A Dataset of Agentic AI Coding Tool Configurations

Matthias Galster
University of Bamberg
Bamberg, Germany
mgalster@ieee.org

Seyedmoein Mohsenimofidi
Heidelberg University
Heidelberg, Germany
s.mohsenimofidi@uni-heidelberg.de

Levi Böhme
University of Bayreuth
Bayreuth, Germany
levi.boehme@uni-bayreuth.de

Jai Lal Lulla
Singapore Management University
Singapore, Singapore
jailal.l.2025@phdcs.smu.edu.sg

Muhammad Auwal Abubakar
University of Bamberg
Bamberg, Germany
muhammad.abubakar@uni-bamberg.de

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

Sebastian Baltes
Heidelberg University
Heidelberg, Germany
sebastian.baltes@uni-heidelberg.de

Abstract

Agentic AI coding tools such as Claude Code and OpenAI Codex execute multi-step coding tasks with limited human oversight. To steer these tools, developers create repository-level configuration artifacts (e.g., Markdown files) for configuration mechanisms such as Context Files, Skills, Rules, and Hooks. There is no curated dataset yet that captures these configurations at scale. This dataset, collected from open-source GitHub repositories, fills that gap. We selected 40,585 actively maintained repositories through metadata filtering, classified them using GPT-5.2 to identify 36,710 as belonging to engineered software projects, and systematically detected configuration artifacts in these repositories. The dataset covers 4,738 repositories across five tools (Claude Code, GitHub Copilot, OpenAI Codex, Cursor, Gemini) and eight configuration mechanisms. We collected 15,591 configuration artifacts, the full content of 18,167 configuration files associated with these configuration artifacts, and 148,519 AI-co-authored commits. The dataset and the construction pipeline are publicly available on Zenodo under CC BY 4.0. An interactive website allows researchers to browse and explore the data. This data supports research on context engineering, AI tool adoption patterns, and human-AI collaboration.

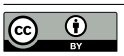
CCS Concepts

• **Software and its engineering**;

Keywords

Software Engineering, Generative AI, AI Agents, Configuration

ACM Reference Format:
Matthias Galster, Seyedmoein Mohsenimofidi, Levi Böhme, Jai Lal Lulla, Muhammad Auwal Abubakar, Christoph Treude, and Sebastian Baltes. 2026. A Dataset of Agentic AI Coding Tool Configurations. In *Proceedings of the 3rd ACM International Conference on AI-Powered Software (AIware '26), July 6–7, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3805760.3814922



AIware '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2601-9/2026/07
https://doi.org/10.1145/3805760.3814922

1 Introduction

Agentic AI coding tools have changed how developers write software [7]. Tools such as OpenAI Codex, Claude Code, and Cursor operate with greater autonomy than traditional AI coding assistants: they plan multi-step tasks, execute code, run tests, and iterate on results with minimal human intervention [5, 12]. These tools break down high-level goals into subtasks and adjust their approach based on real-time feedback [17]. For instance, when asked to implement a new feature, such a tool may identify the relevant parts of the codebase, apply the necessary modifications, run tests, and refine its output based on the results.

As agentic tools see growing adoption [16], developers have started to configure their behavior through *configuration mechanisms*, i.e., means by which developers tailor tool and agent behavior to a project or workflow. Prior work [6] identified and defined eight configuration mechanisms as follows: Context Files are Markdown files loaded into the context of each session; Skills bundle reusable knowledge and invocable workflows; Subagents are specialized agents that operate in parallel to the central agent loop, in their own context; Commands are user-triggered shortcuts for predefined prompts; Rules are system-level instructions to control agent behavior; Settings are JSON/TOML configurations for project-level tool behavior; Hooks are scripts executed at specific agent lifecycle points; and MCP configurations register external tool or data connections via the Model Context Protocol.

A *configuration artifact* is a tangible instance of a mechanism [6]: either a single configuration file (e.g., `CLAUDE.md`, or one subagent file such as `.claude/agents/reviewer.md`) or a directory bundling several configuration files that together define one artifact (e.g., a skill whose directory contains `SKILL.md` alongside scripts, references, and assets). Through configuration artifacts, developers communicate project constraints, coding conventions, and architectural decisions to AI agents [3, 13]. These artifacts

are version-controlled and collaboratively maintained. They represent a new category of software engineering documentation written for AI agents rather than human teammates, functioning as specifications that encode constraints and expected behaviors to guide code generation and modification.

Despite this growing adoption, no curated dataset systematically captures how developers configure agentic AI coding tools across open-source repositories. Without such data, researchers cannot study questions such as which configuration mechanisms developers actually use, how configurations differ across tools and programming languages, or how AI-assisted contributions relate to project characteristics or impact project quality.

We address this gap with a dataset of agentic AI coding tool configurations collected from GitHub. This work builds on a previous study [6], which identified a taxonomy of eight configuration mechanisms and characterized their adoption across 2,853 repositories. Here, we contribute the underlying data and construction pipeline as reusable research artifacts with expanded scope and new data types: (1) a curated sampling frame of 40,585 actively maintained GitHub repositories with rich metadata; (2) LLM-based classification of 36,710 repositories as belonging to *engineered* software projects, including justifications for the classification; (3) 15,591 configuration artifacts from 4,738 repositories across five tools and eight configuration mechanisms, together with the full content of 18,167 configuration files; (4) 148,519 AI-co-authored commits from repositories that use agentic AI coding tools; and (5) a fully reproducible construction pipeline and an interactive website for exploring the dataset.

2 Related Work

Work related to this dataset can be categorized into studies of AI-assisted development behavior, analyses of configuration artifacts, and existing datasets on AI-generated code.

Robbes et al. [16] measured coding-agent adoption on GitHub through file-, commit-, and pull-request-level heuristics over 128,018 repositories, analyzing adoption rates and the size and type of AI-assisted commits. Their file categorization distinguishes structured *configuration files* (such as YAML) from natural-language *rules and guidance files* (including `CLAUDE.md`, `AGENTS.md`, and Cursor rules) and is coarser than our eight-mechanism taxonomy. Their sample applies activity and size thresholds (non-fork, 5,000 lines of code, 100 commits, recent activity) without an engineered-project classification step, so their population includes tutorial and educational repositories that our engineered-project filter excludes; the two populations are complementary and support comparative analysis. Watanabe et al. [19] and Horikawa et al. [9] studied AI-authored pull requests and agentic refactoring, respectively. Both examined what AI tools produce, not how developers configure them. He et al. [8] analyzed Cursor's effect on development velocity and code complexity for a single tool without examining configuration artifacts. Jiang and Nam [10] studied Cursor rule files, covering one configuration mechanism for one tool. These studies produced empirical findings but not reusable, multi-tool datasets.

Two studies target configuration artifacts directly. Chatlatanagulchai et al. [3] conducted an empirical study of agent context files, a single artifact type. Mohsenimofidi et al. [13] studied context engineering practices across open-source repositories, analyzing the textual content of context files. Our prior study [6] defined a taxonomy of eight configuration mechanisms across five tools and analyzed their adoption in 2,853 repositories. That work produced findings and published supplementary scripts and CSV summaries for those repositories. This paper contributes an expanded, standalone dataset with 4,738 repositories, new data types not present in the supplementary material (148,519 AI-co-authored commits, 18,167 raw configuration files with full text content, per-artifact git metadata), and an improved pipeline to classify whether repositories belong to *engineered* software projects, reducing unsure cases (i.e., repositories that could not confidently be classified) by 93%. The raw file contents, combined with per-commit AI authorship, let researchers measure for each configuration artifact how much was written by humans versus by AI tools—a question that file presence alone cannot answer.

Several related datasets capture other aspects of AI-assisted development. The MSR 2026 Mining Challenge dataset (AIDev) [12] provides 932,791 agent-authored pull requests across five AI tools, focusing on code contributions rather than configurations. AgentPack [21] collects 1.8M code edits co-authored by AI agents and humans, primarily to support training future code-editing models. Xiao et al. [20] mined explicit self-admissions of GenAI usage from commits and documentation, providing high-precision but inherently incomplete signals since many projects do not disclose AI tool usage. Repository sampling approaches such as the SEART GitHub Search (GHS) tool [4] and the criteria of Munaiah et al. [14] for identifying engineered projects provide foundations for data collection, but no existing dataset applies them to AI coding tool configurations specifically. To our knowledge, no dataset provides multi-tool AI coding tool configurations with associated metadata, temporal information, and raw artifact contents.

3 Dataset Overview

Throughout this paper, the GitHub repository is the unit of analysis. A single software project may span multiple repositories (e.g., separate frontend and backend repositories). We therefore classify each repository individually as belonging to an engineered software project [14]; the classification prompt extends Munaiah et al.'s original definition and is included in the pipeline archive [2] (see Section 4.2).

The dataset is organized around three nested levels of repositories. At the broadest level, a *sampling frame* of 40,585 GitHub repositories provides metadata for the general population of actively maintained open-source projects. Of these, 36,710 are classified as belonging to engineered software projects. At the narrowest level, our dataset contains 4,738 repositories with detected AI coding tool configurations, corresponding metadata, raw configuration file contents, and AI-co-authored commit histories. Table 1 lists all data files with row counts. We describe each level below.

Sampling frame. All 40,585 repositories include GitHub metadata: primary programming language, star and fork counts, commit totals, contributor counts, license, creation date, and repository topics. The ten programming languages in the sampling frame are C, C#, C++, Go, Java, JavaScript, PHP, Python, Rust, and TypeScript.

Among the 4,738 repositories with detected configurations, the top five languages are TypeScript (24.5%), Python (18.7%), Go (13.8%), Java (8.0%), and C# (7.5%); the remaining 27.5% is split across Rust, JavaScript, C++, PHP, and C. This baseline allows researchers to compare tool-adopting repositories against the broader population and to control for language-specific effects.

Classification. Of the 40,585 repositories, 1,159 were excluded before classification because they had become unavailable (15), lacked a `README` file (74), or had a non-English `README` (1,070). The remaining 39,426 repositories carry classification labels produced by GPT-5.2, indicating whether each repository belongs to an engineered software project. Each label includes a justification text. Of the 39,426 classified repositories, 36,710 belong to engineered software projects, 2,564 do not, and 152 are unsure cases. We discuss this process in more detail in Section 4.2.

Configuration artifacts. Among the 36,710 repositories belonging to engineered software projects, 4,738 contain at least one agentic AI coding tool configuration artifact. The dataset covers five tools (Claude Code, GitHub Copilot, OpenAI Codex, Cursor, and Gemini) and eight configuration mechanisms that we introduced in Section 1: Context Files, Skills, Subagents, Commands, Rules, Settings, Hooks, and MCP configurations. Figure 1 shows the number of repositories per tool (one repository can use more than one tool): Claude Code leads with 2,525 repositories, followed by Copilot (1,397) and Cursor (466). An additional 909 repositories contain only an `AGENTS.md` file without any tool-specific configuration, suggesting use of a tool-agnostic convention. Figure 2 shows the distribution of artifacts across types. Context Files are the most common (found in 4,463 repositories), while Hooks (101) and MCP configurations (124) remain rare.

Figure 3 breaks this down further by showing which configuration mechanisms are used with which tools. Context Files are near-universal across all tools. The clearest tool-specific pattern is Rules, adopted by 63.5% of Cursor repositories compared to at most 17.0% of repositories in any other tool's user base. Claude Code accounts for most Skills, Subagents, and Commands in absolute terms because it has the largest user base (2,525 repositories), but proportional adoption within each tool is broadly comparable: for example, Cursor and Codex users adopt Skills (22.1% and 50.9%) and Commands (17.8% and 28.3%) at rates similar to or higher than Claude Code users (17.1% and 10.3%). This cross-tabulation, derived directly from the dataset, illustrates the configuration practices forming around each tool. Figure 4 shows which combinations of configuration mechanisms co-occur: the majority of repositories use only Context Files, while smaller groups combine Context Files with Settings, Skills, or multiple advanced mechanisms.

Per-artifact CSV files record the file path, git creation date, commit count, and first and last commit SHAs for each artifact. Context Files include additional metadata: detected natural language, line count, whether the file references another file, and AI authorship information for the file's commits. The archive `repos_data.7z` contains the full text of all 18,167 collected configuration files, enabling qualitative content analysis and model training. The file count (18,167) is larger than the artifact count (15,591) because some artifacts are directories: each Skill, Subagent, Command, or Rule can bundle multiple files (e.g., a Skill may pair `SKILL.md` with scripts, references, and assets). All 15,591 artifacts across all eight mechanism types were validated to be non-empty.

Table 1: Dataset files and row counts. Each row in the per-mechanism CSVs (`context_files.csv` through `hooks.csv`) is one configuration artifact, summing to 15,591 artifacts across the eight mechanisms. Large files are distributed as 7z archives.

File	Rows	Description
`repos.csv`	40,585	Sampling frame with metadata
`commits.csv`	148,519	AI-co-authored commits
`context_files.csv`	9,470	Context files
`skills.csv`	2,430	Skills
`commands.csv`	1,098	Commands
`rules.csv`	997	Rules
`subagents.csv`	884	Subagents
`settings.csv`	472	Settings
`mcp.csv`	138	MCP server configurations
`hooks.csv`	102	Hooks
`repos_data.7z`	–	18,167 raw configuration files

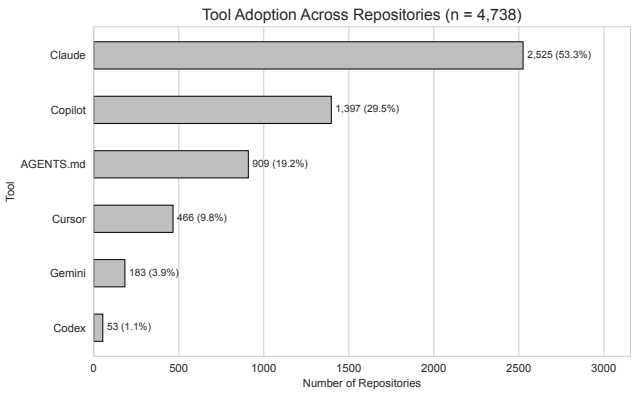


Figure 1: Number of repositories with detected configuration artifacts per agentic AI coding tool. Bars sum to more than 4,738 because a repository can configure multiple tools.

AI-co-authored commits. For every repository with at least one configuration artifact, we collected all AI-co-authored commits (see Section 4.3 for the detection method). The resulting `commits.csv` contains 148,519 commits from 3,392 repositories, each with a timestamp, full commit message, and the detected AI tool. For Context Files, the per-artifact CSV additionally records whether the file's creation commit was AI-co-authored and the total number of AI-co-authored commits that modified the file.

4 Dataset Construction

The dataset was constructed through a three-stage pipeline: repository sampling, LLM-based classification, and configuration artifact extraction. Figure 5 illustrates the pipeline with counts at each stage.

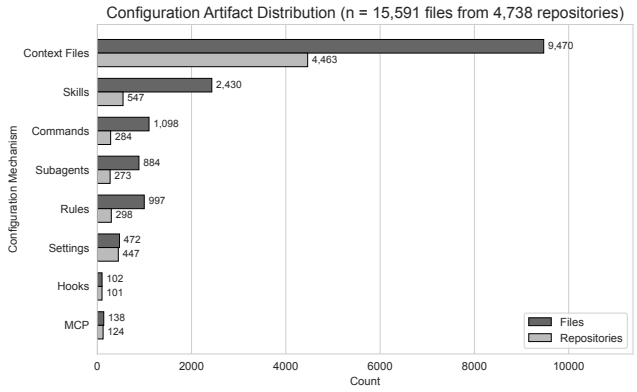


Figure 2: Number of configuration artifacts by type. For each mechanism, the chart shows the total file count and the number of repositories containing at least one artifact.

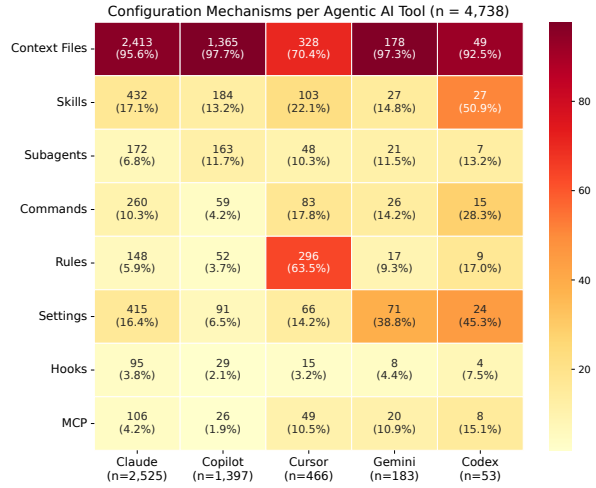


Figure 3: Configuration mechanisms per tool. Each cell shows the count and percentage of repositories using a given tool that also adopt the corresponding mechanism.

4.1 Repository Sampling

We used the GHS[1] dataset [4] as a starting point. GHS provides a regularly updated snapshot of GitHub repository metadata collected via the GitHub API. At the time of our query, it contained over one million repositories with metadata covering repository information, activity metrics, and social engagement indicators.

Our baseline query was designed to avoid common pitfalls in GitHub mining [11]. A repository in our sample had to (1) have a license, (2) not be a fork, (3) have at least two contributors, (4) have at least one pull request, (5) be at least 18 months old, and (6) have had a commit within the past six months. GHS additionally includes only repositories with at least ten stars. Requiring a license enables downstream open-source filtering. Excluding forks avoids duplicated codebases. The contributor and pull request thresholds exclude toy and solo projects. The age and recency requirements filter both immature and stale repositories.

[1] https://seart-ghs.si.usi.ch/

Table 2: Metadata filtering stages and their impact on dataset size, applied to the baseline GHS query result of 187,304 repositories.

Filtering Step	#Excluded	#Remaining
Initial dataset	–	187,304
(1) Excluded archived, disabled, or locked repositories	2,241	185,063
(2) Excluded repositories without OSI-approved licenses	24,646	160,417
(3) Excluded repositories outside the top-10 languages	31,252	129,165
(4) Excluded educational and resource repositories	1,994	127,171
(5) Excluded repositories with < 352 commits or < 6 watchers	86,331	40,840
(6) Excluded duplicates of redirected repositories	255	**40,585**

The query, executed on 2026-03-16, returned 187,304 repositories. We then applied metadata filtering stages (Table 2): (1) removing archived, disabled, or locked repositories; (2) retaining only OSI-approved software licenses per the SPDX License List; (3) keeping only the ten most common programming languages; (4) excluding educational and resource-aggregation repositories, identified by regular-expression patterns over the repository name and the repository's GitHub topics (as recorded by GHS) that match keywords such as *tutorial*, *course*, *awesome-**, and *list*; (5) applying median-based lower thresholds on commits (≥352) and watchers (≥6) [13]; (6) deduplicating redirected repositories. The full regular expressions are in the pipeline archive's YAML configuration [2]. After filtering, 40,585 repositories remained.

4.2 Repository Classification

We classified the sampled repositories to distinguish those belonging to engineered software projects from non-software or toy repositories. After cloning each repository, we excluded 15 that were unavailable, 74 with missing or empty `README` files, and 1,070 with non-English `README`s (detected using `lingua-language-detector`), leaving 39,426 repositories.

For each repository, we assembled three types of inputs for the classification: (1) the `README` content, (2) a summary of the file structure produced by GitHub Linguist, and (3) summaries of external links referenced in the `README`.

For the file structure, we ran GitHub Linguist on each repository and summarized the output by retaining the top seven languages per repository (ranked by byte count; covering the 75th percentile of the language-per-repository distribution) and up to three representative file paths per directory level, capped at depth eight (where 75% of all files across our sample reside). This keeps the representation informative without exceeding the context window of the model used for classification (see below).

For external links, we extracted all URLs from the `README` files (883,997 links across 39,426 repositories). We retained only those

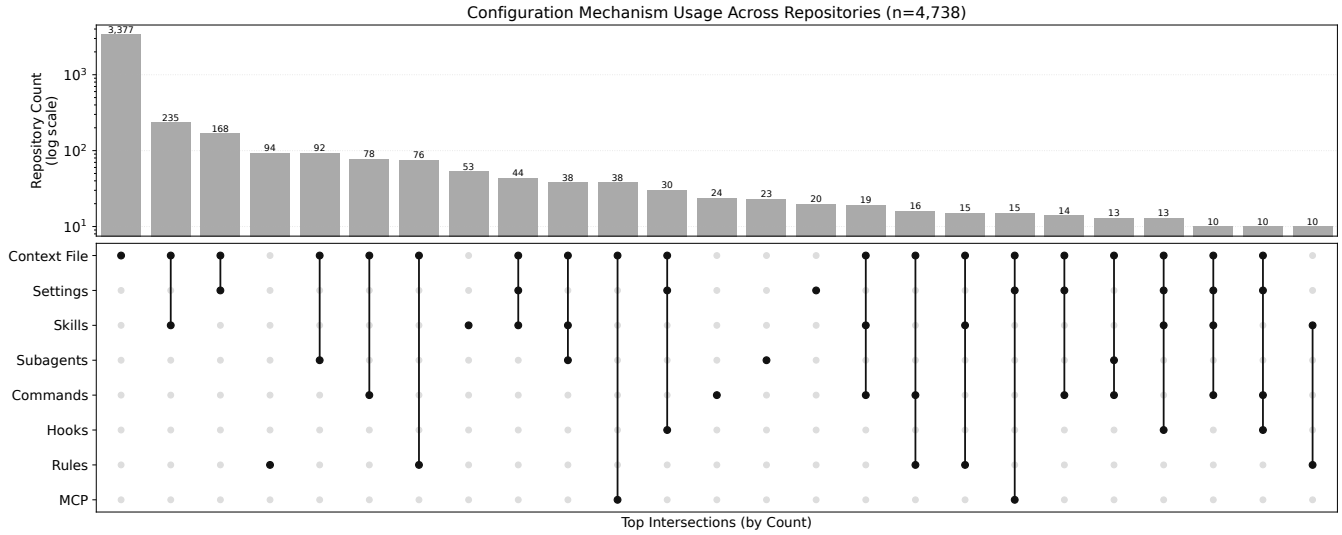


Figure 4: Co-occurrence of configuration mechanisms across repositories. The bar chart shows how many repositories use each combination; the matrix below indicates which mechanisms are included.

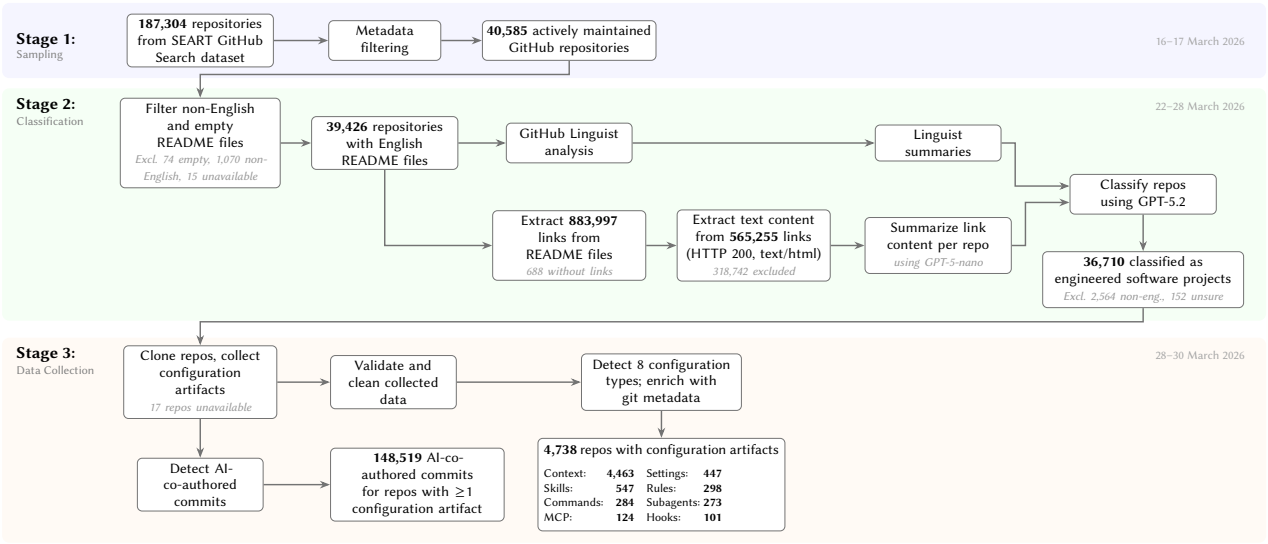


Figure 5: Three-stage data collection pipeline: repository sampling from the SEART GitHub Search dataset, LLM-based classification of engineered software projects, and extraction of AI coding tool configuration artifacts and AI-co-authored commits.

returning HTTP 200 with `text/html` content type (565,255 links) and summarized the first 24 per repository (the 75th percentile of links per repository) using GPT-5-nano. Prana et al. [15] found that links in README files cluster in the "What" and "How" sections that typically appear near the top of the file, so earlier links tend to be more informative about the repository's purpose.

We used GPT-5.2 via the OpenAI Batch API to classify each repository based on these three inputs. The classification prompt extends Munaiah et al.'s [14] definition of engineered software projects with positive indicators (e.g., stated purpose, installation instructions, CI/CD) and explicit exclusion criteria for categories such as tutorials, scaffolding outputs, and code-snippet collections. Two authors spot-checked the assigned labels on a random sample. We piloted both GPT-5.2 and GPT-5-nano; GPT-5.2 showed better adherence to the classification criteria. The exact classification prompt is included in the pipeline archive [2].

The classification labeled 36,710 repositories as belonging to engineered software projects, 2,564 as not belonging to such projects, and 152 as unsure. Compared to our prior study [6], which classified repositories based on README content alone, this improved pipeline

reduced unsure cases from 2,204 to 152. The reduction is attributable to the additional context from file structure summaries and external link summaries, which helped resolve borderline cases.

4.3 Configuration Artifact Extraction

We selected five agentic AI coding tools based on the 2025 Stack Overflow Developer Survey [18]: Claude Code, GitHub Copilot, OpenAI Codex, Cursor, and Gemini. The survey asked developers which AI agent tools they use; we selected the four most popular tools and substituted Codex for ChatGPT, as Codex is the agentic coding tool by the same vendor. We included Cursor given its prominence in recent software engineering research [8, 10].

For each tool, we documented the repository-level files and directories that indicate its presence and developed detection heuristics based on this documentation [6]. We detect the eight configuration mechanisms introduced in Section 1; example artifacts include `CLAUDE.md`, `AGENTS.md`, and `.cursorrules` for Context Files. For GitHub Copilot, Cursor, and Gemini, the detected files apply to both their conversational and agentic interfaces.

We cloned the 36,710 repositories belonging to engineered software projects and applied these heuristics to extract configuration artifacts. After validation and cleaning (removing invalid file paths, recalculating detection flags, and excluding repositories where no valid artifacts remained), 4,738 repositories contained at least one artifact. We enriched each artifact with git metadata: creation date, total commit count, and first and last commit hashes. For Context Files, we additionally detected the natural language, identified files that reference other files, and attributed AI authorship at the commit level.

To assess the extent of active AI tool use in these repositories, we scanned each repository's full commit history for AI-co-authored commits. Detection uses regex-based heuristics across three scopes, applied to every commit reachable from any branch. First, *author identity* patterns match tool-specific bot accounts in the author and committer name and email fields (e.g., `copilot-swe-agent[bot]`, `claude[bot]`). Second, *git trailer* patterns match structured attribution lines such as `Co-authored-by:` and `Generated-by:`, extracted by walking the commit body in reverse to isolate the trailer block from the message body. Third, *message body* patterns match attribution phrases and tool-specific branch prefixes in merge commits. The patterns cover five tools (Claude, Copilot, Cursor, Codex, and Gemini) and are documented with unit tests in the pipeline archive [2]. This produced 148,519 AI-co-authored commits across 3,392 of the 4,738 configured repositories.

4.4 Dataset Validation

Four independent signals from the pipeline's outputs indicate that the detected artifacts correspond to real AI-coding configurations. First, the sampling frame already excludes non-software populations. Repositories must have one of ten mainstream programming languages (C, C#, C++, Go, Java, JavaScript, PHP, Python, Rust, TypeScript) as their primary language, and the GPT-5.2 classification further filters to repositories belonging to engineered software projects (see Section 4.2). Files such as `AGENTS.md` also appear outside coding-agent contexts (e.g., in agent-framework experiments or prompt libraries), but such uses concentrate in repositories that these filters exclude. Second, 71.6% of the 4,738 configured repositories (3,392) contain at least one AI-co-authored commit, a lower bound because commit detection requires explicit attribution signals that not every tool or workflow leaves behind. Third, each commit is checked against three independent signals: the author identity, git trailers, and patterns in the commit message body. Author-identity matches correspond to verified bot accounts such as `copilot-swe-agent[bot]`. Fourth, the prior study [6] manually inspected every short configuration file (≤10 lines) in its own 2,853-repo sample and found that 87.8% (396 of 451) contained substantive configuration content. The per-mechanism breakdown is preserved in the pipeline archive [2]. Per-repository justification texts are shipped with the dataset so users can re-audit any label.

5 Use Cases

The dataset supports a range of empirical research questions, each illustrated below with the data fields that enable it.

Tool adoption and evolution. Each configuration artifact in the dataset carries a `created_at` timestamp and commit count, and each AI-co-authored commit has a timestamp. Researchers can use these fields to study when repositories adopted specific tools, in what order they added configuration mechanisms, and whether tool adoption correlates with changes in commit activity. Figure 6 shows the cumulative number of repositories adopting each artifact type, illustrating the steep growth of `CLAUDE.md` and `AGENTS.md` from late 2024 and the later adoption of Skills and Subagents.

Configuration practices. The raw configuration files in `repos_data.7z` allow qualitative and quantitative content analysis. What instructions do developers give AI agents? Do configuration patterns differ by programming language or repository size? How do configuration files for different tools compare in length, structure, and content? Researchers can also compare how the eight configuration mechanisms are combined within repositories: our prior study [6] found that most repositories use only Context Files, while advanced mechanisms such as Skills and Subagents remain rare.

The Context File metadata records whether a file is a reference to another file (e.g., a `CLAUDE.md` that simply points to a shared `AGENTS.md`). Figure 7 shows the resulting reference network: `AGENTS.md` is the dominant reference target, receiving incoming references from tool-specific context files of all five tools. This pattern indicates that `AGENTS.md` is adopted across tools as a tool-agnostic standard that tool-specific files reference.

AI-co-authored commits. The `commits.csv` file enables analysis of AI-co-authored commits: their frequency, timing, commit message style, and distribution across repositories. Figure 8 shows the corresponding growth in AI-co-authored commits, with monthly volumes rising from near zero in early 2025 to over 40,000 by March 2026. The rise coincides with the availability of agentic features in Claude Code and GitHub Copilot, and suggests that configuration artifact adoption and active AI tool usage are closely linked.

Each commit record includes the full commit message and the detected AI tool, allowing researchers to compare the writing style and scope of AI-generated commits with human-authored ones. Researchers can also measure what fraction of a repository's total commits are AI-co-authored, and how this fraction changes over time.

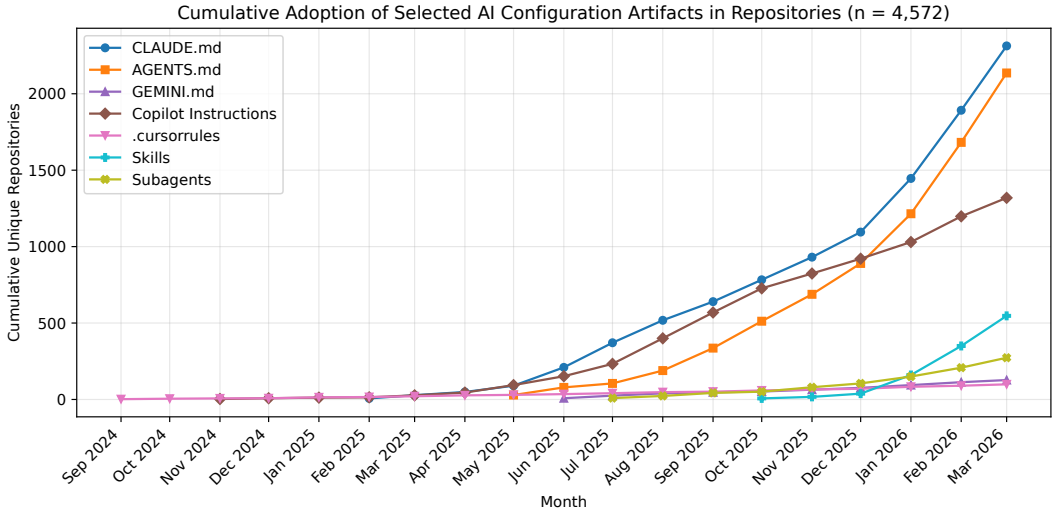


Figure 6: Cumulative adoption of selected AI configuration artifacts over time. Each curve shows the cumulative number of repositories in which the artifact type was first observed.

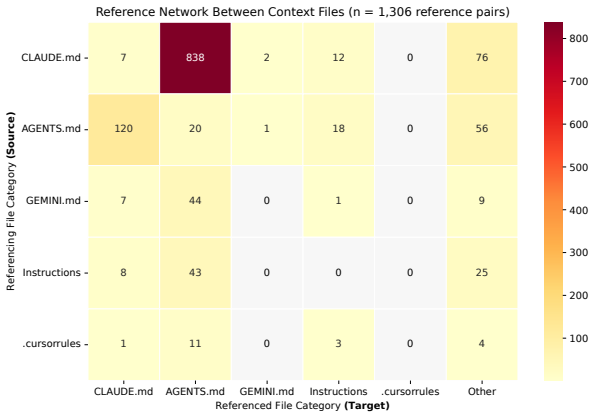


Figure 7: Reference network between context file types. Each cell shows how many files of the source type (row) reference a file of the target type (column).

Configuration authorship evolution. Researchers can join `repos_data.7z` (raw file contents at HEAD) with `commits.csv` and each configuration artifact's first and last commit SHAs to trace whether a `CLAUDE.md`, `AGENTS.md`, or other artifact was originally authored by a human or by an AI tool and how it evolved afterwards. The `context_files.csv` metadata already surfaces the first-commit author and a per-file AI-commit count; `git blame` against the stored files yields line-level attribution. This supports questions such as: What fraction of typical `CLAUDE.md` content is human-written versus AI-written? Do AI-authored additions focus on mechanical details (build commands, test runners) while humans write architectural guidance? Which configuration mechanisms are most often initialized by humans and later extended by AI tools, and which ones are AI-initiated from the start?

Repository characteristics. Because `repos.csv` covers both tool-adopting and non-adopting repositories with full GitHub metadata, researchers can study what distinguishes repositories that adopt AI coding tools, using metrics such as repository size, contributor count, language, and activity level. The classification justifications in `repos.csv` can also serve as training data for automated repository classification approaches.

Governance and safety. The raw configuration files encode how developers constrain AI agent behavior: which files agents may modify, which actions require human approval, and what security or coding practices agents must follow. Researchers can study how governance and safety constraints are expressed across repositories, identify under-specified areas, and detect risks such as ambiguity, over-delegation, or missing safeguards. Repositories that configure multiple tools simultaneously allow studying whether instructions across artifacts are consistent or contradictory.

Model training and evaluation. The 18,167 configuration files can serve as a corpus for training or fine-tuning language models on developer-written instructions for AI agents. The commit-level AI authorship metadata lets researchers filter for human-authored configurations when building training sets. The per-artifact metadata (creation date, commit count, language) supports stratified sampling across these dimensions.

6 Reproducibility and Availability

The dataset is archived on Zenodo under CC BY 4.0 [1]. It includes all CSV files listed in Table 1 and the `repos_data.7z` archive containing the raw text of all 18,167 collected configuration files. Large files (`repos.csv` and `commits.csv`) are additionally distributed as 7z archives for faster download. Users can load the data directly with standard tools and join across files:

```
import pandas as pd
repos = pd.read_csv("repos.csv")
```

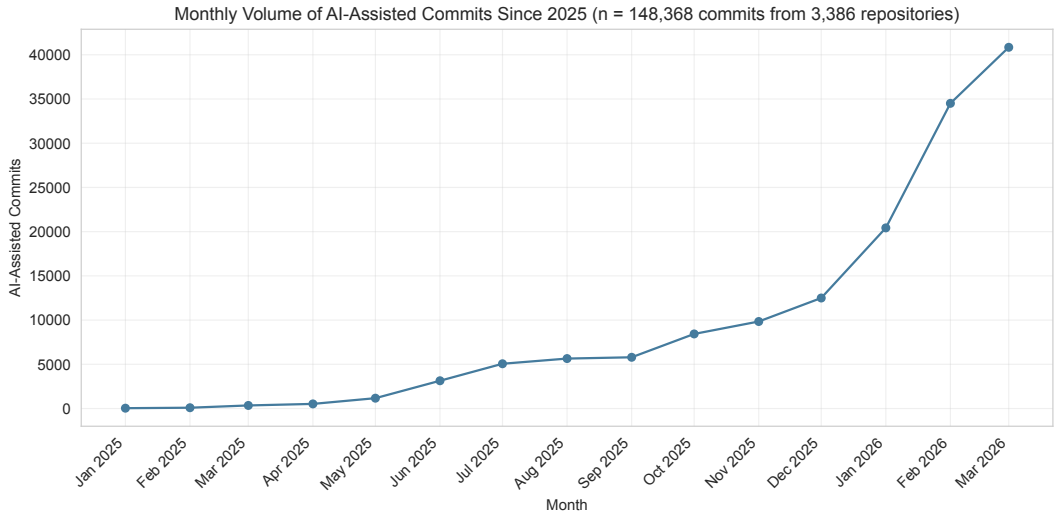


Figure 8: Monthly volume of AI-co-authored commits across the 4,738 repositories with detected configuration artifacts.

```
configured = repos[repos["scanned_at"].notna()]
claude_repos = configured[configured["claude"] ==
    True]
commits = pd.read_csv("commits.csv")
claude_commits = commits[commits["ai_tool"].str.
    contains("Claude")]
```

The complete construction pipeline (all Python scripts and YAML configuration files across the three stages shown in Figure 5) is archived separately on Zenodo [2]. The pipeline is organized into four directories matching the methodology sections. Each stage reads from the previous stage's output and produces self-contained results. The sampling stage is driven by a declarative YAML configuration file that specifies all filter parameters, making queries reproducible given the same GHS snapshot. The classification stage depends on the OpenAI API (GPT-5.2 and GPT-5-nano), introducing potential non-determinism across runs; we include the classification justifications in the dataset so that users can inspect and verify individual labels. The artifact extraction stage uses deterministic file-matching heuristics documented in a versioned specification, with unit tests for the AI commit detection logic.

An interactive website[2] provides a searchable interface for browsing repositories, viewing configuration artifacts, exploring AI-co-authored commits, and consulting usage examples without downloading the full dataset.

7 Limitations

As with any large-scale mined dataset, this dataset has limitations stemming from sampling decisions, automated classification, and heuristic-based extraction.

The dataset covers only public GitHub repositories and does not include proprietary codebases, where agentic AI coding tools may be adopted at a larger scale and under different organizational constraints. AI tool adoption on other platforms (GitLab, Bitbucket) may follow different patterns; extending the pipeline to other platforms would require adapting the repository sampling and cloning stages.

LLM-based classification of engineered projects introduces non-determinism: re-running GPT-5.2 on the same input may produce slightly different labels. We include justification texts in the dataset, and two authors spot-checked labels on a random sample. The 152 unsure repositories are included in `repos.csv` with their labels, so users can apply their own inclusion criteria.

Tool detection relies on file and directory naming conventions documented by each tool's vendor. Repositories that use non-standard locations or that removed configuration files before our March 2026 snapshot are missed. The detection heuristics are versioned in the pipeline archive and can be extended as tools evolve.

AI-co-authored commit detection depends on identifiable markers in author fields, git trailers, and commit messages. Tools or workflows that do not leave such traces produce false negatives, meaning the true volume of AI-assisted contributions is likely underestimated. Some tools do not expose consistent attribution signals, which points to the need for more systematic approaches to identify AI-authored commits regardless of whether explicit attribution is present. The detection patterns are documented and extensible. The dataset reflects a single point-in-time snapshot (March 2026). We designed the pipeline for periodic re-execution with updated data.

8 Conclusion

We present a curated dataset of agentic AI coding tool configurations from 4,738 open-source GitHub repositories, covering five tools, eight configuration mechanisms, 15,591 configuration artifacts with their full content, and 148,519 AI-co-authored commits. The dataset and the construction pipeline are publicly available on Zenodo. An interactive website provides detailed usage examples and an alternative entry point for browsing and exploring the dataset without downloading it.

[2] https://se-uhd.de/ai-config-dataset/

We intend to continuously update the dataset as new tools and configuration mechanisms are introduced, expanding the detection heuristics accordingly. Future versions will also improve the classification pipeline by incorporating quantitative repository metadata (stars, forks, activity levels) and the GitHub repository description alongside the `README` content, which should further reduce the number of unsure cases. Beyond maintenance, future work can link configuration practices to downstream outcomes such as code quality and review efficiency.

References

[1] Sebastian Baltes, Seyedmoein Mohsenimofidi, Levi Böhme, Jai Lal Lulla, Muhammad Auwal Abubakar, Christoph Treude, and Matthias Galster. 2026. A Dataset of Agentic AI Coding Tool Configurations. doi:10.5281/zenodo.19375880

[2] Sebastian Baltes, Seyedmoein Mohsenimofidi, Levi Böhme, Jai Lal Lulla, Muhammad Auwal Abubakar, Christoph Treude, and Matthias Galster. 2026. A Dataset of Agentic AI Coding Tool Configurations (Pipeline). doi:10.5281/zenodo.19375429

[3] Worawalan Chatlatanagulchai, Hao Li, Yutaro Kashiwa, Brittany Reid, Kundjanasith Thonglek, Pattara Leelaprute, Arnon Rungsawang, Bundit Manaskasemsak, Bram Adams, Ahmed E. Hassan, and Hajimu Iida. 2025. Agent READMEs: An Empirical Study of Context Files for Agentic Coding. arXiv:2511.12884 [cs.SE] doi:10.48550/arXiv.2511.12884

[4] Ozren Dabic, Emad Aghajani, and Gabriele Bavota. 2021. Sampling Projects in GitHub for MSR Studies. In *18th IEEE/ACM International Conference on Mining Software Repositories, MSR 2021, Madrid, Spain, May 17-19, 2021*. IEEE, Madrid, Spain, 560–564. doi:10.1109/MSR52588.2021.00074

[5] Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2024. Self-Collaboration Code Generation via ChatGPT. *ACM Trans. Softw. Eng. Methodol.* 33, 7 (2024), 189:1–189:38. doi:10.1145/3672459

[6] Matthias Galster, Seyedmoein Mohsenimofidi, Jai Lal Lulla, Muhammad Auwal Abubakar, Christoph Treude, and Sebastian Baltes. 2026. Configuring Agentic AI Coding Tools: An Exploratory Study. arXiv:2602.14690 [cs.SE] doi:10.48550/arXiv.2602.14690 To appear at the 3rd ACM International Conference on AI-powered Software (AIware 2026).

[7] Ahmed E. Hassan, Hao Li, Dayi Lin, Bram Adams, Tse-Hsun Chen, Yutaro Kashiwa, and Dong Qiu. 2025. Agentic Software Engineering: Foundational Pillars and a Research Roadmap. arXiv:2509.06216 [cs.SE] doi:10.48550/arXiv.2509.06216

[8] Hao He, Courtney Miller, Shyam Agarwal, Christian Kästner, and Bogdan Vasilescu. 2026. Speed at the Cost of Quality: How Cursor AI Increases Short-Term Velocity and Long-Term Complexity in Open-Source Projects. arXiv:2511.04427 [cs.SE] doi:10.48550/arXiv.2511.04427 To appear at the 23rd IEEE/ACM International Conference on Mining Software Repositories (MSR 2026).

[9] Kosei Horikawa, Hao Li, Yutaro Kashiwa, Bram Adams, Hajimu Iida, and Ahmed E. Hassan. 2025. Agentic Refactoring: An Empirical Study of AI Coding Agents. arXiv:2511.04824 [cs.SE] doi:10.48550/arXiv.2511.04824

[10] Shaokang Jiang and Daye Nam. 2026. Beyond the Prompt: An Empirical Study of Cursor Rules. arXiv:2512.18925 [cs.SE] doi:10.48550/arXiv.2512.18925 To appear at the 23rd IEEE/ACM International Conference on Mining Software Repositories (MSR 2026).

[11] Eirini Kalliamvakou, Georgios Gousios, Kelly Blincoe, Leif Singer, Daniel M. Germán, and Daniela E. Damian. 2014. The promises and perils of mining GitHub. In *11th Working Conference on Mining Software Repositories, MSR 2014, Proceedings, May 31 - June 1, 2014, Hyderabad, India*, Premkumar T. Devanbu, Sung Kim, and Martin Pinzger (Eds.). ACM, Hyderabad, India, 92–101. doi:10.1145/2597073.2597074

[12] Hao Li, Haoxiang Zhang, and Ahmed E. Hassan. 2026. AIDev: Studying AI Coding Agents on GitHub. arXiv:2602.09185 [cs.SE] doi:10.48550/arXiv.2602.09185

[13] Seyedmoein Mohsenimofidi, Matthias Galster, Christoph Treude, and Sebastian Baltes. 2026. Context Engineering for AI Agents in Open-Source Software. arXiv:2510.21413 [cs.SE] doi:10.48550/arXiv.2510.21413 To appear at the 23rd IEEE/ACM International Conference on Mining Software Repositories (MSR 2026).

[14] Nuthan Munaiah, Steven Kroh, Craig Cabrey, and Meiyappan Nagappan. 2017. Curating GitHub for engineered software projects. *Empir. Softw. Eng.* 22, 6 (2017), 3219–3253. doi:10.1007/S10664-017-9512-6

[15] Gede Artha Azriadi Prana, Christoph Treude, Ferdian Thung, Thushari Atapattu, and David Lo. 2019. Categorizing the Content of GitHub README Files. *Empir. Softw. Eng.* 24, 3 (2019), 1296–1327. doi:10.1007/S10664-018-9660-3

[16] Romain Robbes, Théo Matricon, Thomas Degueule, André C. Hora, and Stefano Zacchiroli. 2026. Agentic Much? Adoption of Coding Agents on GitHub. arXiv:2601.18341 [cs.SE] doi:10.48550/arXiv.2601.18341

[17] Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. 2025. AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges. arXiv:2505.10468 [cs.AI] doi:10.48550/arXiv.2505.10468

[18] Stack Exchange Inc. 2025. Stack Overflow Developer Survey 2025: AI Agent out-of-the-box tools. https://survey.stackoverflow.co/2025/ai/#3-ai-agent-out-of-the-box-tools.

[19] Miku Watanabe, Hao Li, Yutaro Kashiwa, Brittany Reid, Hajimu Iida, and Ahmed E. Hassan. 2025. On the Use of Agentic Coding: An Empirical Study of Pull Requests on GitHub. arXiv:2509.14745 [cs.SE] doi:10.48550/arXiv.2509.14745

[20] Tao Xiao, Youmei Fan, Fabio Calefato, Christoph Treude, Raula Gaikovina Kula, Hideaki Hata, and Sebastian Baltes. 2025. Self-Admitted GenAI Usage in Open-Source Software. arXiv:2507.10422 [cs.SE]

[21] Yangtian Zi, Zixuan Wu, Aleksander Boruch-Gruszecki, Jonathan Bell, and Arjun Guha. 2025. AgentPack: A Dataset of Code Changes, Co-Authored by Agents and Humans. arXiv:2509.21891 [cs.SE]

Received 2026-02-15; accepted 2026-03-28