Evaluating AGENTS.md: Are Repository-Level Context Files Helpful for Coding Agents?

Title: Evaluating AGENTS.md: Are Repository-Level Context Files Helpful for Coding Agents?

Authors: Thibaud Gloaguen, Niels Mündler (ETH Zurich), Veselin Raychev, Mark Müller, Martin Vechev (LogicStar.ai / ETH Zurich)

Published: June 2026 — arXiv:2602.11988

Link: https://arxiv.org/abs/2602.11988

Summary

A widespread practice in software development is to tailor coding agents to repositories using context files (AGENTS.md, CLAUDE.md), strongly encouraged by agent developers. This paper presents the first rigorous evaluation of whether such context files actually improve task completion performance.

Key findings:

- **LLM-generated context files do NOT improve success rates** — they slightly reduce them (−0.5% on SWE-Bench, −2% on CTX-Bench), though not statistically significantly
- **They increase inference cost by over 20%** on average, as agents follow instructions and run more tests/exploration
- **Developer-written context files outperform LLM-generated ones** by ~7%, but the improvement over having no file is marginal (p=21%)
- **Repository overviews are not helpful** — agents don't discover relevant files faster with an overview section
- **Instructions ARE followed** — if a tool is mentioned in the context file, usage increases dramatically. The cost increase comes from more thorough testing, not ignoring

instructions

The authors created **CTX BENCH**, a new benchmark of 138 real-world Python tasks from 12 niche open-source repos with developer-committed context files, complementing SWE-Bench Lite evaluation.

Conclusion: Context files should only contain non-standard coding practices not already in the README. LLM-generated ones are not worth it yet. Any attempts to improve performance should be rigorously evaluated before deployment.

Revision #3
Created 2026-08-30 21:50:57 UTC by Clive
Updated 2026-08-30 22:00:19 UTC by Clive