants, translating opaque policies into a quantitative, comparable measure. The analysis revealed a 20-point performance gap and systemic industry weaknesses, such as pervasive opt-out consent models and a lack of automated secret filtering. The scorecard framework stands as the primary contribution—a reusable tool for holding providers accountable and empowering organizations to make evidence-based decisions. While the findings are a time-sensitive snapshot limited to five proprietary tools, this work establishes a new benchmark for transparency. Future work should extend this analysis to open-source models and track policy changes over time. Ultimately, this research advocates for a necessary industry shift toward more transparent and user-centric privacy standards.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The replication package for this study is openly available on Google Sheets at: https://docs.google.com/spreadsheets/d/1mpVCiv_1WSV33ovtnb-qmf1s3pRjFTonCluhGu56HDo/edit?usp=sharing (accessed on 16 May 2025).

Acknowledgments: The author thanks Abdulatif Alabdulatif and Rashid Alharbi for validating the scorecard framework. OpenAI's GPT-4o was used as a writing assistant for prose refinement and formatting; the author retains full responsibility for all content and findings.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. ResearchAndMarkets.com. Generative Artificial Intelligence Coding Assistants—Global Strategic Business Report. 2025. Available online: <https://www.businesswire.com/news/home/20250319490646/en/> (accessed on 15 May 2025).
2. CodeSignal. Developers and AI Coding Assistant Trends. 2024. Available online: <https://codesignal.com/report-developers-and-ai-coding-assistant-trends/> (accessed on 16 May 2025).
3. Petkauskas, V. Lessons Learned from ChatGPT's Samsung Leak. Organization: CyberNews. 2023. Available online: <https://cybernews.com/security/chatgpt-samsung-leak-explained-lessons/> (accessed on 16 May 2025).
4. DeGeurin, M. Oops: Samsung Employees Leaked Confidential Data to ChatGPT. Organization: Gizmodo. 2023. Available online: <https://gizmodo.com/chatgpt-ai-samsung-employees-leak-data-1850307376> (accessed on 16 May 2025).
5. Ferrari, A.T. The Italian Data Protection Authority halts ChatGPT's Data Processing Operations. Organization: Clifford Chance. 2023. Available online: <https://www.cliffordchance.com/insights/resources/blogs/talking-tech/en/articles/2023/04/the-italian-data-protection-authority-halts-chatgpt-s-data-proce.html> (accessed on 16 May 2025).
6. Xu, W.; Gao, K.; He, H.; Zhou, M. A first look at license compliance capability of LLMs in code generation. *arXiv* **2024**, arXiv:2408.08381. [CrossRef]
7. Wang, C.; Chen, Z.; Li, T.; Zhang, Y.; Liu, Y. Towards Trustworthy LLMs for Code: A Data-Centric Synergistic Auditing Framework. In Proceedings of the 47th IEEE/ACM International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER), Ottawa, Ontario, Canada, 27 April–3 May 2025; Presented on 1 May 2025, in the SE for AI track.

8. Madampe, K.; Grundy, J.; Arachchilage, N. How Are We Doing With Using AI-Based Programming Assistants For Privacy-Related Code Generation? The Developers' Experience. *arXiv* **2025**, arXiv:2503.03988. [[CrossRef](#)]
9. Vu-Minh, A.; Nguyen, H.H.; Tran, T. Comparative Analysis of Privacy Policies in Large Language Models: Compliance with Data Privacy Laws. In Proceedings of the PACIS 2025 Proceedings, Kuala Lumpur, Malaysia, 5–9 July 2025; Paper 9.
10. Chadwick, R.; Blundell, S.; Prendergast, E. Assessments of the Privacy Compliance in Commercial Large Language Models. OSF Preprint. 2024. Available online: https://osf.io/preprints/osf/ugmfw_v1 (accessed on 1 July 2025).
11. Pearce, H.; Ahmad, B.; Tan, B.; Dolan-Gavitt, B.; Karri, R. Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions. *Commun. ACM* **2025**, *68*, 96–105. [[CrossRef](#)]
12. Carlini, N.; Tramèr, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, Ú.; et al. Extracting Training Data from Large Language Models. In Proceedings of the 30th USENIX Security Symposium (USENIXD Security 21), USENIX Association, Virtual, 11–13 August 2021; pp. 2633–2650.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.