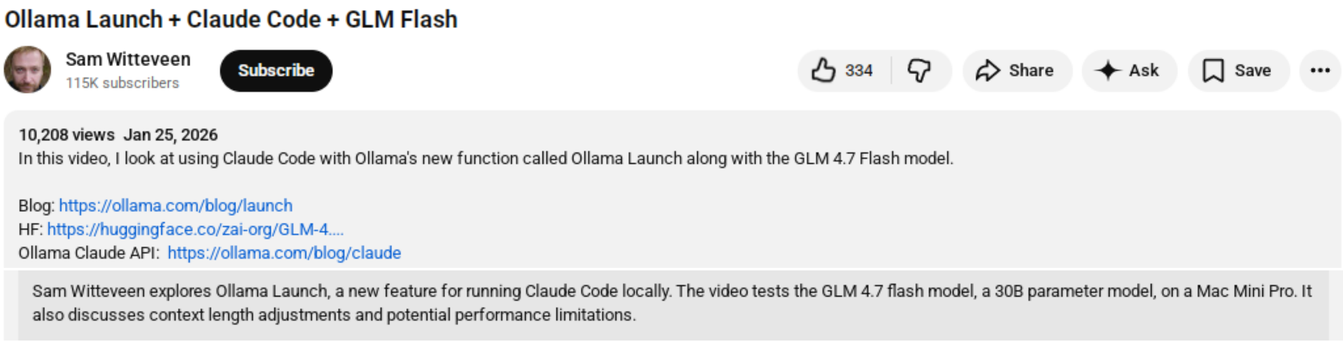


Ollama Launch + Claude Code + GLM Flash

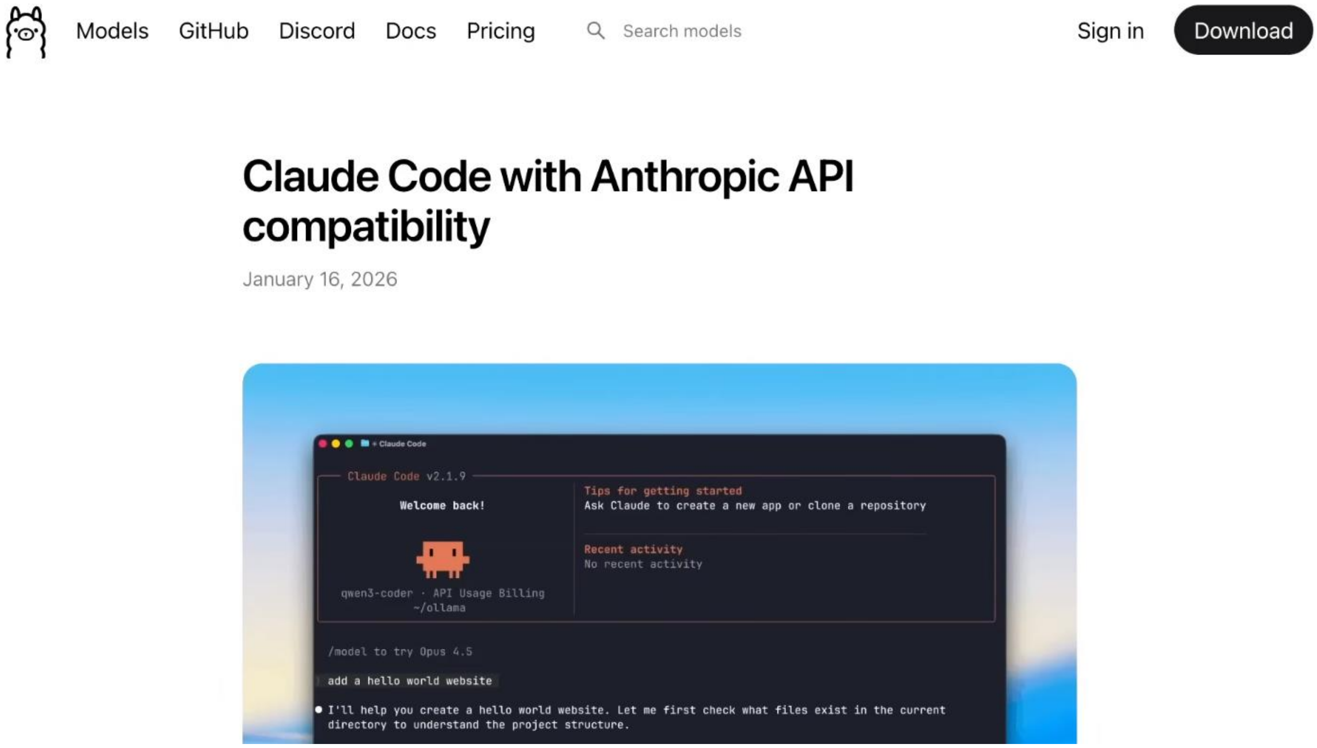
Sam Witteveen
115K subscribers

10,208 views Jan 25, 2026

In this video, I look at using Claude Code with Ollama's new function called Ollama Launch along with the GLM 4.7 Flash model.

Blog: https://ollama.com/blog/launch
HF: https://huggingface.co/zai-org/GLM-4....
Ollama Claude API: https://ollama.com/blog/claude

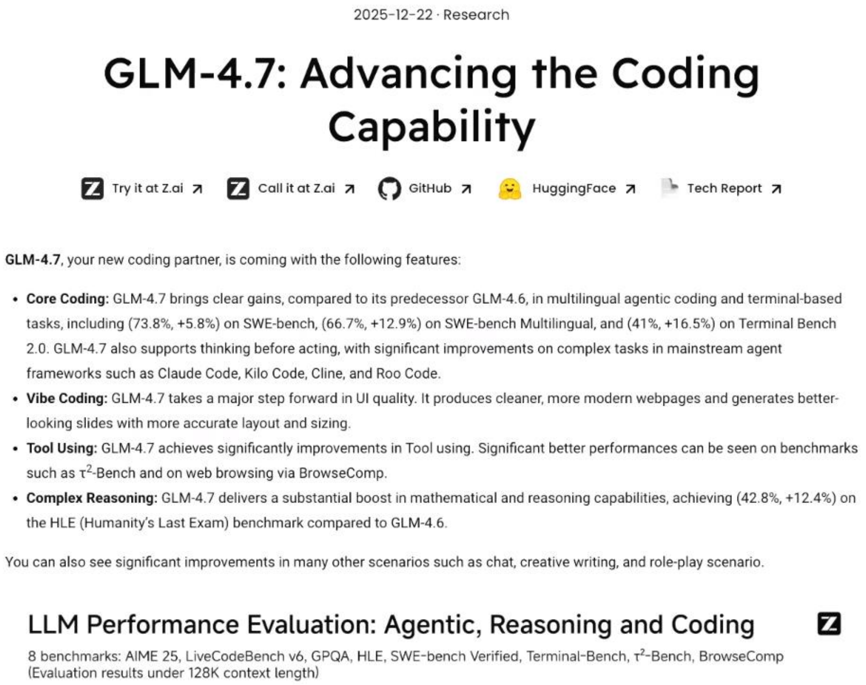
Sam Witteveen explores Ollama Launch, a new feature for running Claude Code locally. The video tests the GLM 4.7 flash model, a 30B parameter model, on a Mac Mini Pro. It also discusses context length adjustments and potential performance limitations.



Claude Code with Anthropic API compatibility

January 16, 2026

```
Claude Code v2.1.9
        Welcome back!              Tips for getting started
                                   Ask Claude to create a new app or clone a repository

                                   Recent activity
                                   No recent activity
 qwen3-coder · API Usage Billing
          ~/ollama

/model to try Opus 4.5

add a hello world website

● I'll help you create a hello world website. Let me first check what files exist in the current
  directory to understand the project structure.
```



2025-12-22 · Research

GLM-4.7: Advancing the Coding Capability

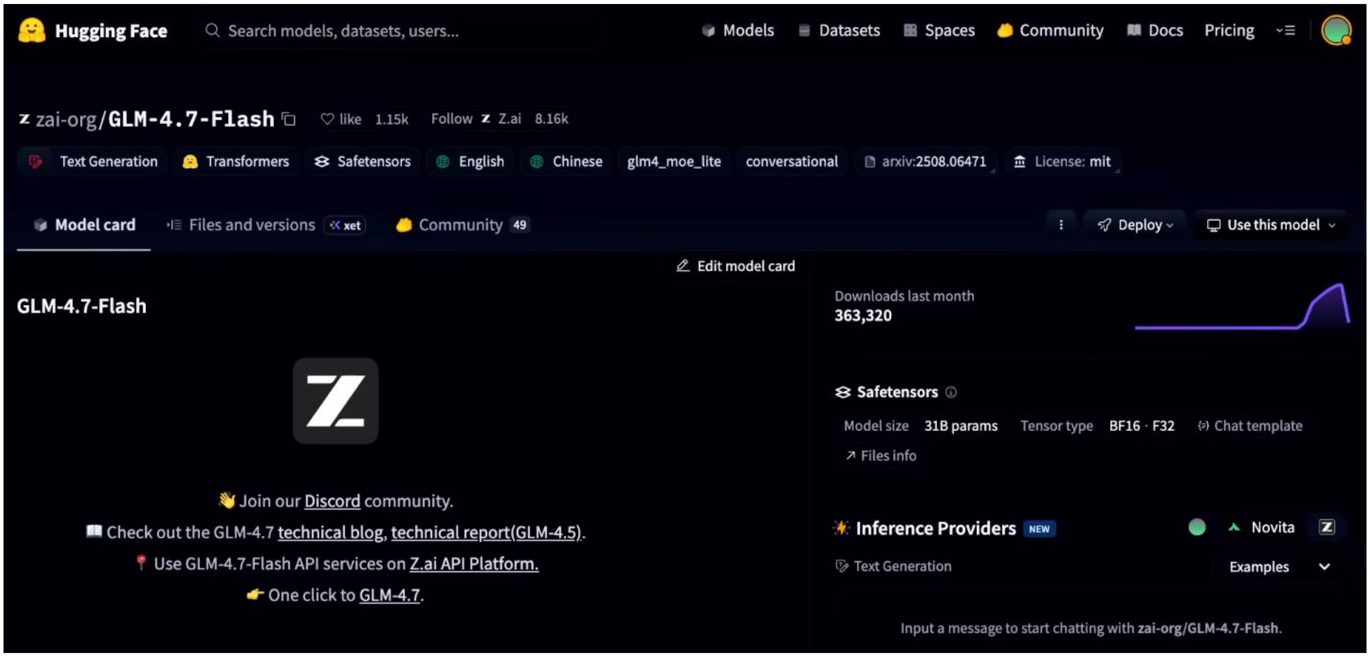
GLM-4.7, your new coding partner, is coming with the following features:

- **Core Coding:** GLM-4.7 brings clear gains, compared to its predecessor GLM-4.6, in multilingual agentic coding and terminal-based tasks, including (73.8%, +5.8%) on SWE-bench, (66.7%, +12.9%) on SWE-bench Multilingual, and (41%, +16.5%) on Terminal Bench 2.0. GLM-4.7 also supports thinking before acting, with significant improvements on complex tasks in mainstream agent frameworks such as Claude Code, Kilo Code, Cline, and Roo Code.
- **Vibe Coding:** GLM-4.7 takes a major step forward in UI quality. It produces cleaner, more modern webpages and generates better-looking slides with more accurate layout and sizing.
- **Tool Using:** GLM-4.7 achieves significantly improvements in Tool using. Significant better performances can be seen on benchmarks such as τ^2-Bench and on web browsing via BrowseComp.
- **Complex Reasoning:** GLM-4.7 delivers a substantial boost in mathematical and reasoning capabilities, achieving (42.8%, +12.4%) on the HLE (Humanity's Last Exam) benchmark compared to GLM-4.6.

You can also see significant improvements in many other scenarios such as chat, creative writing, and role-play scenario.

LLM Performance Evaluation: Agentic, Reasoning and Coding

8 benchmarks: AIME 25, LiveCodeBench v6, GPQA, HLE, SWE-bench Verified, Terminal-Bench, τ^2-Bench, BrowseComp
(Evaluation results under 128K context length)



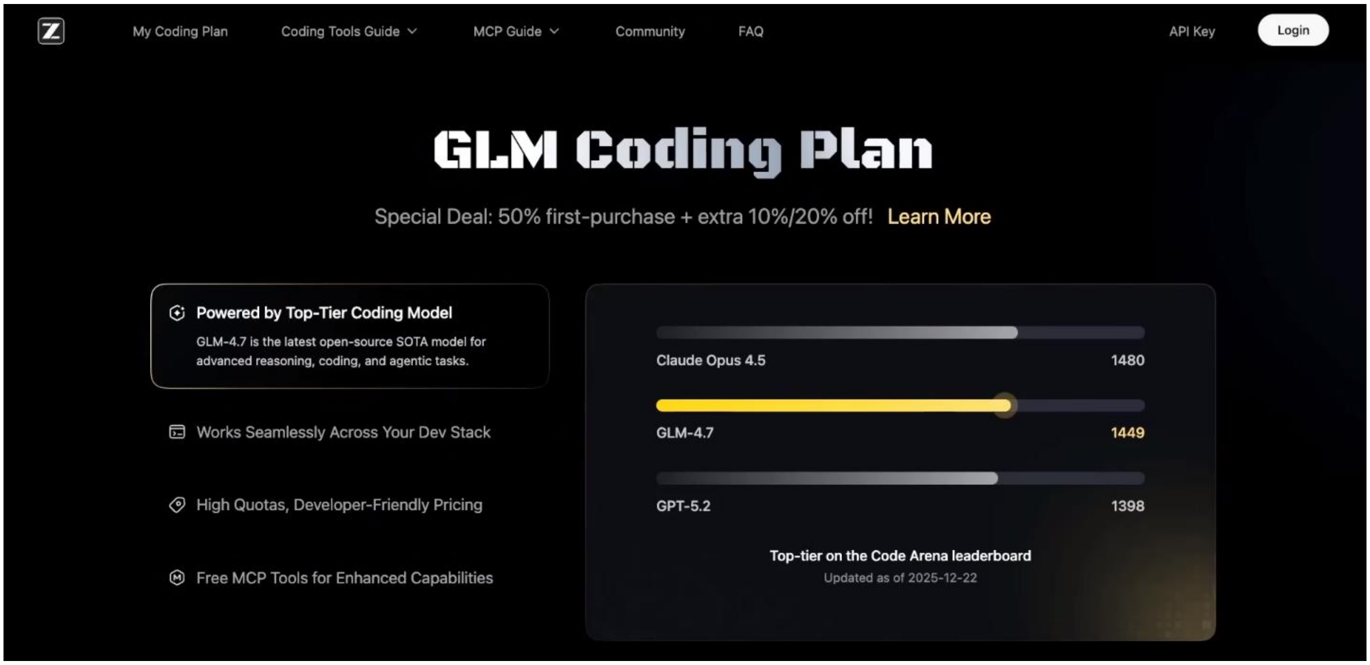
GLM-4.7-Flash

👋 Join our Discord community.

📖 Check out the GLM-4.7 technical blog, technical report(GLM-4.5).

📍 Use GLM-4.7-Flash API services on Z.ai API Platform.

👉 One click to GLM-4.7.



GLM Coding Plan

Special Deal: 50% first-purchase + extra 10%/20% off! Learn More

Powered by Top-Tier Coding Model

GLM-4.7 is the latest open-source SOTA model for advanced reasoning, coding, and agentic tasks.

Works Seamlessly Across Your Dev Stack

High Quotas, Developer-Friendly Pricing

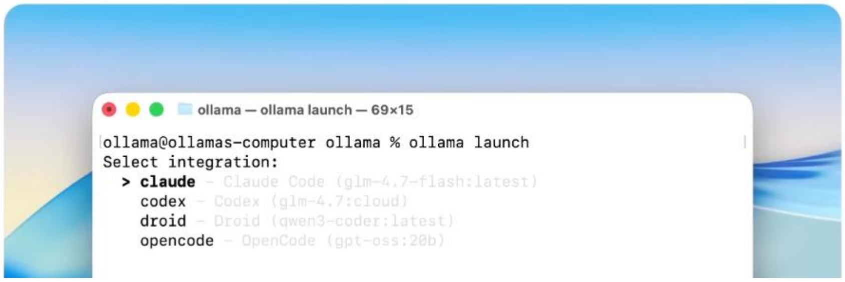
Free MCP Tools for Enhanced Capabilities

Top-tier on the Code Arena leaderboard

Updated as of 2025-12-22

ollama launch

January 23, 2026



```
ollama@ollamas-computer ollama % ollama launch
Select integration:
  > claude – Claude Code (glm-4.7-flash:latest)
    codex – Codex (glm-4.7:cloud)
    droid – Droid (qwen3-coder:latest)
    opencode – OpenCode (gpt-oss:20b)
```

`ollama launch` is a new command which sets up and runs your favorite coding tools like Claude Code, OpenCode, and Codex with local or cloud models. No environment variables or config files needed.

Get started

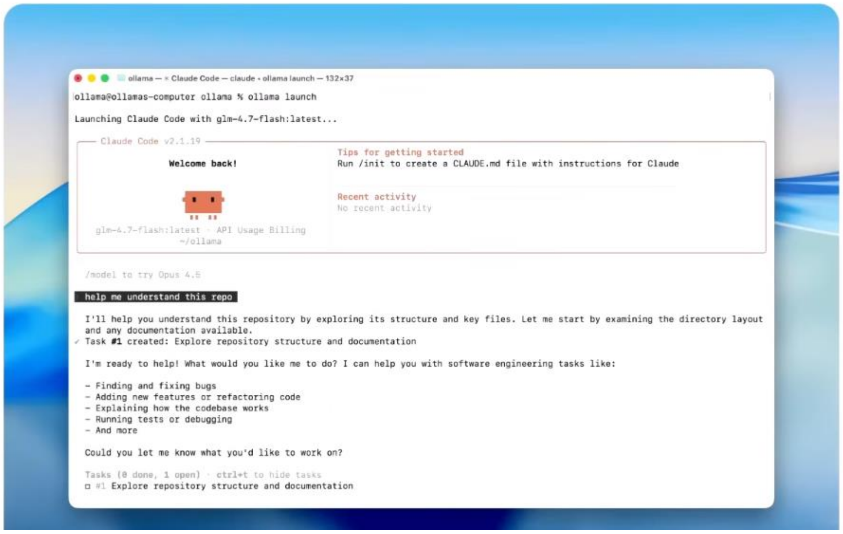
Download Ollama v0.15+, then open a terminal and run:

```
# ~23 GB VRAM required with 64000 tokens context length
ollama pull glm-4.7-flash

# or use a cloud model (with full context length)
ollama pull glm-4.7:cloud
```

One command setup

Claude Code:

```
ollama launch claude
```



```
ollama@ollamas-computer ollama % ollama launch

Launching Claude Code with glm-4.7-flash:latest...

── Claude Code v2.1.19 ──────────────────────────────────────────────────────────
                                         Tips for getting started
           Welcome back!                 Run /init to create a CLAUDE.md file with instructions for Claude

                                         Recent activity
                                         No recent activity

 glm-4.7-flash:latest · API Usage Billing
              ~/ollama

  /model to try Opus 4.5

 help me understand this repo

  I'll help you understand this repository by exploring its structure and key files. Let me start by examining the directory layout
  and any documentation available.
✓ Task #1 created: Explore repository structure and documentation

  I'm ready to help! What would you like me to do? I can help you with software engineering tasks like:

  - Finding and fixing bugs
  - Adding new features or refactoring code
  - Explaining how the codebase works
  - Running tests or debugging
  - And more

  Could you let me know what you'd like to work on?

  Tasks (0 done, 1 open) · ctrl+t to hide tasks
  □ #1 Explore repository structure and documentation
```

Supported integrations

- Claude Code
- OpenCode
- Codex
- Droid

Recommended models for coding

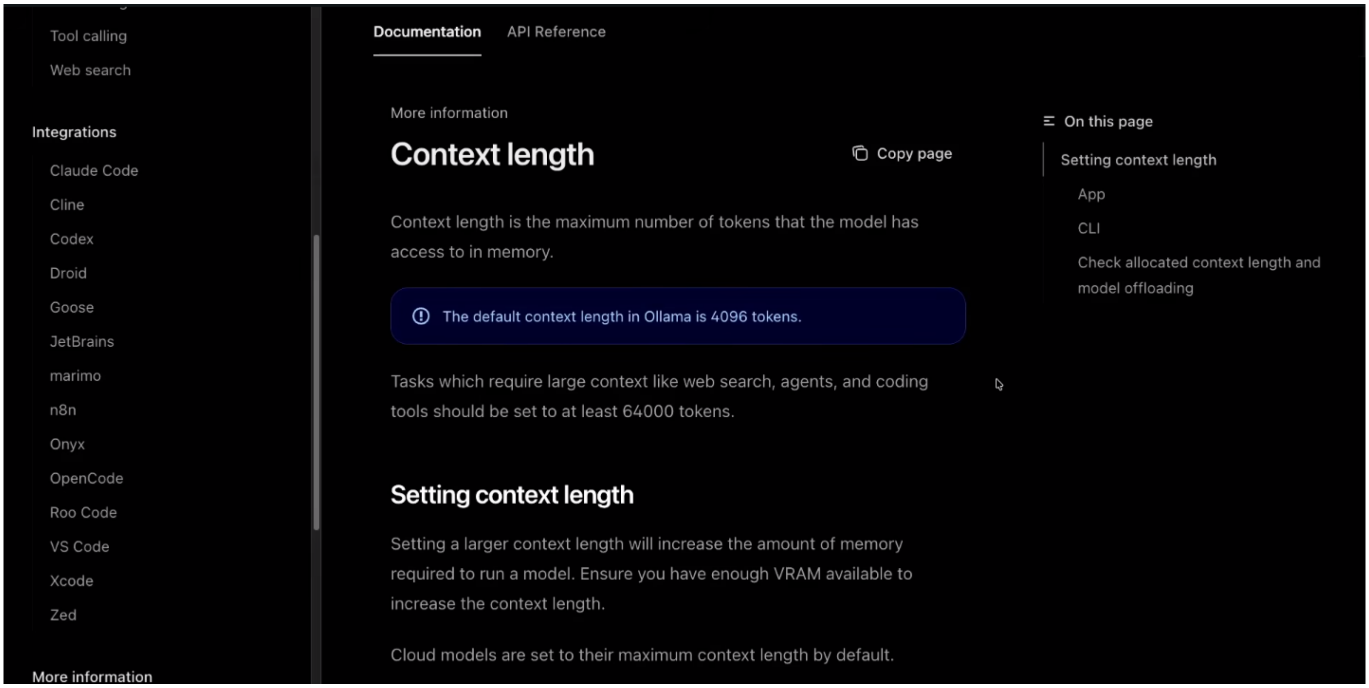
Note: Coding tools work best with a full context length. Update the context length in Ollama's settings to at least 64000 tokens. See the context length documentation on how to make changes.

Local models:

- `glm-4.7-flash`
- `qwen3-coder`
- `gpt-oss:20b`

Cloud models:

- `glm-4.7:cloud`
- `minimax-m2.1:cloud`



More information

Context length

Context length is the maximum number of tokens that the model has access to in memory.

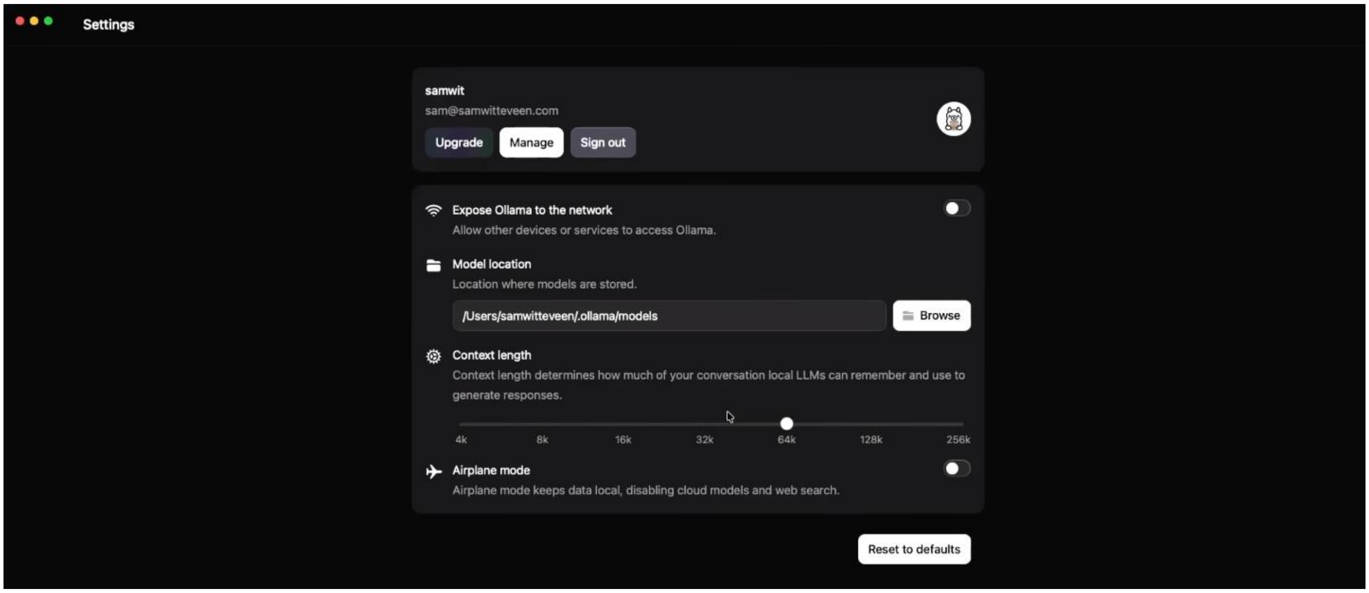
The default context length in Ollama is 4096 tokens.

Tasks which require large context like web search, agents, and coding tools should be set to at least 64000 tokens.

Setting context length

Setting a larger context length will increase the amount of memory required to run a model. Ensure you have enough VRAM available to increase the context length.

Cloud models are set to their maximum context length by default.



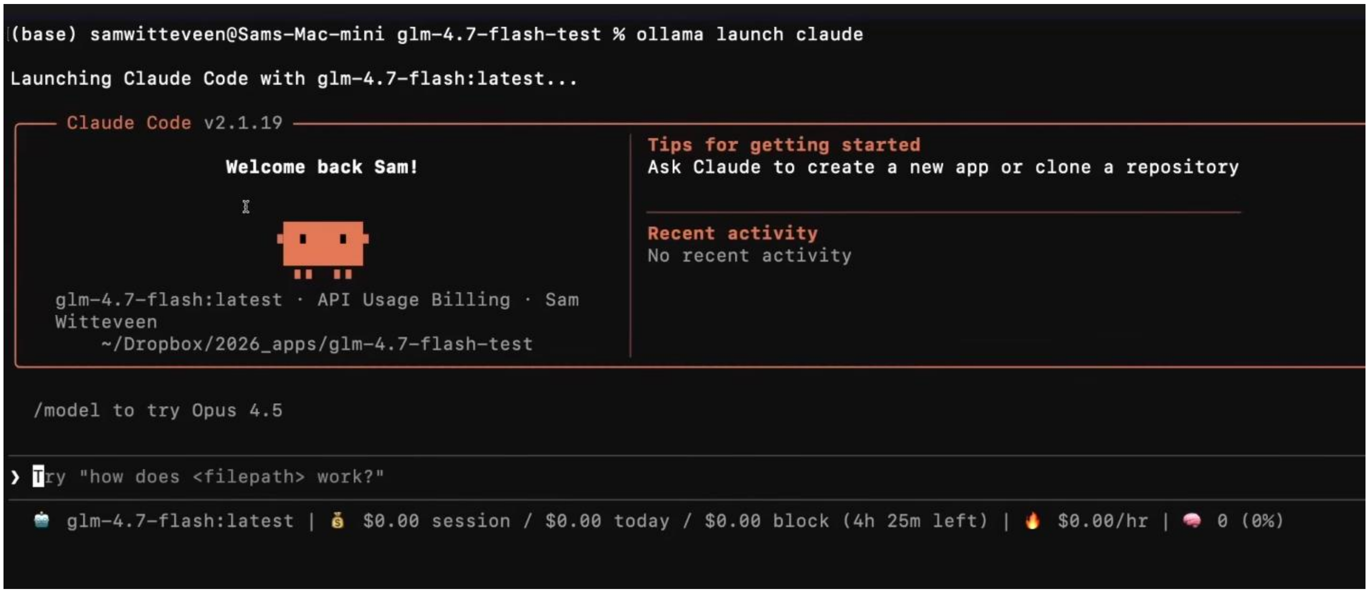
samwit
sam@samwitteveen.com

Expose Ollama to the network ○
Allow other devices or services to access Ollama.

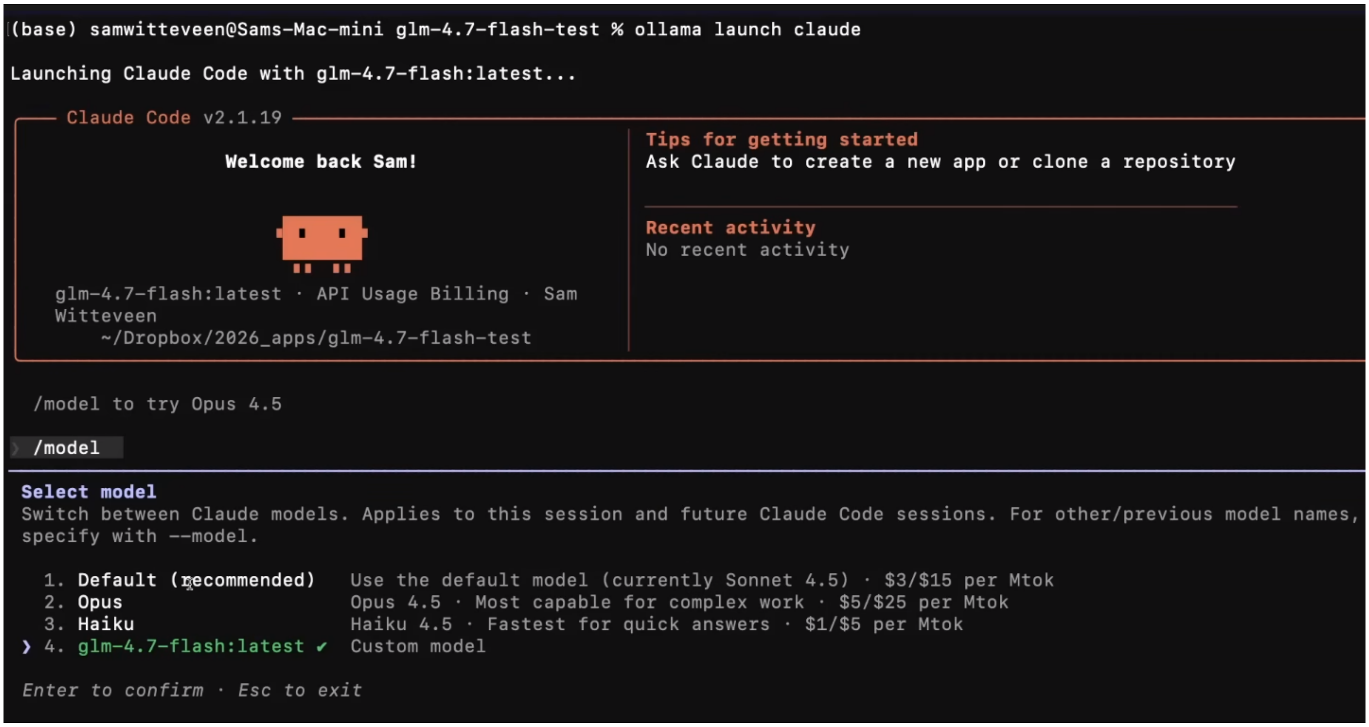
Model location
Location where models are stored.
/Users/samwitteveen/.ollama/models

Context length
Context length determines how much of your conversation local LLMs can remember and use to generate responses.
64k

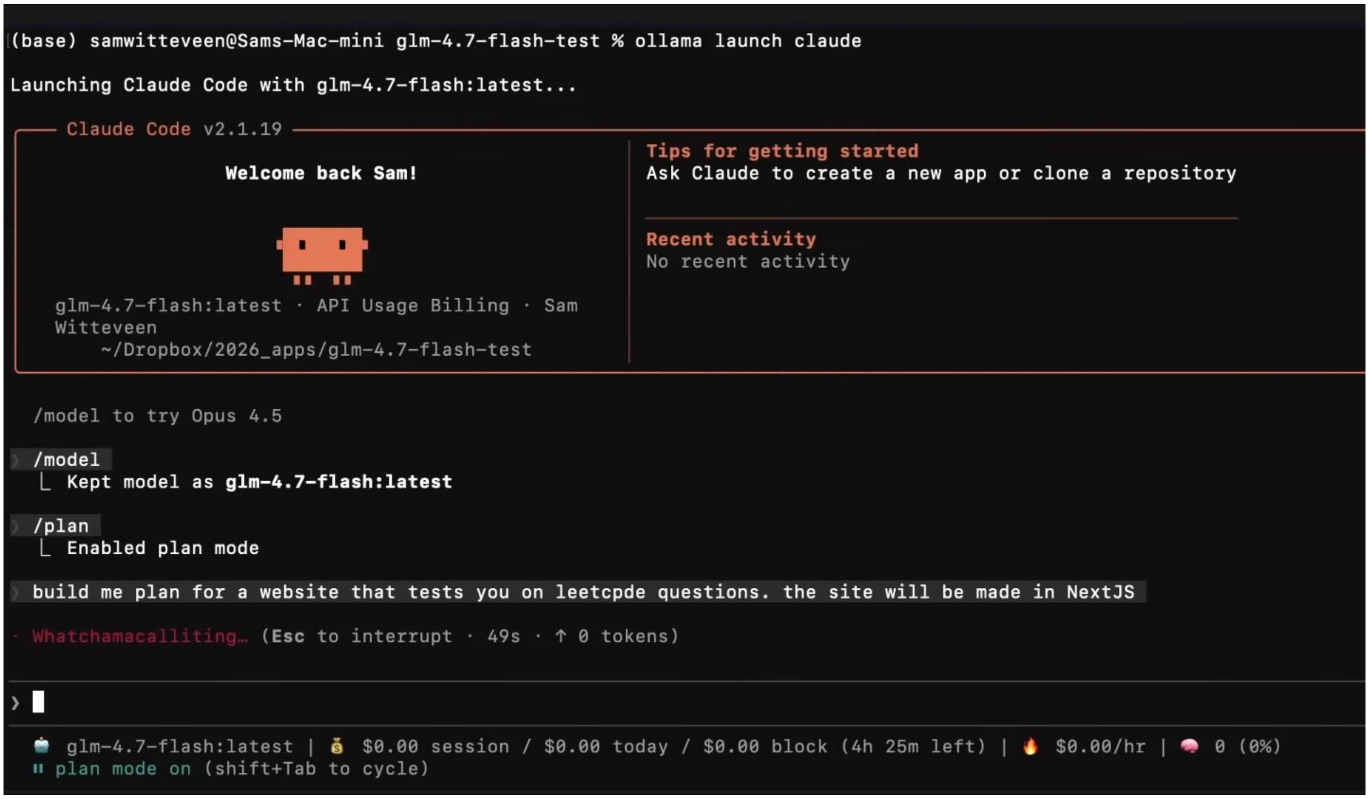
Airplane mode ○
Airplane mode keeps data local, disabling cloud models and web search.



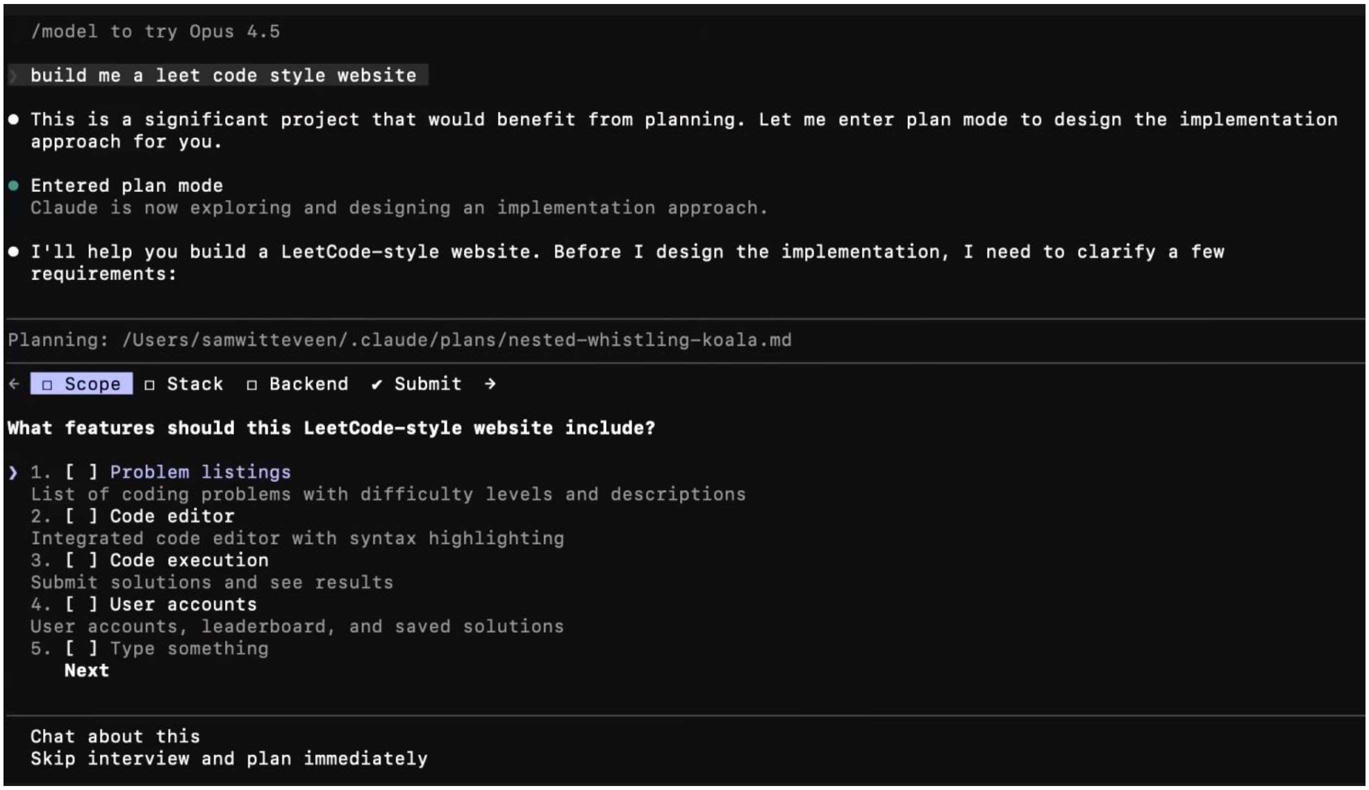
```
(base) samwitteveen@Sams-Mac-mini glm-4.7-flash-test % ollama launch claude

Launching Claude Code with glm-4.7-flash:latest...

── Claude Code v2.1.19 ──────────────────────────────────────────────────────────

              Welcome back Sam!            │ Tips for getting started
                                           │ Ask Claude to create a new app or clone a repository
                                           │ ─────────────────────────────────────────────────
                                           │ Recent activity
                                           │ No recent activity

  glm-4.7-flash:latest · API Usage Billing · Sam
  Witteveen
      ~/Dropbox/2026_apps/glm-4.7-flash-test

  /model to try Opus 4.5

❯ Try "how does <filepath> work?"

  🤖 glm-4.7-flash:latest | 💰 $0.00 session / $0.00 today / $0.00 block (4h 25m left) | 🔥 $0.00/hr | 🧠 0 (0%)
```



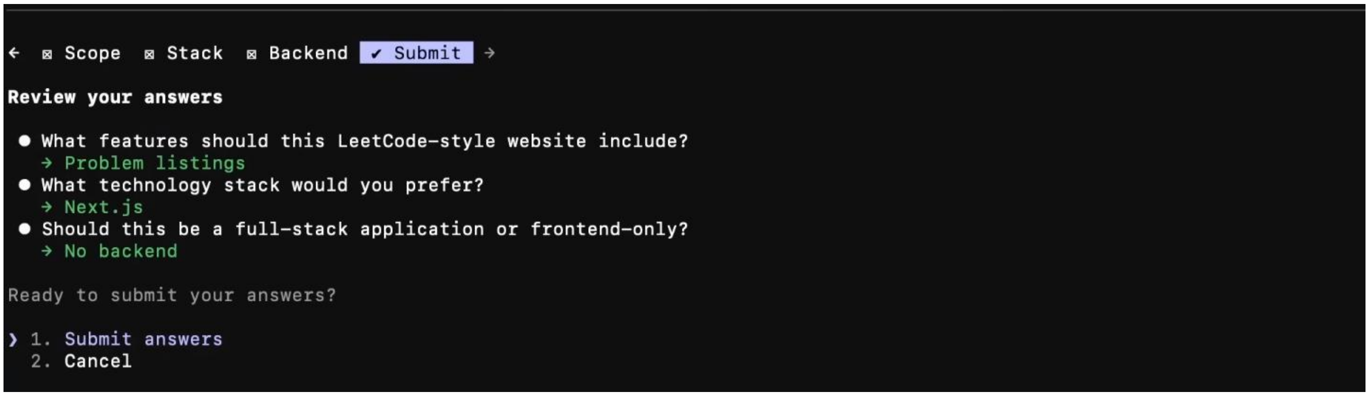
```
(base) samwitteveen@Sams-Mac-mini glm-4.7-flash-test % ollama launch claude

Launching Claude Code with glm-4.7-flash:latest...

── Claude Code v2.1.19 ──────────────────────────────────────────────────────────

              Welcome back Sam!            │ Tips for getting started
                                           │ Ask Claude to create a new app or clone a repository
                                           │ ─────────────────────────────────────────────────
                                           │ Recent activity
                                           │ No recent activity

  glm-4.7-flash:latest · API Usage Billing · Sam
  Witteveen
      ~/Dropbox/2026_apps/glm-4.7-flash-test

  /model to try Opus 4.5

❯ /model
─────────────────────────────────────────────────────────────────────────────────
 Select model
 Switch between Claude models. Applies to this session and future Claude Code sessions. For other/previous model names,
 specify with --model.

   1. Default (recommended)    Use the default model (currently Sonnet 4.5) · $3/$15 per Mtok
   2. Opus                     Opus 4.5 · Most capable for complex work · $5/$25 per Mtok
   3. Haiku                    Haiku 4.5 · Fastest for quick answers · $1/$5 per Mtok
 ❯ 4. glm-4.7-flash:latest ✔   Custom model

 Enter to confirm · Esc to exit
```



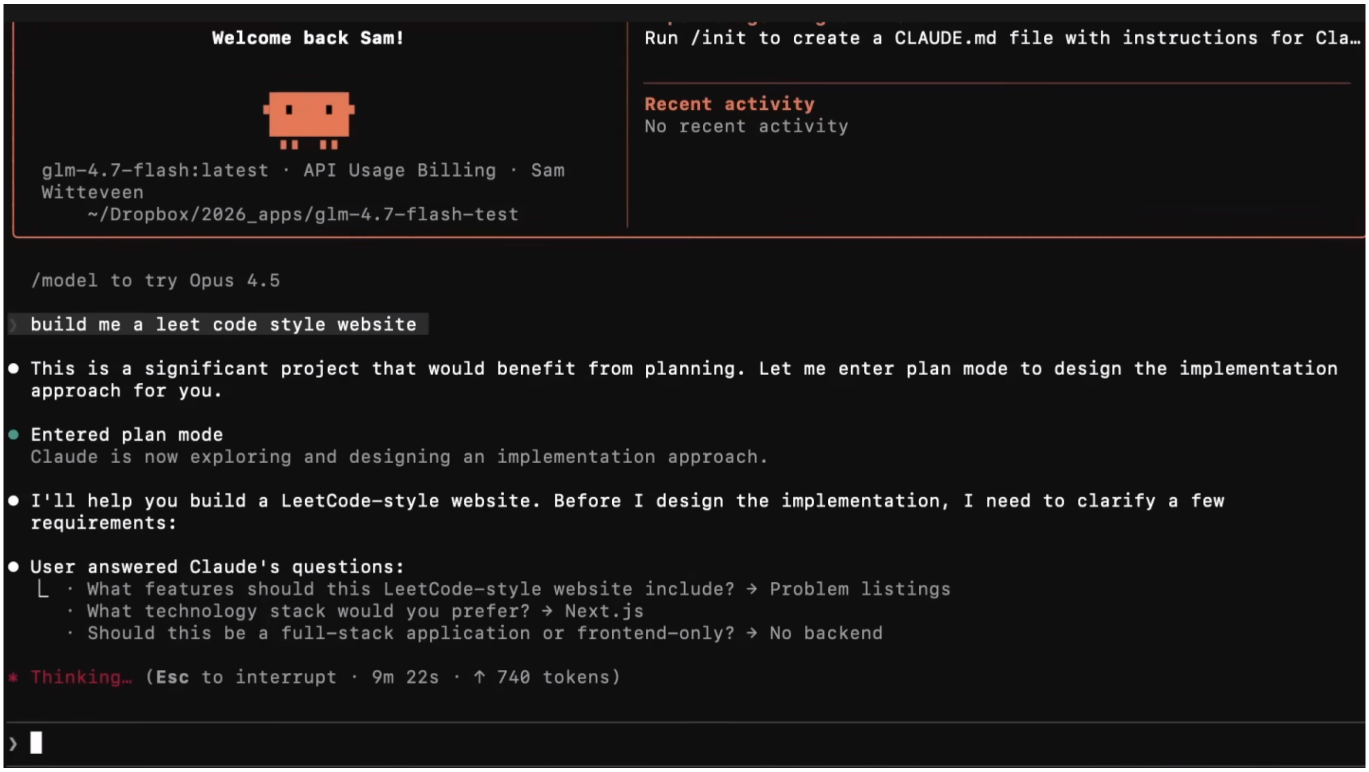
```
(base) samwitteveen@Sams-Mac-mini glm-4.7-flash-test % ollama launch claude

Launching Claude Code with glm-4.7-flash:latest...

╭─── Claude Code v2.1.19 ──────────────────────────────────────────────────────────────
│                                        │ Tips for getting started
│           Welcome back Sam!            │ Ask Claude to create a new app or clone a repository
│                                        │ ─────────────────────────────────────────────────────
│                                        │ Recent activity
│                                        │ No recent activity
│                                        │
│  glm-4.7-flash:latest · API Usage Billing · Sam │
│  Witteveen                             │
│      ~/Dropbox/2026_apps/glm-4.7-flash-test     │
╰──────────────────────────────────────────────────────────────────────────────────────

  /model to try Opus 4.5

> /model
  ⎿ Kept model as glm-4.7-flash:latest

> /plan
  ⎿ Enabled plan mode

> build me plan for a website that tests you on leetcpde questions. the site will be made in NextJS

· Whatchamacalliting… (Esc to interrupt · 49s · ↑ 0 tokens)

────────────────────────────────────────────────────────────────────────────────────────
> █
────────────────────────────────────────────────────────────────────────────────────────
  🤖 glm-4.7-flash:latest | 💰 $0.00 session / $0.00 today / $0.00 block (4h 25m left) | 🔥 $0.00/hr | 🧠 0 (0%)
  ⏸ plan mode on (shift+Tab to cycle)
```



```
  /model to try Opus 4.5

> build me a leet code style website

● This is a significant project that would benefit from planning. Let me enter plan mode to design the implementation
  approach for you.

● Entered plan mode
  Claude is now exploring and designing an implementation approach.

● I'll help you build a LeetCode-style website. Before I design the implementation, I need to clarify a few
  requirements:

────────────────────────────────────────────────────────────────────────────────────────
Planning: /Users/samwitteveen/.claude/plans/nested-whistling-koala.md
────────────────────────────────────────────────────────────────────────────────────────
←  □ Scope  □ Stack  □ Backend  ✔ Submit  →

What features should this LeetCode-style website include?

❯ 1. [ ] Problem listings
  List of coding problems with difficulty levels and descriptions
  2. [ ] Code editor
  Integrated code editor with syntax highlighting
  3. [ ] Code execution
  Submit solutions and see results
  4. [ ] User accounts
  User accounts, leaderboard, and saved solutions
  5. [ ] Type something
     Next

────────────────────────────────────────────────────────────────────────────────────────
  Chat about this
  Skip interview and plan immediately
```



```
← ☒ Scope  ☒ Stack  ☒ Backend  ✔ Submit  →

Review your answers

 ● What features should this LeetCode-style website include?
   → Problem listings
 ● What technology stack would you prefer?
   → Next.js
 ● Should this be a full-stack application or frontend-only?
   → No backend

Ready to submit your answers?

❯ 1. Submit answers
  2. Cancel
```



```
                  Welcome back Sam!                  │ Run /init to create a CLAUDE.md file with instructions for Cla…
                                                      │ ────────────────────────────────────────────────────────────
                                                      │ Recent activity
                                                      │ No recent activity

   glm-4.7-flash:latest · API Usage Billing · Sam
   Witteveen
       ~/Dropbox/2026_apps/glm-4.7-flash-test

  /model to try Opus 4.5

> build me a leet code style website

● This is a significant project that would benefit from planning. Let me enter plan mode to design the implementation
  approach for you.

● Entered plan mode
  Claude is now exploring and designing an implementation approach.

● I'll help you build a LeetCode-style website. Before I design the implementation, I need to clarify a few
  requirements:

● User answered Claude's questions:
  ⎿  · What features should this LeetCode-style website include? → Problem listings
     · What technology stack would you prefer? → Next.js
     · Should this be a full-stack application or frontend-only? → No backend

* Thinking… (Esc to interrupt · 9m 22s · ↑ 740 tokens)

────────────────────────────────────────────────────────────────────────────────────
> 
```