Microsoft FastContext: A Scout So Your Coding Agent Stops Burning Tokens

Kabui, Charles

2026-06-29

Read at ToKnow.ai

Microsoft released FastContext, a small helper model that does one job for an AI coding agent: find the right code in a large repository. Normally the same model both hunts for relevant files and writes the fix, so dozens of reads and searches pile into its limited working memory (the context window) and crowd out the real task. FastContext splits the two apart. A specialized explorer, built in sizes from 4B to 30B parameters, runs read-only searches in parallel and

hands back only file paths and line ranges, not whole files. Dropped into the Mini-SWE-Agent coding loop across three SWE-bench tests, it raised end-to-end fix rates up to 5.5% while cutting the main agent's token use up to 60.3%.

Tokens are the hidden bill on every agent task, and crawling a repo is where most of them go. By moving search to a cheap 4B model that can run locally through Ollama, a developer keeps the expensive frontier model focused on solving, pays less, and gets more bugs fixed. The explorer touches nothing, using only read and grep tools, so it cannot break anything while it looks.

It also marks a clear direction: instead of one giant model doing everything, agents are splitting into specialists trained for narrow jobs. We previously covered the same instinct in CodeGraph, which indexes a repo so agents read fewer tokens.

Sources:

- FastContext: Training Efficient Repository Explorer for Coding Agents (arXiv)
- microsoft/fastcontext on GitHub
- FastContext paper page on Hugging Face
- SWE-bench
