GHAgentFiles: A dataset of coding agent file histories in GitHub repositories

Guillaume Cardoen[1], Tom Mens[1] and Alexandre Decan[1,2]

[1]*Software Engineering Lab, University of Mons, Belgium*
[2]*F.R.S.-FNRS Research Associate*

Large language models (LLMs) have recently become widely used, particularly in collaborative software development, where coding LLMs are used to review or generate code automatically [1]. In an ideal scenario, the LLM acts autonomously, attempting to complete a task specified by a *prompt* in natural language. Such semi-autonomous LLMs are commonly called *agents*. Semi-autonomous coding agents are used to automate a wide variety of tasks in software engineering [2].

Equipping agents with project-specific knowledge improves both their autonomy and their performance, and *context files* have emerged as one of the most common mechanisms to this end [3]. Typically written in natural language and stored as Markdown documents in predefined locations within the repository, these files provide coding agents with background information while they perform their tasks. The scope of such files varies widely: they can cover the project's conventions and architecture, style guidelines, instructions for building, testing, and deploying, recommendations regarding quality and security, and the boundaries within which agents should operate. Several publicly available examples exist online.[1]

Prior empirical studies have examined the structure, maintenance, adoption, and impact of context files on agents [1, 2, 4, 5], but have largely relied on small datasets. Galster et al. [3, 6] extracted 4,768 context files from 2,853 repositories across five major coding tools, and found that context files serve as the primary configuration mechanism for agents, with AGENTS.md emerging as an open standard. In a follow-up work [6], they extended their dataset to 4,738 repositories. Mohsenimofidi et al. [7] analyzed 453 repositories and found that only 5% of them use context files, which primarily document conventions, guidelines, and project architecture, and that changes to these files consist mostly of adding instructions. Chatlatanagulchai et al. [8], analyzing more than 2,300 files, reported that context files follow a consistent hierarchy of Markdown headers but tend to be long and hard to read, and that they are actively maintained to provide actionable guidance.

Given the increasing popularity of AI agents and their potential impact on the collaborative software development process, we contribute a dataset of coding-agent file histories comprising 142K+ file histories from 27K+ of the 165K+ repositories considered and 401K+ revisions, tracked until 1 July 2026. In addition to context files, the dataset includes *skill files*, which are reusable step-by-step descriptions of how to perform specific tasks, as well as *subagent files*, which serve as context files for specialized agents to which an orchestrating agent delegates tasks. By releasing this corpus, we aim to facilitate both the replication of existing studies on a substantially larger dataset and new empirical studies at scale.

To identify and extract coding-agent files (context, skill, and subagent files) from GitHub repositories, we developed cofee, an open-source CLI tool that extracts file histories from a local or remote git repository based on glob patterns declared in a .toml configuration file. cofee traverses the first-parent history from a given reference (defaulting to HEAD) and stores every touched file whose path matches the configured patterns, together with its git metadata (e.g.,, author name, commit hash, commit date, …). These patterns currently target six coding agents (Claude, Codex, GitHub Copilot, Gemini, Windsurf,

BENEVOL 2026: The 25th Belgium-Netherlands Software Evolution Workshop, 26-27 November 2026, Madrid, Spain
guillaume.cardoen@umons.ac.be (G. Cardoen); tom.mens@umons.ac.be (T. Mens); alexandre.decan@umons.ac.be (A. Decan)
0009-0005-2008-3565 (G. Cardoen); 0009-0005-2008-3565 (T. Mens); 0000-0002-5824-5823 (A. Decan)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]See, for example, https://agents.md

and Cursor), identified as the most prevalent in previous work [2, 3, 5], with an extra generic group for the emerging AGENTS.md and SKILL.md standards.

This dataset can be used to conduct empirical studies of coding-agent files, such as those initiated in [3, 7, 8]. This presentation will primarily focus on introducing the dataset and its main characteristics, including the categories of agent files it covers and the metadata tracked for each revision. We will discuss how the presented dataset can support replication and new studies at scale. To close, we will summarize preliminary results from our ongoing empirical work on agent files, such as the main topics covered by agent files, and the extent of duplication across those files.

The paper [9], dataset [10], and cofee tool[2] are freely accessible.

References

[1] R. Robbes, T. Matricon, T. Degueule, A. Hora, S. Zacchiroli, Agentic much? Adoption of coding agents on GitHub, ACM Transactions on Software Engineering and Methodology (2026).

[2] R. Robbes, T. Matricon, T. Degueule, A. Hora, S. Zacchiroli, Promises, perils, and (timely) heuristics for mining coding agent activity, in: Int'l Conf. Mining Software Repositories, IEEE, 2026.

[3] M. Galster, S. Mohsenimofidi, J. L. Lulla, M. A. Abubakar, C. Treude, S. Baltes, Configuring agentic ai coding tools: An exploratory study, in: 3rd ACM International Conference on AI-Powered Software, AIware '26, Association for Computing Machinery, 2026, pp. 11–20.

[4] J. L. Lulla, S. Mohsenimofidi, M. Galster, J. M. Zhang, S. Baltes, C. Treude, On the impact of AGENTS.md files on the efficiency of AI coding agents, preprint arXiv:2601.20404 (2026).

[5] H. Li, H. Zhang, A. E. Hassan, The rise of AI teammates in software engineering (SE) 3.0: How autonomous coding agents are reshaping software engineering, arXiv:2507.15003 (2025).

[6] M. Galster, S. Mohsenimofidi, L. Böhme, J. L. Lulla, M. A. Abubakar, C. Treude, S. Baltes, A dataset of agentic AI coding tool configurations, in: Int'l Conf. AI-Powered Software, 2026.

[7] S. Mohsenimofidi, M. Galster, C. Treude, S. Baltes, Context engineering for AI agents in open-source software, in: Int'l Conf. Mining Software Repositories, ACM, 2026.

[8] W. Chatlatanagulchai, H. Li, Y. Kashiwa, B. Reid, K. Thonglek, P. Leelaprute, A. Rungsawang, B. Manaskasemsak, B. Adams, A. E. Hassan, H. Iida, Agent READMEs: An empirical study of context files for agentic coding, preprint arXiv:2511.12884 (2025).

[9] G. Cardoen, T. Mens, A. Decan, GHAgentFiles: A dataset of coding agent file histories in GitHub repositories, in: 42nd International Conference on Software Maintenance and Evolution, IEEE, 16 September 2026. URL: https://orbi.umons.ac.be/handle/20.500.12907/56423.

[10] G. Cardoen, T. Mens, A. Decan, GHAgentFiles: A dataset of coding agent context file histories in GitHub repositories, 2026. URL: https://doi.org/10.5281/zenodo.20396489.

[2]The cofee tool can be accessed at https://github.com/sgl-umons/cofee