Deepseek Harness - Công cụ code hiệu quả

- Cài đặt
- Cấu hình
- Giảm chi phí với Ollama free token hàng tuần.

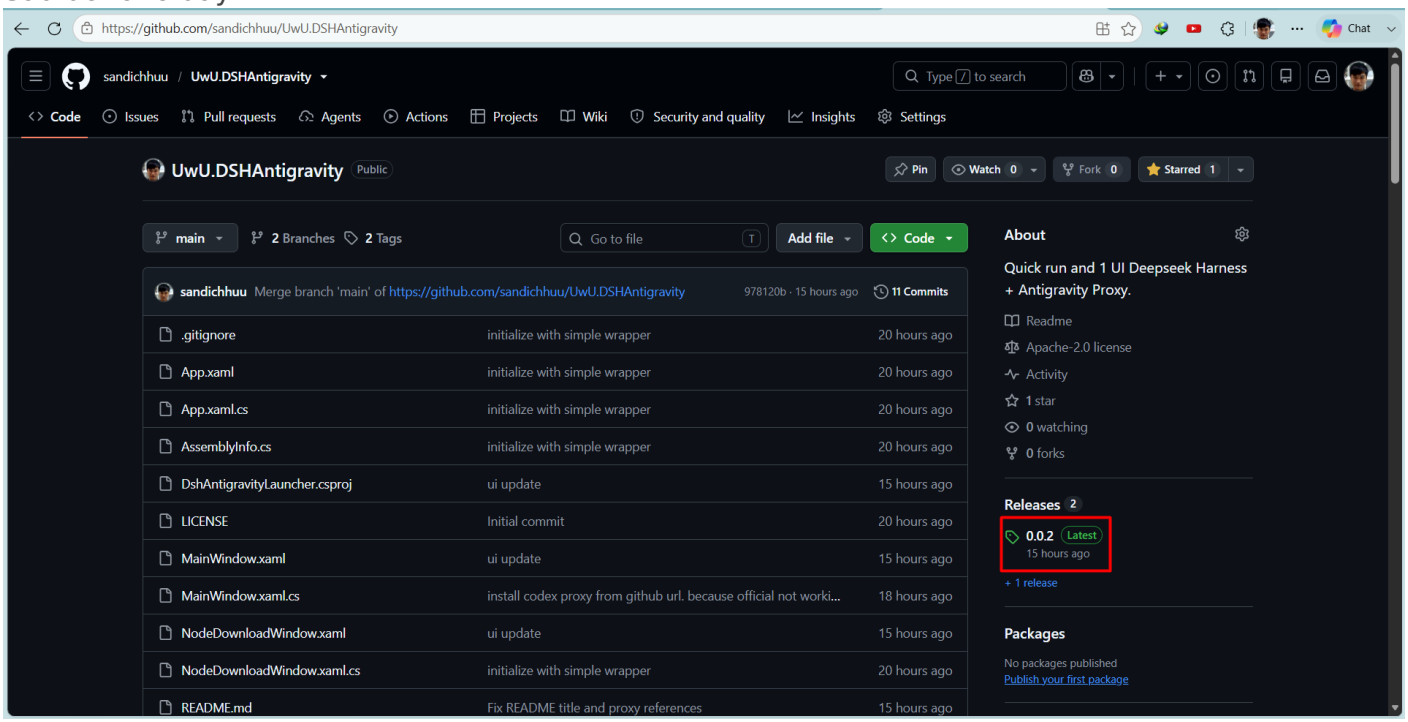
Cài đặt

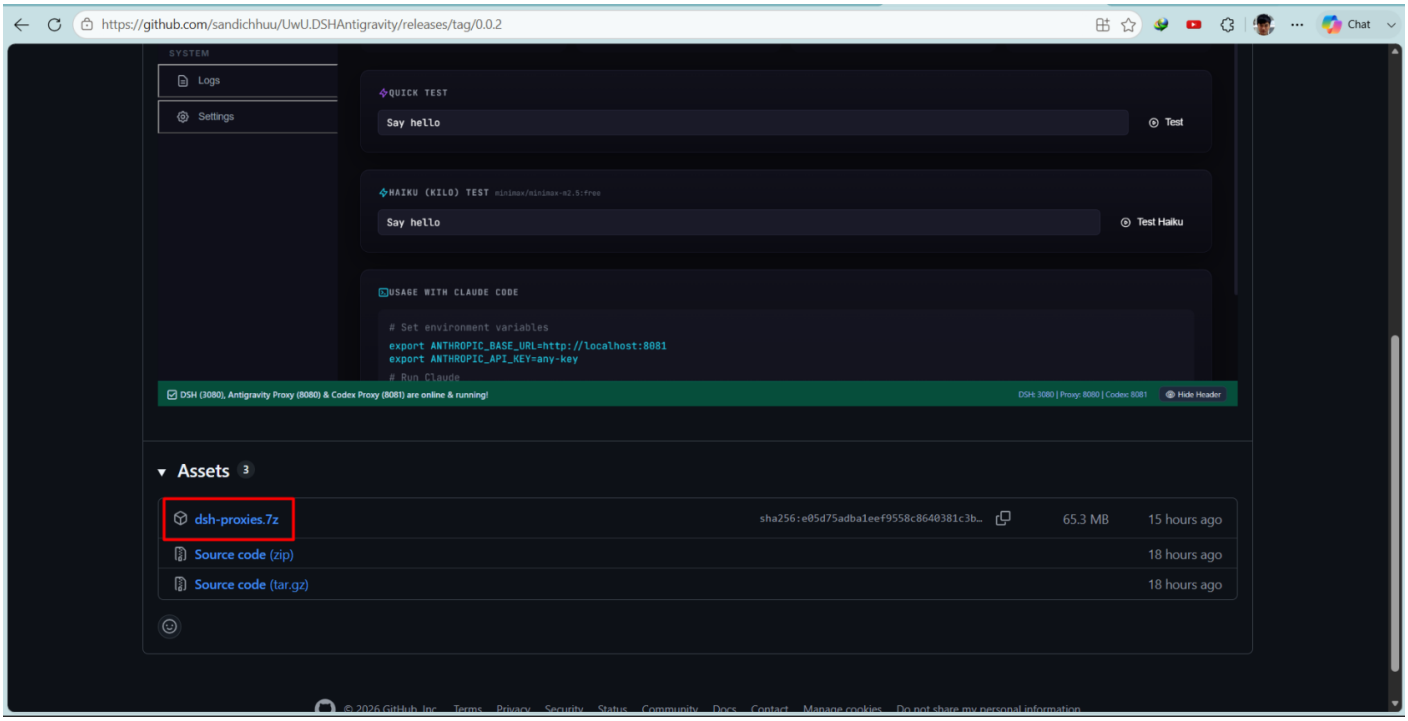
Tải phần mềm DSH Proxies

Hãy vô trang này: github

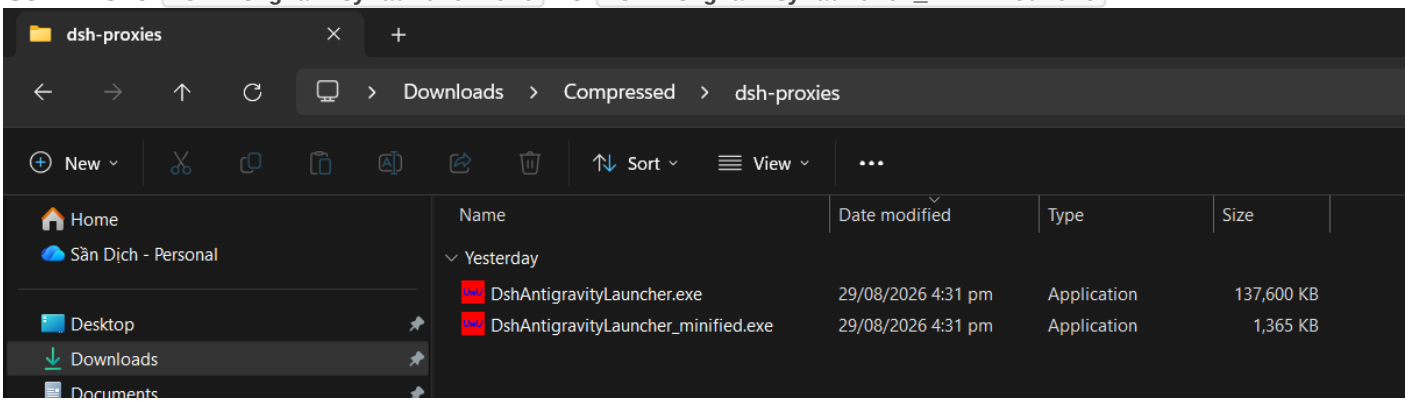
> Tải nhanh từ link này: Download

Sau đó tải ở đây:





Sau khi tải xong thì giải nén ra.
Có 2 file là `DshAntigravityLauncher.exe` và `DshAntigravityLauncher_minified.exe`.

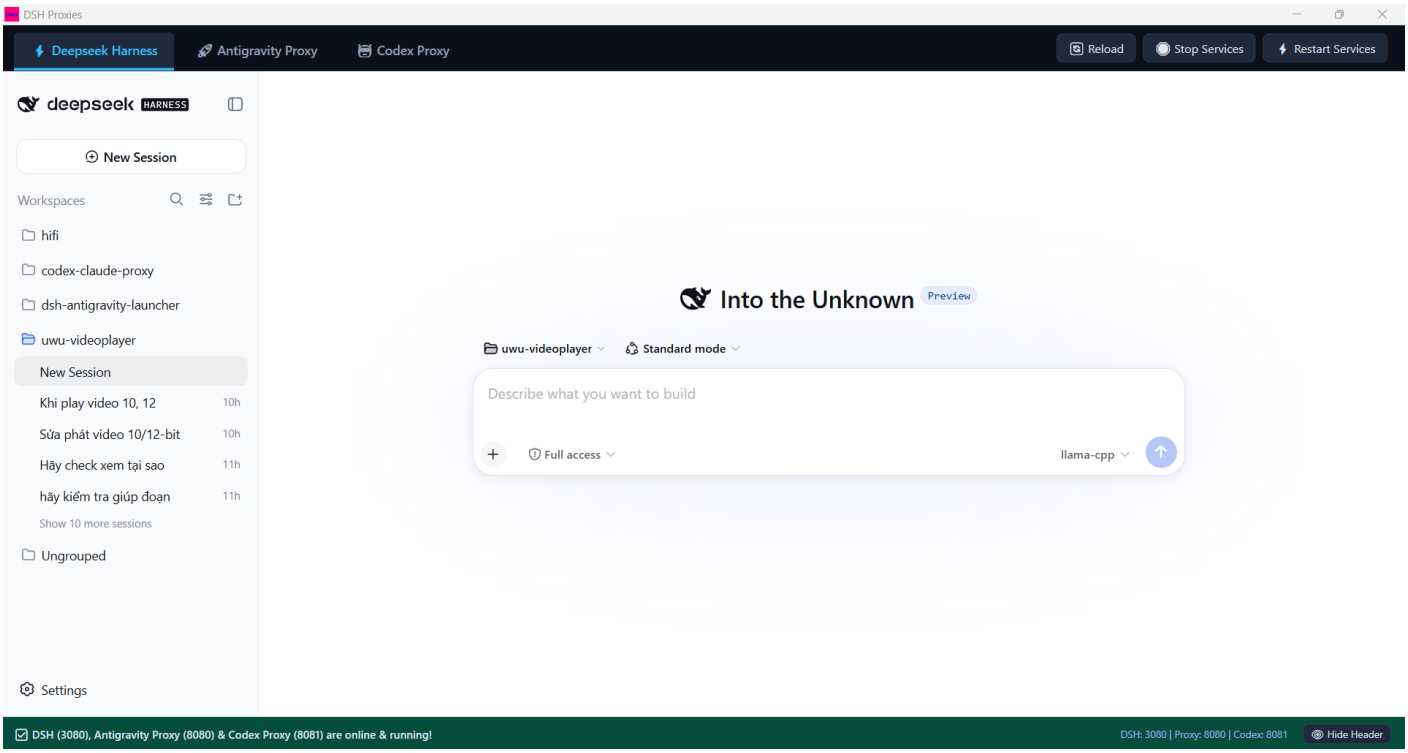


File `DshAntigravityLauncher.exe` thì nặng hơn, đảm bảo đầy đủ thư viện, mở lên là chạy được.
File `DshAntigravityLauncher_minified.exe` thì siêu nhẹ, cần phải có windows 11 và cài đặt `.NET 10 Runtime` trước khi sử dụng.

Cấu hình

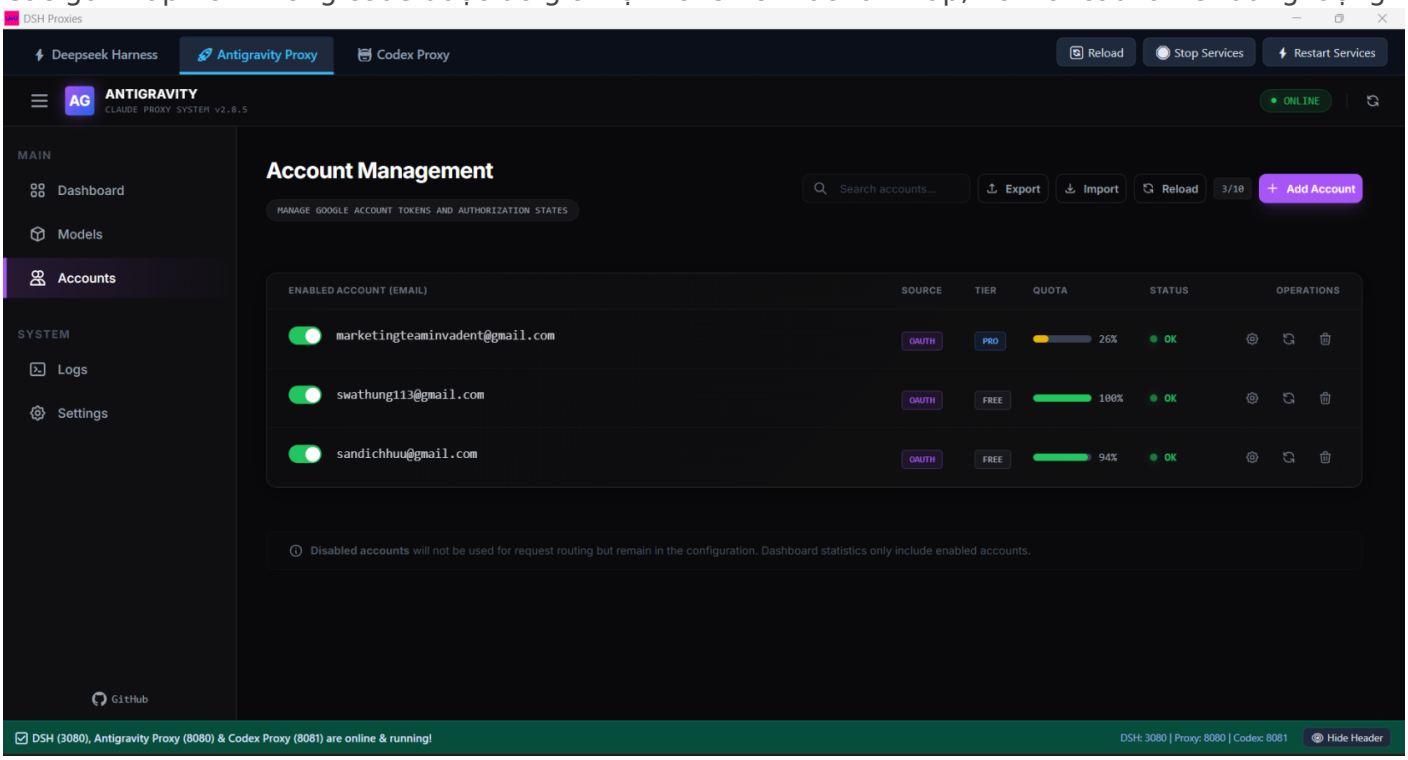
Giao diện

Phần giao diện khá đơn giản.
Ở phía trên sẽ có 3 tab là `Deepseek Harness`, `Antigravity Proxy` và `Codex Proxy`.
Ở phía dưới có nút `Hide Header` để ẩn thanh phía trên đi để tăng diện tích sử dụng.
Khi mới vào cần đợi tự động cài đặt, khi nào thanh bên dưới chuyển màu xanh lá cây mới sử dụng được.

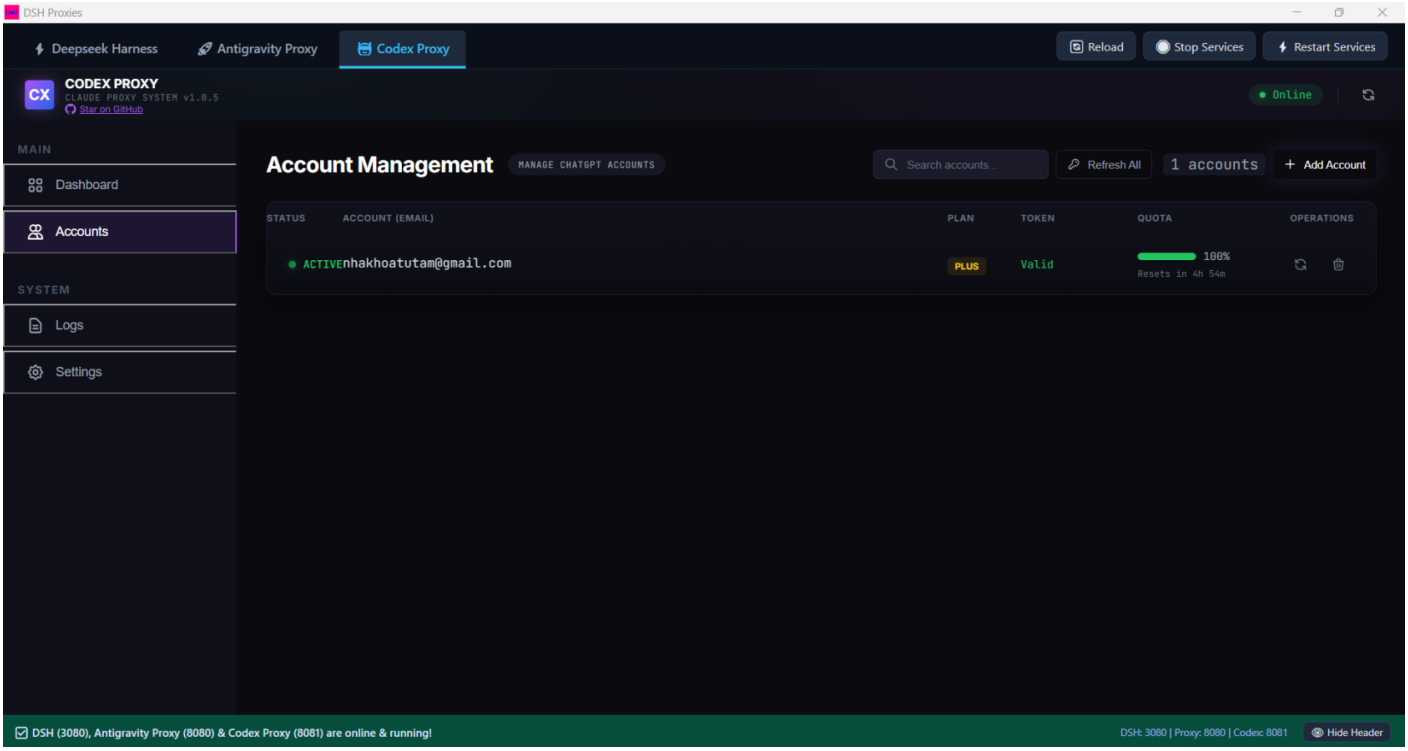


Thêm tài khoản

Đầu tiên hãy thêm tài khoản Google đã mua gói Plus hoặc Pro vào.
Các gói thấp hơn không code được do giới hạn token ở mức rất thấp, hỏi vài câu là hết dung lượng.



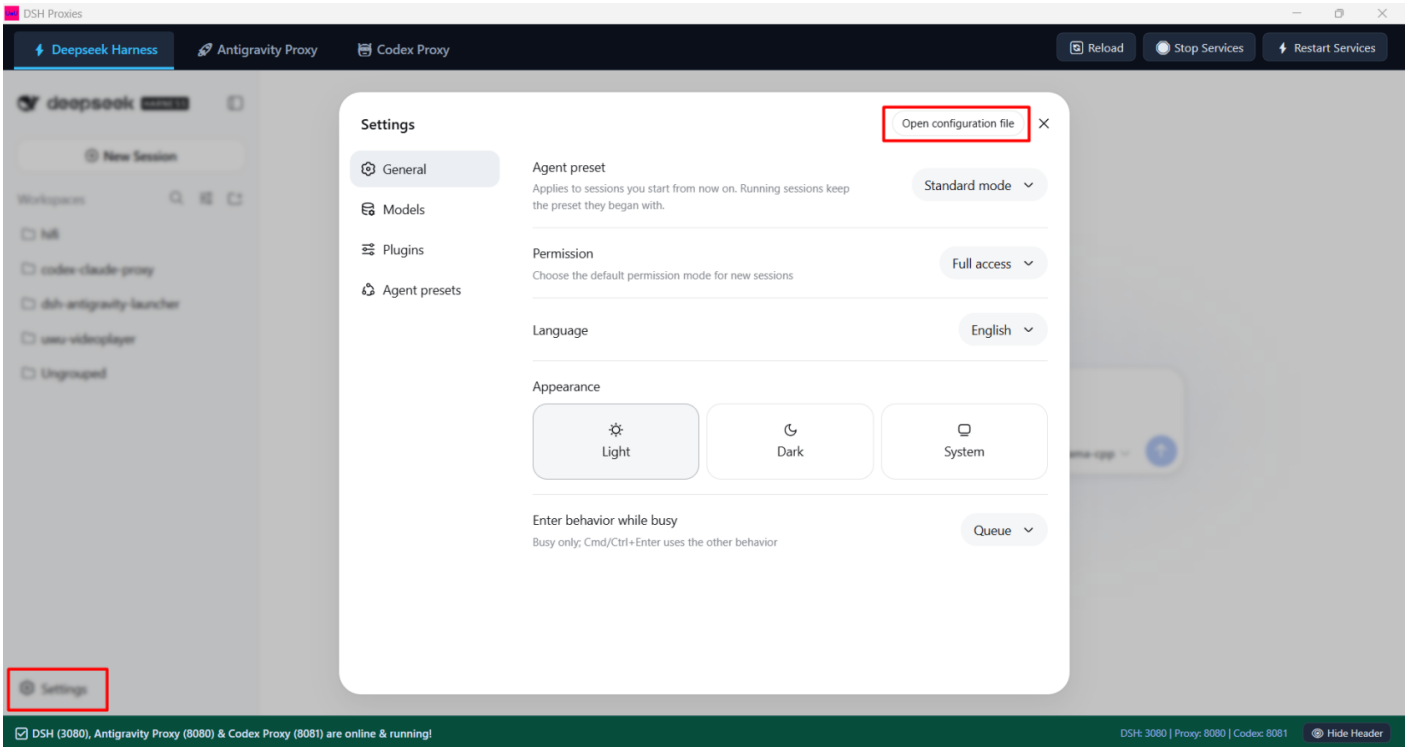
Tiếp theo thêm gói Codex đã mua bản Plus / Pro / ... vào.



> Lưu ý phần mềm hoạt động hiệu quả nhất với Antigravity, Codex hoạt động kém hơn.
> Antigravity có thể đạt tốc độ 3500 token/s, còn Codex chỉ đạt 100-350 token/s thôi. Code chậm hơn rất nhiều.

Cấu hình DSH

Đầu tiên hãy vào tab Deekseek Harness.
Ấn nút bánh răng ở góc dưới bên trái.
Rồi ấn nút `Open configuration file`.



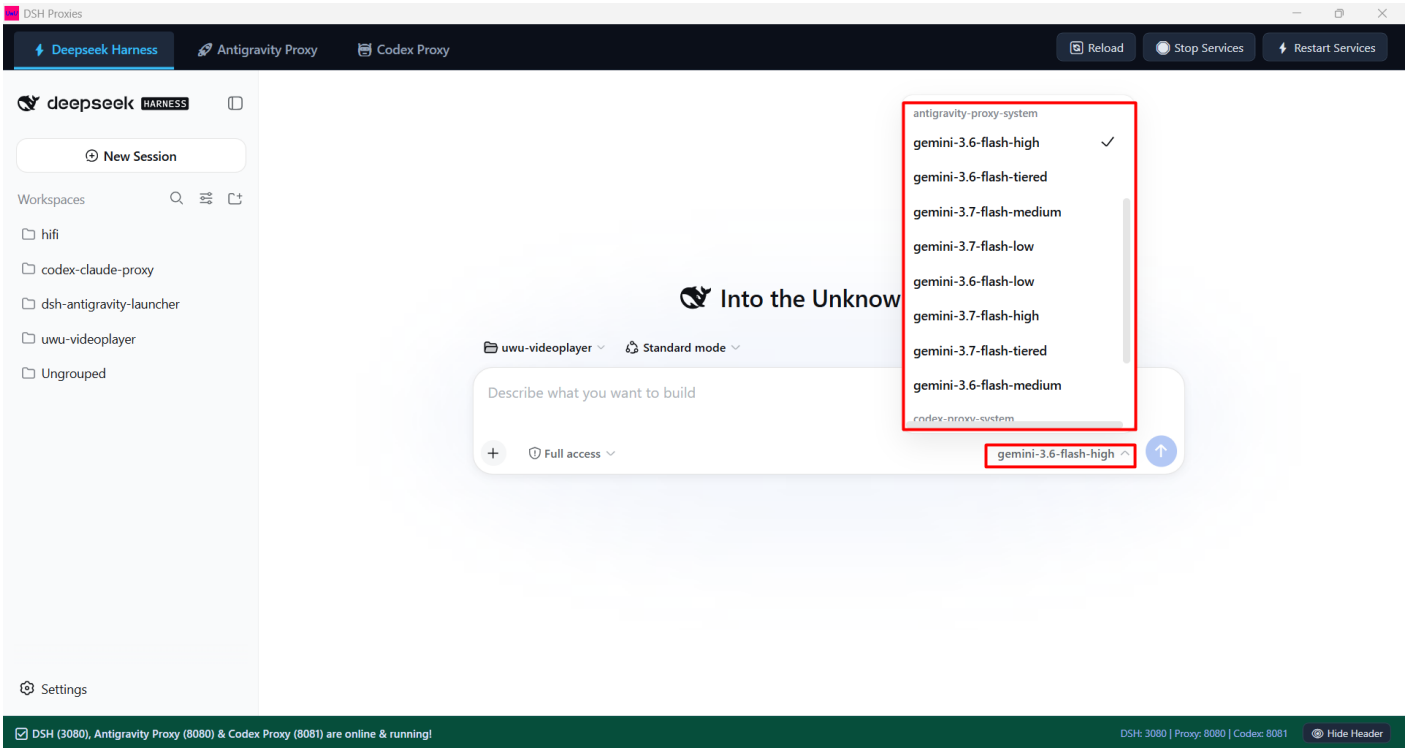
Tiếp theo hãy nó sẽ tự mở một file text ra bằng notepad cho mình sửa.
Hãy paste đoạn config sau vào để chạy:

```
ui-onboarding:
  welcomeNoticeVersion: 2026-08-13.1
llm-pi-ai:
  providers:
    {
      antigravity-proxy-system:
        {
          displayName: antigravity-proxy-system,
          apiKeyEnv: ANTIGRAVITY_PROXY_SYSTEM_API_KEY,
          api: anthropic-messages,
          baseURL: http://127.0.0.1:8080,
          models:
            [
              { id: gemini-3.6-flash-high, contextWindow: 1048576, maxTokens: 65536 },
              { id: gemini-3.6-flash-tiered, contextWindow: 1048576, maxTokens: 65536 },
              { id: gemini-3.7-flash-medium, contextWindow: 1048576, maxTokens: 65536 },
              { id: gemini-3.7-flash-low, contextWindow: 1048576, maxTokens: 65536 },
              { id: gemini-3.6-flash-low, contextWindow: 1048576, maxTokens: 65536 },
              { id: gemini-3.7-flash-high, contextWindow: 1048576, maxTokens: 65536 },
              { id: gemini-3.7-flash-tiered, contextWindow: 1048576, maxTokens: 65536 },
              { id: gemini-3.6-flash-medium, contextWindow: 1048576, maxTokens: 65536 }
            ]
        },
      codex-proxy-system:
```

```
        {
          displayName: codex-proxy-system,
          apiKeyEnv: CODEX_PROXY_SYSTEM_API_KEY,
          api: anthropic-messages,
          baseURL: http://localhost:8081,
          models:
            [
              { id: gpt-5.6-sol, contextWindow: 1048576, maxTokens: 65536 },
              { id: gpt-5.6-terra, contextWindow: 1048576, maxTokens: 65536 },
              { id: gpt-5.6-luna, contextWindow: 1048576, maxTokens: 65536 }
            ]
        }
    }
permission:
  defaultPreset: danger-full-access
ui-theme:
  preference: light
agent-default-model:
  provider: antigravity-proxy-system
  model: gemini-3.6-flash-high
agent-presets:
  default: standard
```

Chọn mô hình để hoạt động.

Lúc này DSH sẽ hiển thị toàn bộ các mô hình của Antigravity và Codex để có thể bắt đầu sử dụng.



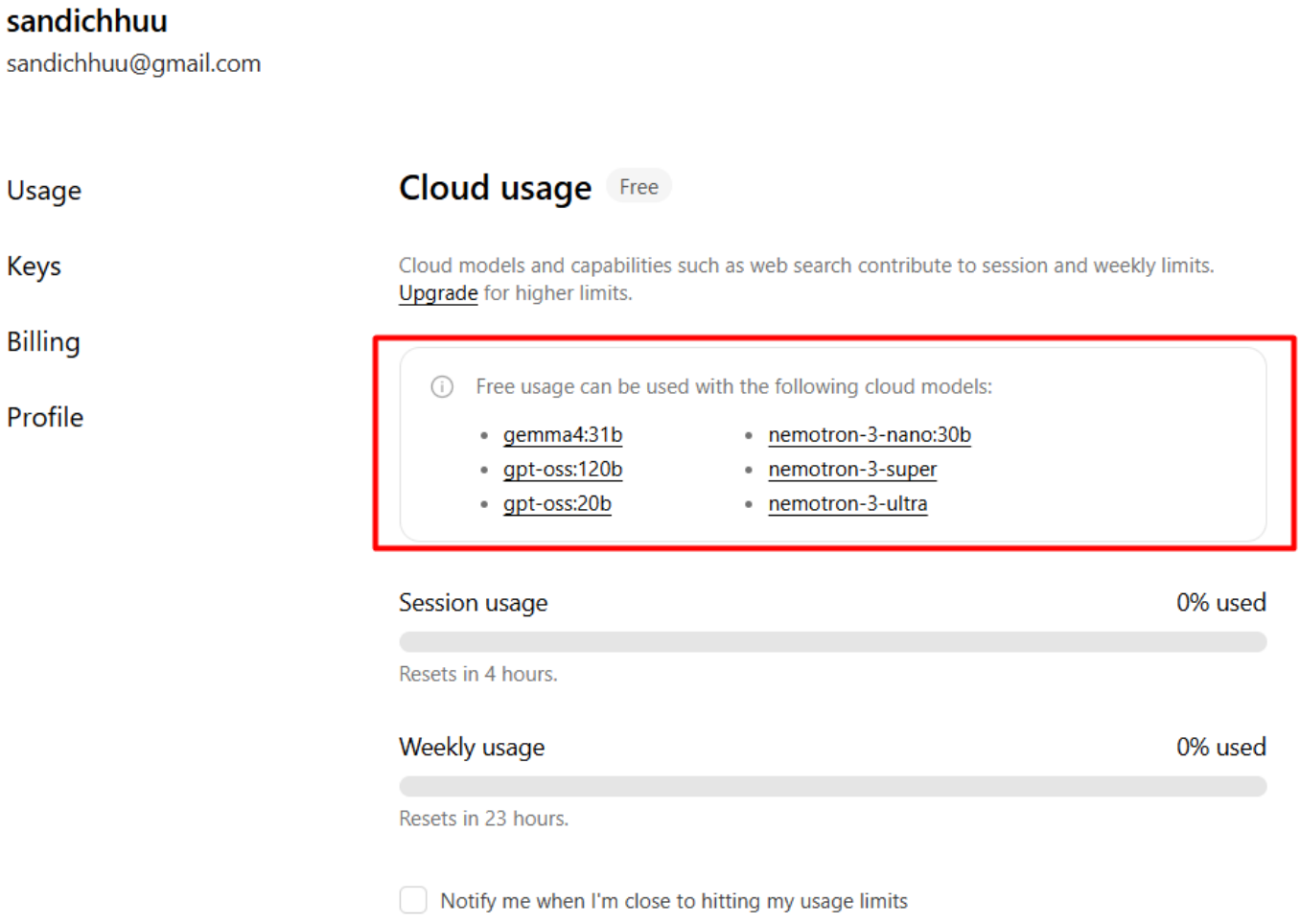
> Lưu ý, nếu có dư kinh phí thì nên mua thêm Deepseek Flash V4 để sử dụng tốt hơn. Nạp 500k một người sử dụng nhiều tháng mới hết dung lượng.

Giảm chi phí với Ollama free token hàng tuần.

Ollama

Đầu tiên hãy tạo tài khoản Ollama ở đây: Ollama

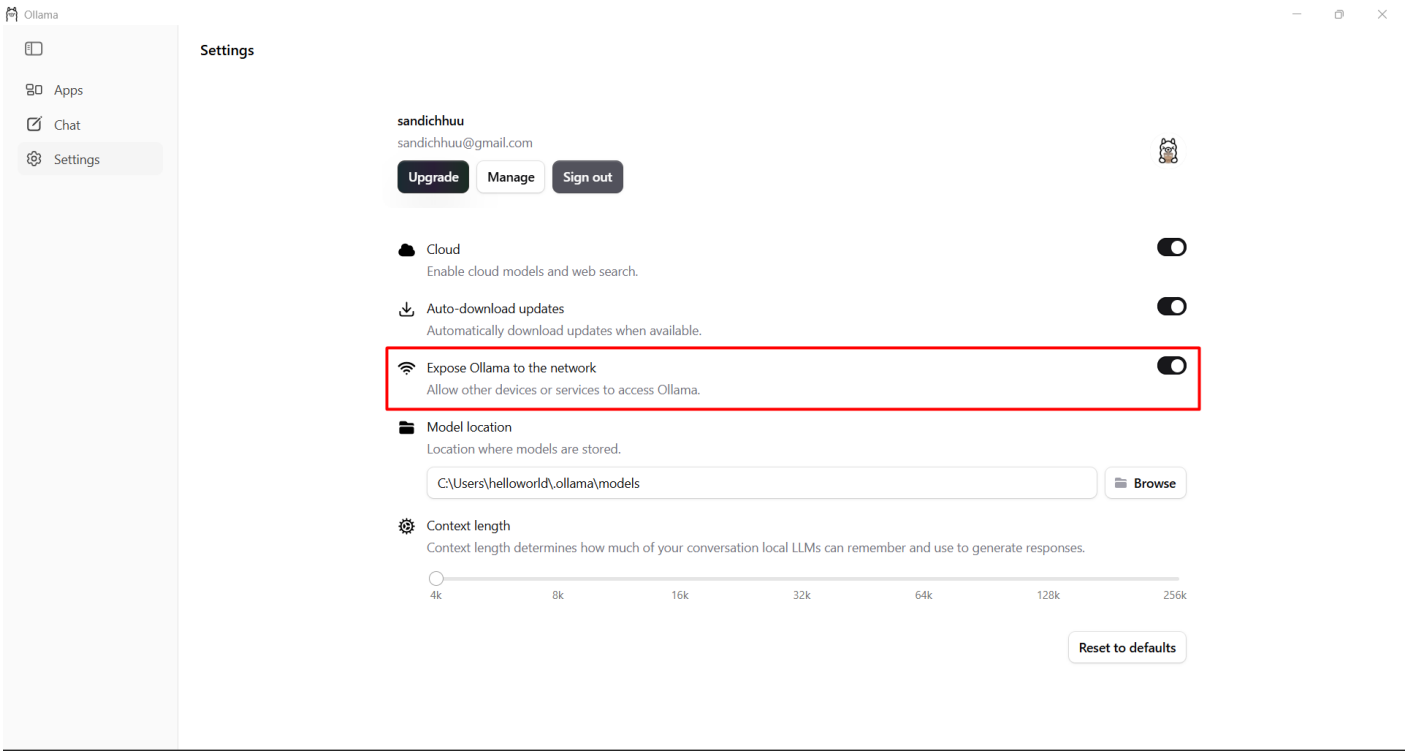
Tiếp theo hãy vào đây xem mô hình nào đang free: Ollama



Ví dụ ở đây họ đang cho free 6 mô hình này.
Trong đó chúng ta sẽ chừa tụi gpt-oss ra, mấy mô hình này hơi dởm khi code.
Khuyến nghị dùng mô hình Gemma4:31b.

Cài đặt & Cấu hình

Cài ollama bằng lệnh này trong PowerShell: `irm https://ollama.com/install.ps1 | iex`.
Sau đó vào Settings mở cái phần khoanh đỏ lên để DSH kết nối được.

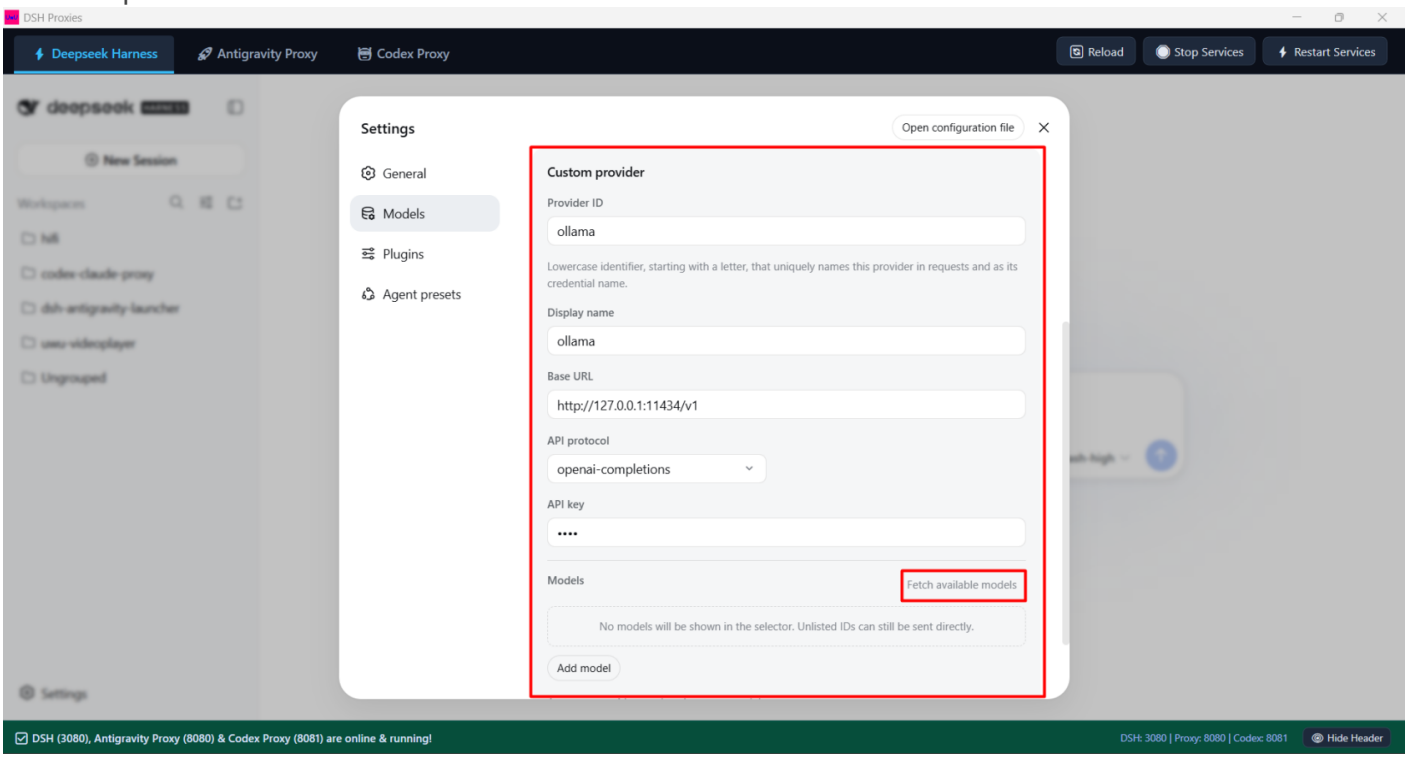


Tiếp theo chạy lần lượt các lệnh sau trong PowerShell để dùng các mô hình free:

```
ollama pull gemma4:31b-cloud
ollama pull gpt-oss:120b-cloud
ollama pull gpt-oss:20b-cloud
ollama pull nemotron-3-nano:30b-cloud
ollama pull nemotron-3-super:cloud
ollama pull nemotron-3-ultra:cloud
```

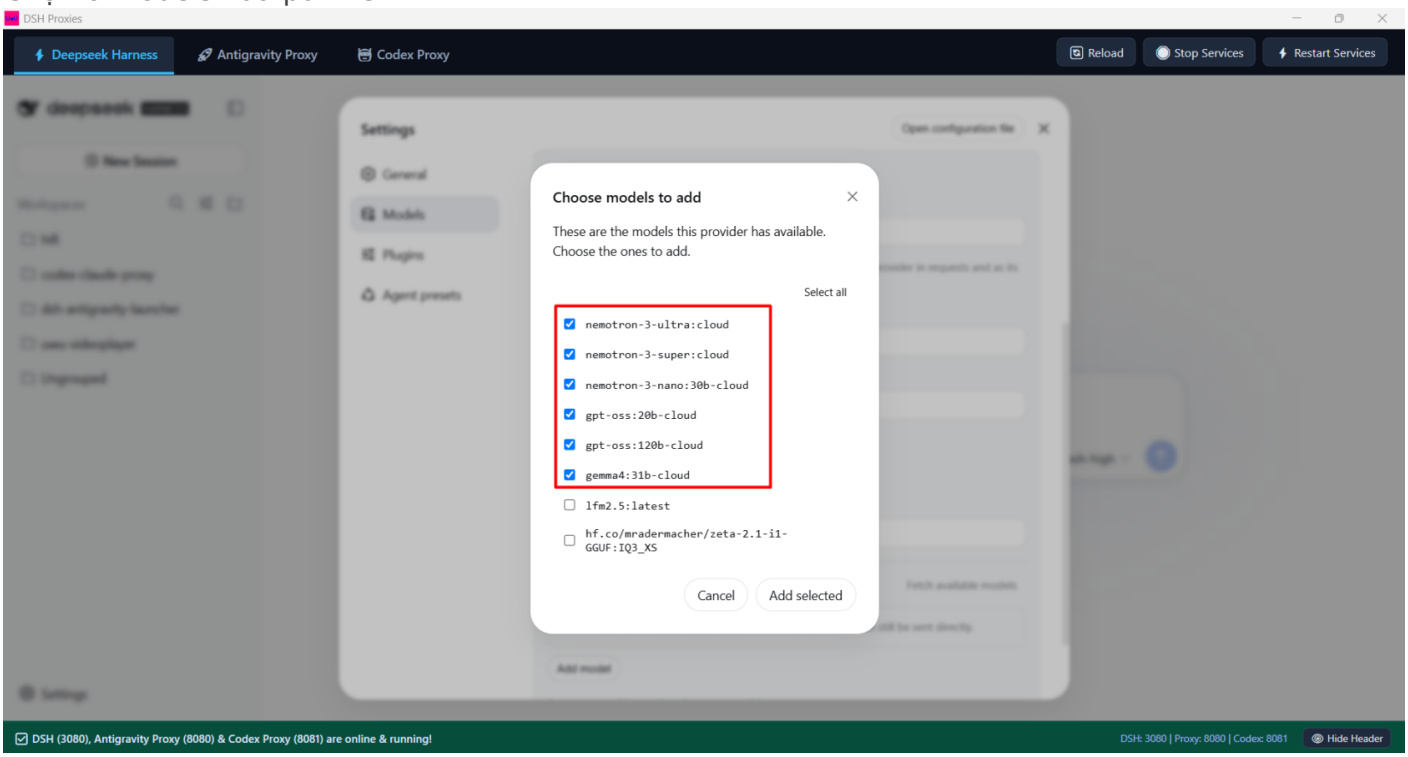
Deepseek Harness

Vào Deepseek Harness để cấu hình thêm mô hình như ảnh:



Sau đó chọn nút `Fetch Available Models`:

Chọn 6 models vừa pull về



Hiện ra như này là ok

