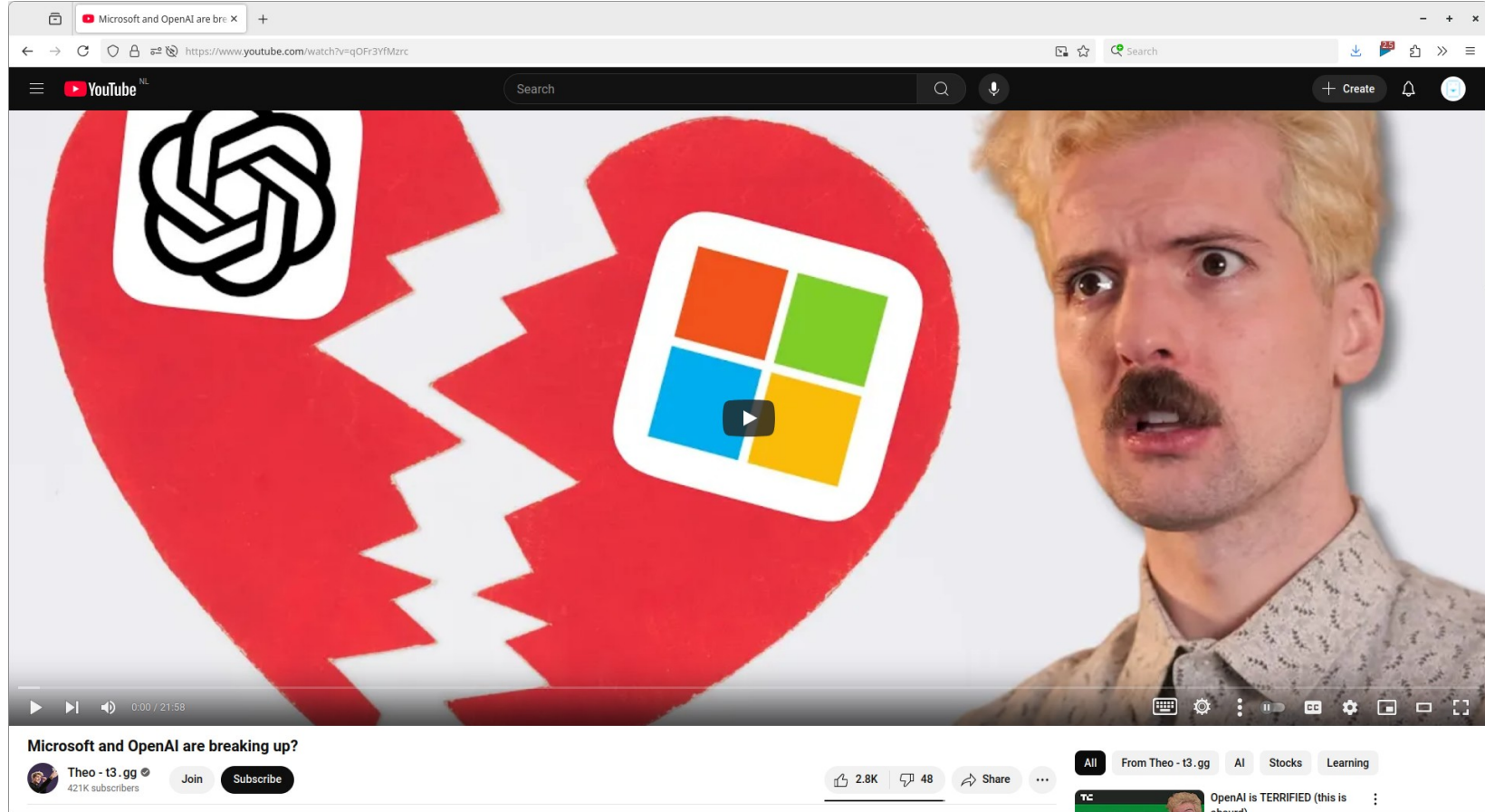


Motivation

- AI Code suggestions can be very useful
 - But the AIs for tools like Copilot and CodeWhisperer are run on the servers of big corporations
 - Big corporations that have a track record of not respecting copyright laws (to build their AI)
 - Where you work may not feel comfortable sending (important) company code to get code suggestions!

AI Commoditization



Llama Coder

The screenshot shows the Visual Studio Code (VS Code) interface. On the left, the 'EXTENSIONS' sidebar is open, displaying a list of installed and recommended extensions. The 'Llama Coder' extension by Steve Korshakov is highlighted in the installed list. The main editor area shows the 'Extension: Llama Coder' details page. The extension's icon is a purple circle with a white llama. The title 'Llama Coder' is prominently displayed, followed by the author 'Steve Korshakov', a download count of 56,772, and a 5-star rating from 2 reviews. A description states it is a 'Better and self-hosted Github Copilot replacement'. Below this are buttons for 'Disable', 'Uninstall', and a checked 'Auto Update' option. The 'DETAILS' tab is selected, showing a description of the extension as a 'better and self-hosted Github Copilot replacement for VS Code' that uses 'Ollama' and 'codellama'. It also includes a 'VS Code Plugin' link. The 'FEATURES' section lists three bullet points: 'As good as Copilot', 'Fast. Works well on consumer GPUs. Apple Silicon or RTX 4090 is recommended for best performance.', and 'No telemetry or tracking'. The 'RECOMMENDED hardware' section specifies a minimum of 16GB RAM and provides instructions for installation on different hardware. The 'LOCAL Installation' section mentions installing Ollama locally. On the right side of the details page, there are sections for 'Installation' (with a table of metadata), 'Marketplace' (with publication and release dates), 'Categories' (Machine Learning, Programming Languages), and 'Resources' (Marketplace, Issues, Repository, License, and the author's profile).

EXTENSIONS

Search Extensions in Marketplace

INSTALLED (53)

- docker** (Microsoft) - Makes it easy to create, manage, and de...
- Extension Pack for Java** (Microsoft) - Popular extensions for Java development...
- Gradle for Java** (Microsoft) - Manage Gradle Projects, run Gradle task...
- Language Support for Java(TM)** (Red Hat) - Java Linting, Intellisense, formatting, refa...
- Llama Coder** (Steve Korshakov) - Better and self-hosted Github Copilot re...
- Lombok Annotations Support f...** (Microsoft) - Refactor code with Lombok annotations, ...
- Maven for Java** (Microsoft) - Manage Maven projects, execute goals, ...
- npm Intellisense**

RECOMMENDED (7)

- Dev Containers** (Microsoft) - Open any folder or repository inside a D...
- markdownlint** (David Anson) - Markdown linting and style checking for ...
- Git History** - View git log, file history, compare branch...

Extension: Llama Coder | Settings

Llama Coder
Steve Korshakov | 56,772 | ★★★★★ (2)
Better and self-hosted Github Copilot replacement

Disable | Uninstall | Auto Update

DETAILS | FEATURES

Llama Coder

Llama Coder is a better and self-hosted Github Copilot replacement for [VS Code](#). Llama Coder uses [Ollama](#) and [codellama](#) to provide autocomplete that runs on your hardware. Works best with Mac M1/M2/M3 or with RTX 4090.

[VS Code Plugin](#)

Features

- As good as Copilot
- Fast. Works well on consumer GPUs. Apple Silicon or RTX 4090 is recommended for best performance.
- No telemetry or tracking
- Works with any language coding or human one.

Recommended hardware

Minimum required RAM: 16GB is a minimum, more is better since even smallest model takes 5GB of RAM. The best way: dedicated machine with RTX 4090. Install [Ollama](#) on this machine and configure endpoint in extension settings to offload to this machine. Second best way: run on MacBook M1/M2/M3 with enough RAM (more == better, but 10gb extra would be enough). For windows notebooks: it runs good with decent GPU, but dedicated machine with a good GPU is recommended. Perfect if you have a dedicated gaming PC.

Local Installation

Install [Ollama](#) on local machine and then launch the extension in VSCode, everything should work as it is.

Installation

| | |
|--------------|----------------------|
| Identifier | ex3ndr.llama-coder |
| Version | 0.0.14 |
| Last Updated | 2025-03-29, 12:03:32 |
| Size | 182.25KB |

Marketplace

| | |
|---------------|----------------------|
| Published | 2023-11-21, 03:40:50 |
| Last Released | 2024-04-07, 19:27:36 |

Categories

- Machine Learning
- Programming Languages

Resources

- [Marketplace](#)
- [Issues](#)
- [Repository](#)
- [License](#)
- [Steve Korshakov](#)

Settings

The screenshot shows the Visual Studio Code interface with the 'Settings' window open for the 'Llama Coder' extension. The left sidebar lists various settings categories, with 'Extensions' expanded and 'Llama coder' selected. The main panel displays the 'Extensions' settings for 'Language Support for Java(TM) by Red Hat'. Under the 'Llama coder' section, there are three inference settings: 'Endpoint', 'Model', and 'Temperature'. The 'Endpoint' is a text input field. The 'Model' is a dropdown menu set to 'stable-code:3b-code-q4_0'. The 'Temperature' is a text input field set to '0.2'. Below these, there are sections for 'Custom: Model' and 'Custom: Format', both with dropdown menus. The 'Custom: Model' dropdown is empty, and the 'Custom: Format' dropdown is set to 'stable-code'. The status bar at the bottom shows 'main' and 'java: Ready'.

File Edit Selection View Go Run Terminal Help ← → SpringAI

Extension: Llama Coder Settings x

Search settings

User Workspace Last synced: 32 secs ago

> Features
> Application
> Security
▼ Extensions
 .ipynb Support
 Cloudfoundry Manifest Language Server C...
> CSS Language Features
 Docker
 Emmet
 Extension Pack for Java
 Firefox debug
 Git
 GitHub
 GitHub Enterprise Server Authentication P...
 Gradle
 Grunt
 Gulp
 HTML
 Jake
 Java Debugger
 Java Project Manager
 JavaScript Debugger
 JSON
> Language Support for Java(TM) by Red Hat
 Llama coder
 Markdown
 Markdown Math
 Maven for Java
 Media Previewer
 Merge Conflict
> Microsoft Account

Extensions

Language Support for Java(TM) by Red Hat

Llama coder

Inference: Endpoint
Ollama Server Endpoint. Empty for local instance. Example: `http://192.168.0.100:11434`

Inference: Model
Inference model to use
stable-code:3b-code-q4_0

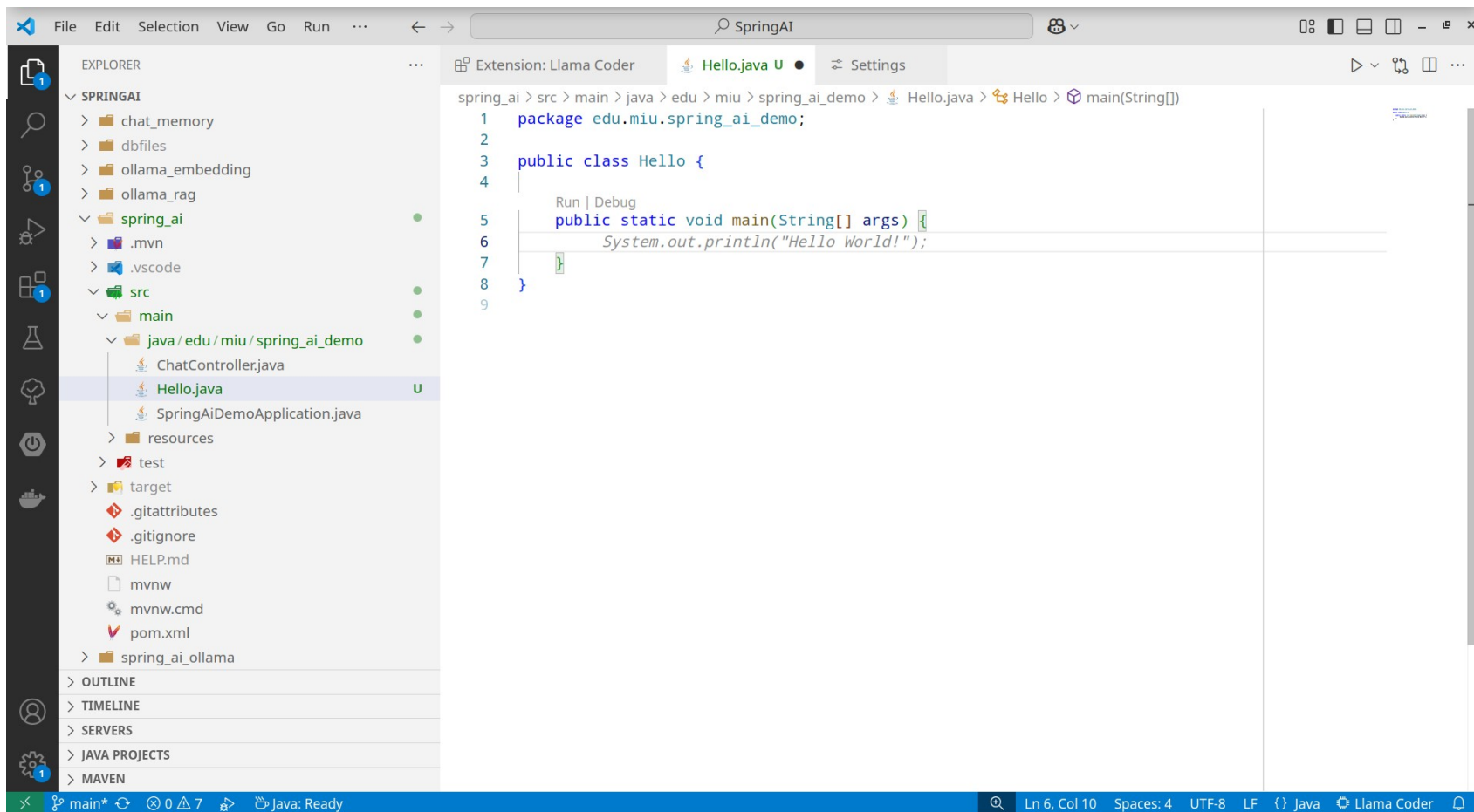
Inference: Temperature
Temperature of the model. Increasing the temperature will make the model answer more creatively.
0.2

Inference > Custom: Model
Custom model name

Inference > Custom: Format
Custom model prompt format
stable-code

main java: Ready Llama Coder

Demo



IntelliJ ProxyAI


Proxy AI Plugin for JetBrains I x

+

← → ↻ 🔒 https://plugins.jetbrains.com/plugin/21056-proxy-ai


🔍 Search

📄 ⬇️ 🗑️ ⌵ ≡

 **JETBRAINS Marketplace**

Edu Courses Themes Plugin Ideas Build Plugins Sign In ? 🔍

Code Tools Fun Stuff Code Editing +2 more



Proxy AI

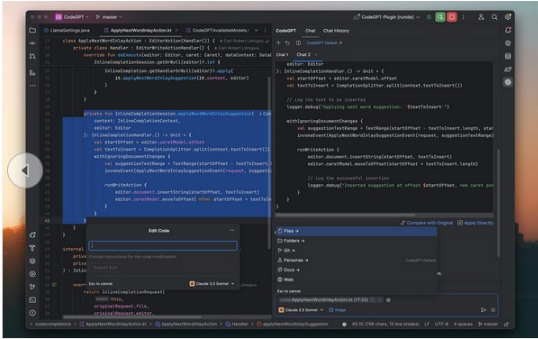
★★★★☆

Proxy OÜ

Overview Versions Reviews

Get

Compatible with IntelliJ IDEA (Ultimate, Community), Android Studio and 17 more



Multi-line edits

Get edits and multi-line changes based on your recent changes.

```
interface FancyButtonProps {
  text: string;
  icon?: React.ReactNode;
  onClick?: () => void;
  size?: 'sm' | 'md' | 'lg';
  isLoading?: boolean;
}

const FancyButton: React.FC<FancyButtonProps> = ({
  text,
  icon,
  onClick,
```

ProxyAI is an AI-powered code assistant designed to help you with various programming activities. It's a powerful alternative to GitHub Copilot, AI Assistant, Codiumate, and other JetBrains plugins.

Highly Configurable

Access top-tier language models (LLMs) with your own API key or use privately hosted models within your corporate network.

- **Cloud Providers & Custom Setups:** Integrate top-tier models from OpenAI, Anthropic, Azure, Mistral, or use self-hosted models for offline use.

