Copilot CLI

- Install
- Usage with Ollama
  - Quick setup
  - Run directly with a model
- Non-interactive (headless) mode
- Manual setup

GitHub Copilot CLI is GitHub's AI coding agent for the terminal. It can understand your codebase, make edits, run commands, and help you build software faster.

Open models can be used with Copilot CLI through Ollama, enabling you to use models such as `qwen3.5`, `glm-5.3:cloud`, `kimi-k3:cloud` See all supported models here.

Install

Install Copilot CLI:

```
brew install copilot-cli
```

```
npm install -g @github/copilot
```

```
curl -fsSL https://gh.io/copilot-install | bash
```

```
winget install GitHub.Copilot
```

Usage with Ollama

Quick setup

```
ollama launch copilot
```

Run directly with a model

```
ollama launch copilot --model kimi-k2.5:cloud
```

Non-interactive (headless) mode

Run Copilot CLI without interaction for use in Docker, CI/CD, or scripts:

```
ollama launch copilot --model kimi-k2.5:cloud --yes -- -p "how does this repository work?"
```

The `--yes` flag auto-pulls the model, skips selectors, and requires `--model` to be specified. Arguments after `--` are passed directly to Copilot CLI.

Manual setup

Copilot CLI connects to Ollama using the OpenAI-compatible API via environment variables.

1. Set the environment variables:

```
export COPILOT_PROVIDER_BASE_URL=http://localhost:11434/v1
export COPILOT_PROVIDER_API_KEY=
export COPILOT_PROVIDER_WIRE_API=responses
export COPILOT_MODEL=qwen3.5
```

1. Run Copilot CLI:

```
copilot
```

Or run with environment variables inline:

```
COPILOT_PROVIDER_BASE_URL=http://localhost:11434/v1 COPILOT_PROVIDER_API_KEY=
COPILOT_PROVIDER_WIRE_API=responses COPILOT_MODEL=glm-5:cloud copilot
```

Note: Copilot requires a large context window. We recommend at least 64k tokens. See the context length documentation for how to adjust context length in Ollama.

□□□ #2
□ □□□ □□□ 2026-10-31 13:17:51 UTC
□ □□□ □□□ 2026-10-31 13:17:52 UTC