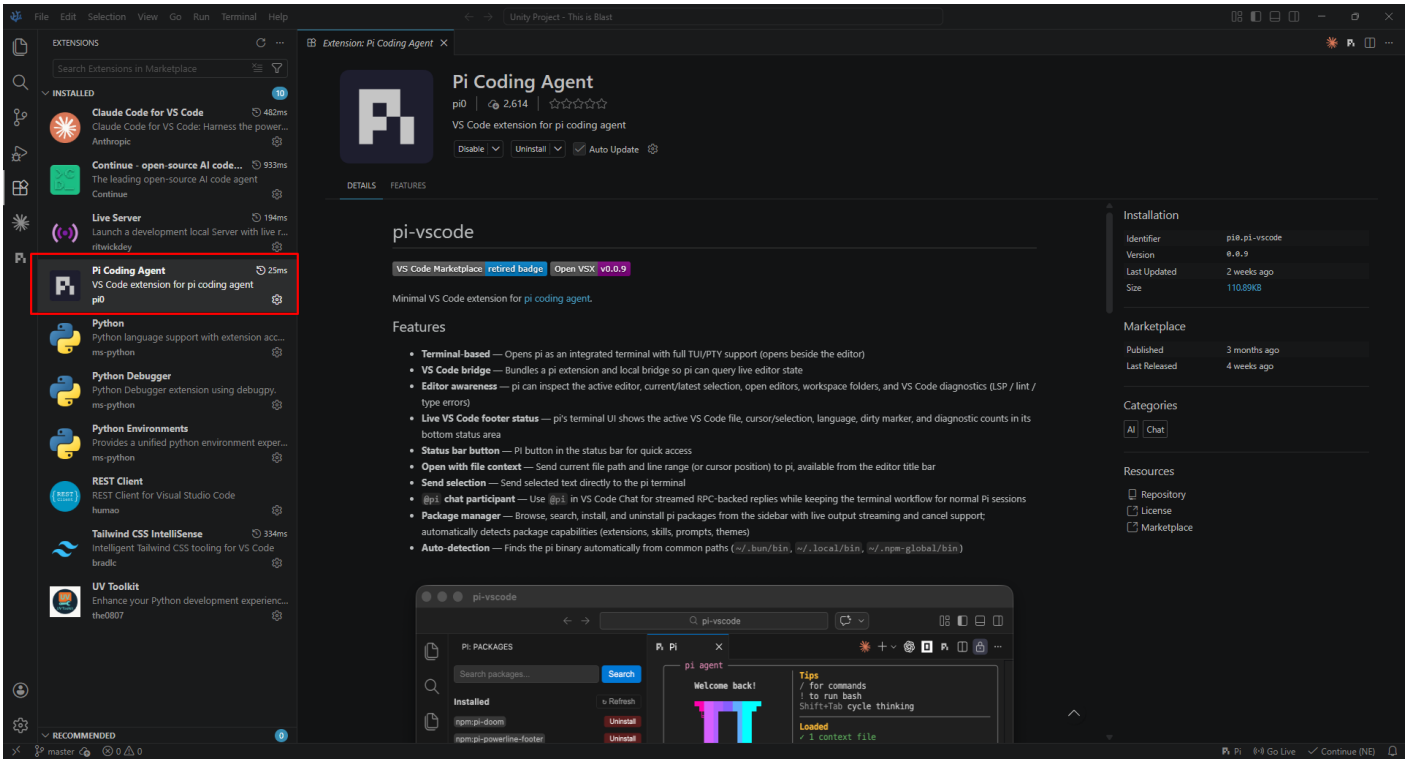
□□ pi.dev | Claude alternative

1. Follow this page to install:

[pi.dev github]

2. Install extension: **Pi Coding Agent**



3. Connect to hosted llm service:

- Edit file: **"C:\Users\helloworld\.pi\agent\models.json"**

```
{
  "providers": {
 "llama-cpp": {
   "baseUrl": "http://localhost:8888/v1",
   "api": "openai-completions",
```

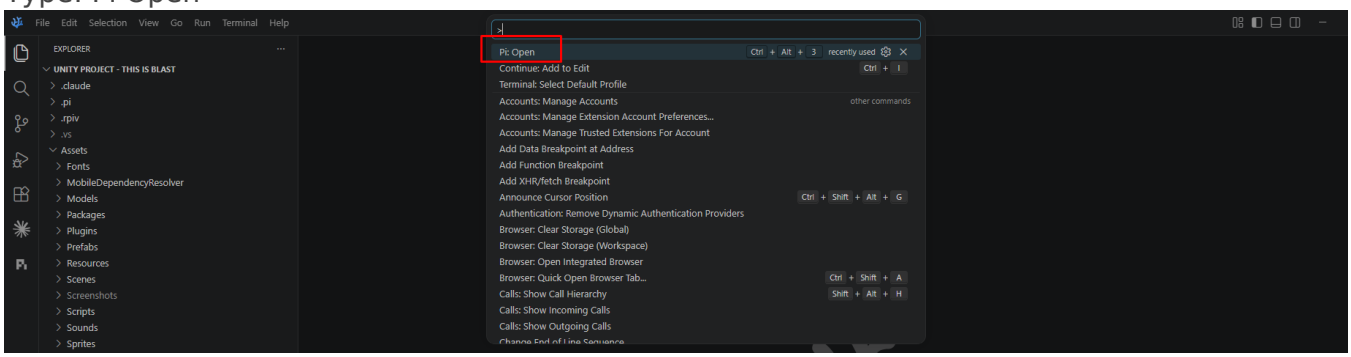
```
   "apiKey": "none",
   "compat": {
  "supportsDeveloperRole": false
   },
   "models": [
  {
    "id": "unsloth/gemma-4-E2B-it-GGUF:Q4_K_M",
    "input": [
   "text",
   "image"
    ],
    "contextWindow": 131072,
          "maxTokens": 32000,
    "reasoning": true
  }
   ]
    }
  }
}
```

> ⚠ Important: Do not using Ollama because tool calling is buggy and not stable for coding agent. Use llama.cpp / vllm / sglang for best compatibility and performance.

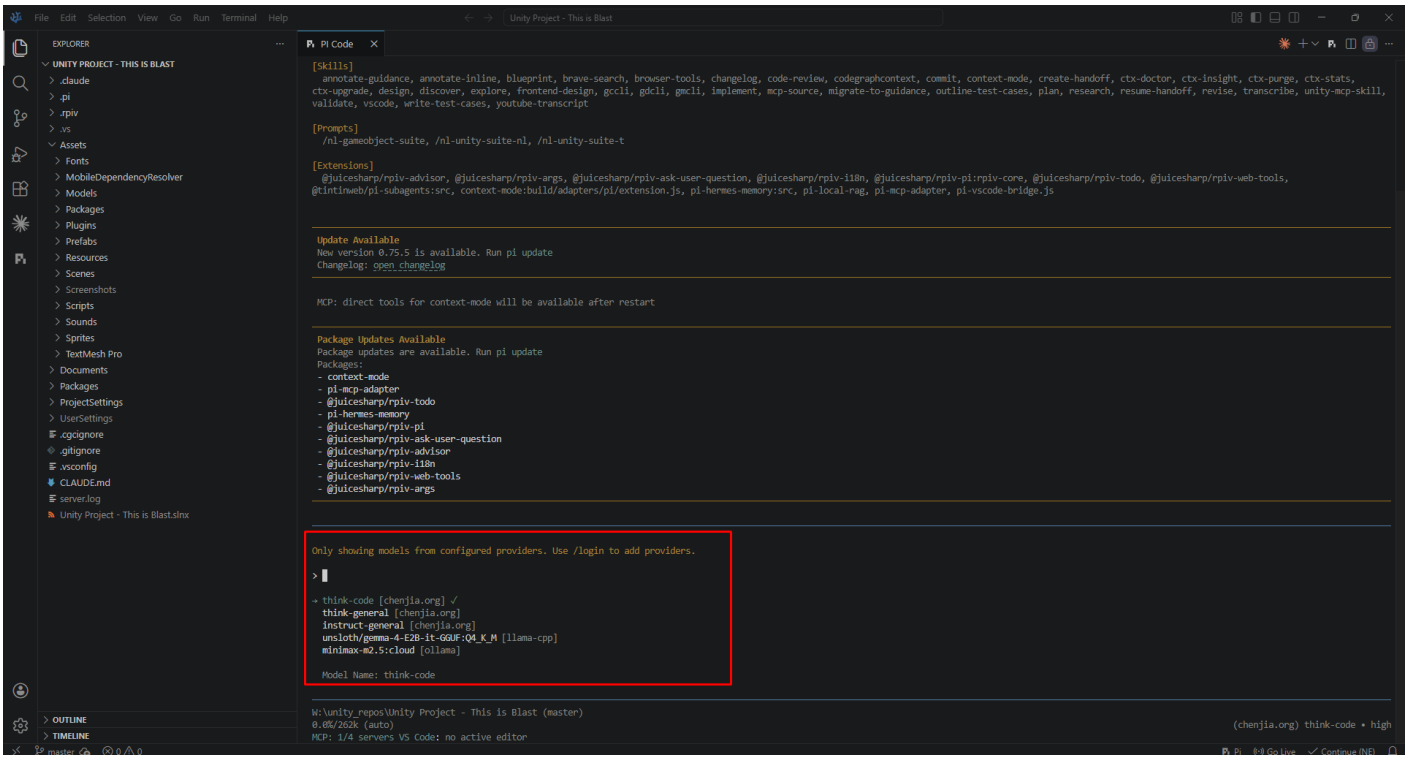
---

4. Select model and go

- Open vscode or codium, then press Ctrl + Shift + T
- Type: Pi Open




```
/model # select model
```



5. Recommend models

- Qwen3.6-35B-A3B-MTP
- Qwen3.6-27B-MTP
- Gemma4-26B-A4B-IT
- Gemma4-31B-IT

Revision #3
Created 2026-05-24 01:30:07 UTC by sandichhuu
Updated 2026-05-28 05:58:16 UTC by sandichhuu