ローカルLLMに15パズルを開発させてみた

みやはら とおる

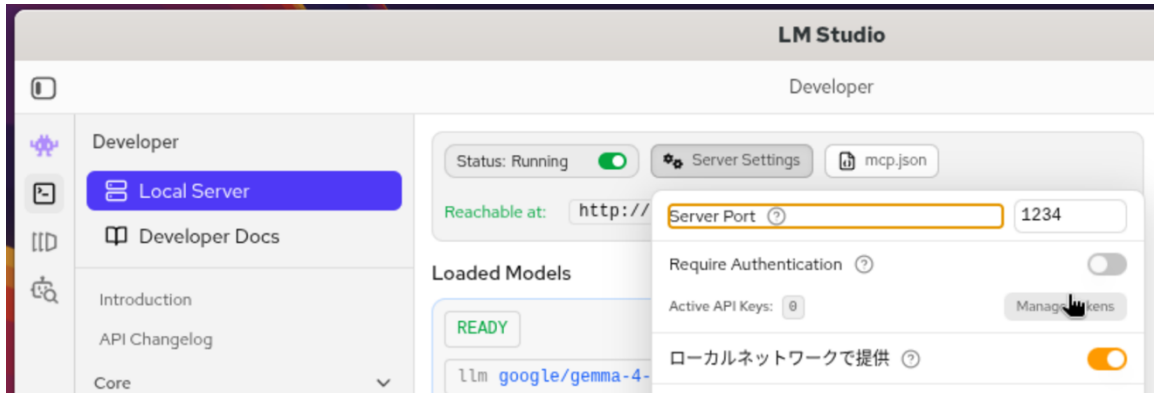
@tmiyahar

AIコーディングしてみよう

LTなのでいろいろ端折ります

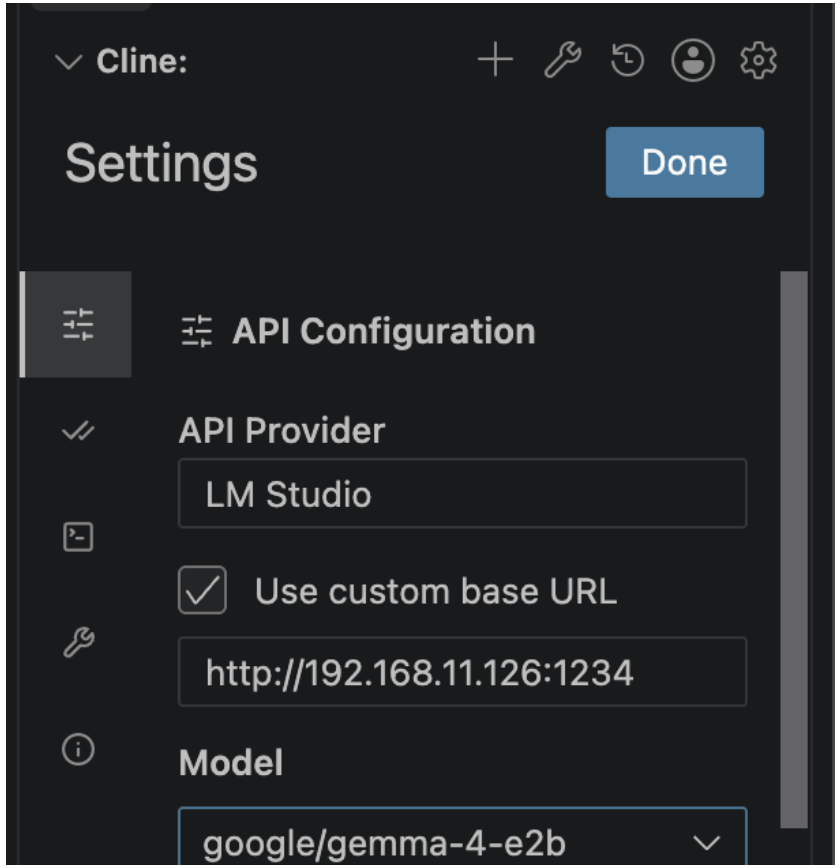
VS Code＋Cline＋LM StudioでAIコーディング

- LM StudioでLLM Serverを動作
  - Developerモードで起動
  - Server Settingsで「ローカルネットワークで提供」をオン
  - ファイアーウォールの設定を変更
    - $ sudo firewall-cmd --permanent --add-port=1234/tcp
    - $ sudo firewall-cmd --reload



Clineプラグインの設定

- VS CodeにClineプラグインをインストール
- API設定でLM Studioに接続
  - 同じマシンで動作させていれば簡単
  - 別のマシンの場合カスタムURLで
  - LLMサーバーのファイアーウォールを通せるようにするように注意
  - 接続するとモデルが選択できます



Clineの使用

- Planモードで仕様を検討
  - 細かい仕様を聞いてくるので答えてあげる
  - ファイルの書き込みなどはできない
- Actモード
  - 実装を進めてくれる
  - ファイルの書き込みに失敗するなどストレスが多い

15パズルを作ろう

15パズルとは（プロンプト）

- 15パズルを開発してください。
- Webアプリケーションです。
- パネルは15枚です。
- クリックすると空きに移動します。
- JavaScriptで開発し、フレームワークは使用しません。
- タイマーをつけてください。
- 手数をカウントしてください

テスト環境

- MacBook Pro 14”
  - M5 Max
  - メモリ128GB

- LM Studio
- VS Code
- Cline

比較したLLMモデル

- Genma 4 31B QAT Q4_0
  - サイズ：18.85GB
- Genma 4 31B Q8_0
  - サイズ：33.84GB
- Genma 4 26B A4B Q4_K_M
  - サイズ：17.99GB
- Genma 4 E4B Q4_K_M
  - サイズ6.33GB

Genma 4 31B QAT Q4_0

- 一見頑張っているように見えるが・・
- ループに入ってしまい完成させられず・・

- QAT（Quantization-Aware Training）が良くないのかな？
  - サイズ：18.85GB
- QATしていないGenmba 4 31Bを試してみよう

Genma 4 31B Q8_0

- サイズ： 33.84GB
  - QAT Q4_0の約2倍

- うまくいった！

- ただ、サイズがでかいので、普通のGPUじゃ乗らないな・・
  - GPUメモリ32GBは必要？
- モデルサイズが小さいしてみよう

Genma 4 26B A4B Q4_K_M

- サイズ： 17.99GB
  - Genma 4 31B QAT Q4_0の18.85GBと同程度
- A4Bなので、一部GPUでもそこそこ動くかもしれない？
  - 今回は全部UMAに乗っちゃいましたが

- うまくいった！

- でもやっぱりもっと少ないGPUメモリで動かしたい
- Genma 4 E4Bを試そう
  - Genma E2Bもあるけど、そこまでは・・

Genma 4 E4B Q4_K_M

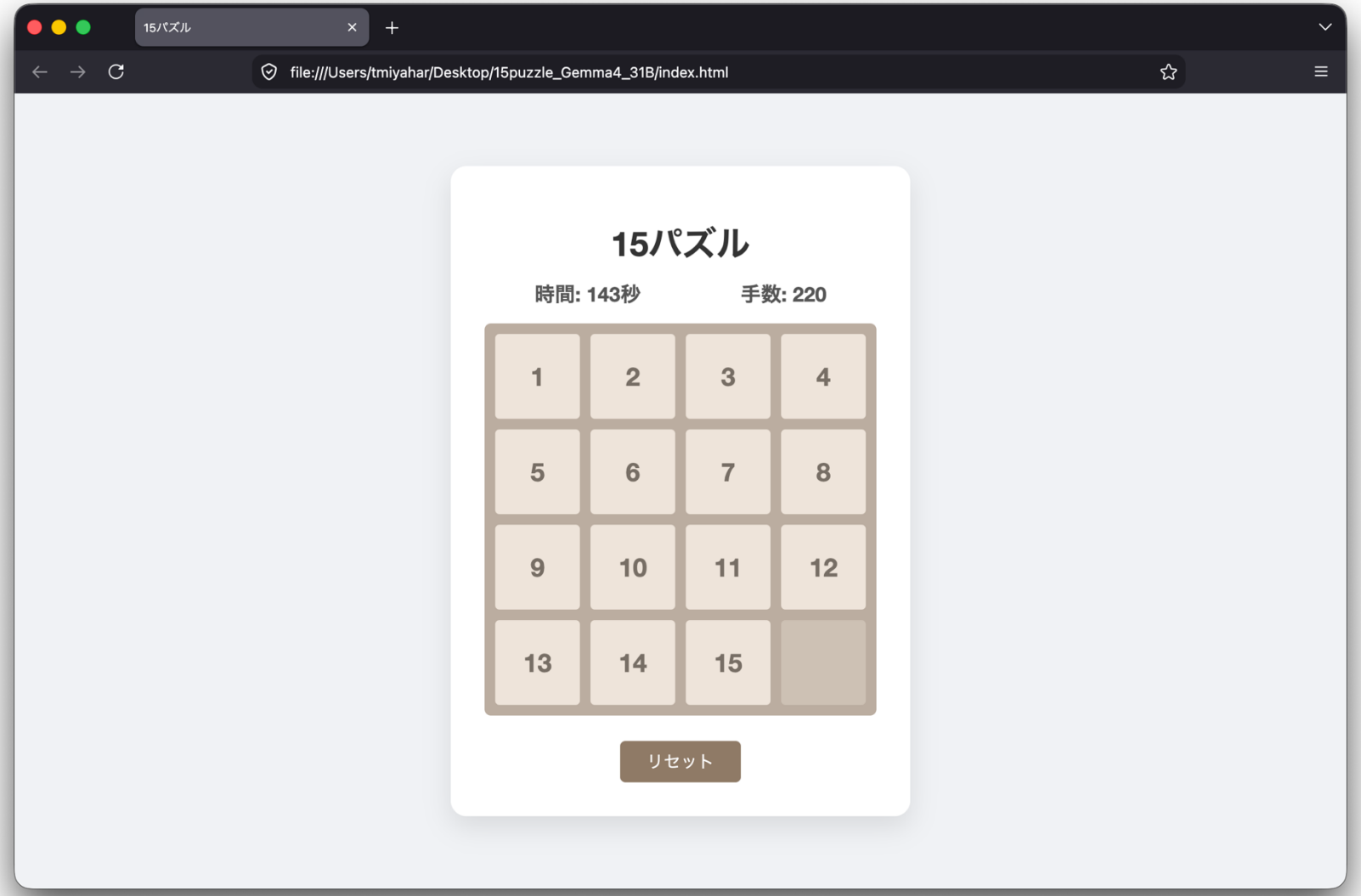
- サイズ： 6.33GB

- index.htmlが作成されない
  - コピペしろとぬかす
  - 小さいモデルだと、Tool Useが正常動作しないことが多い

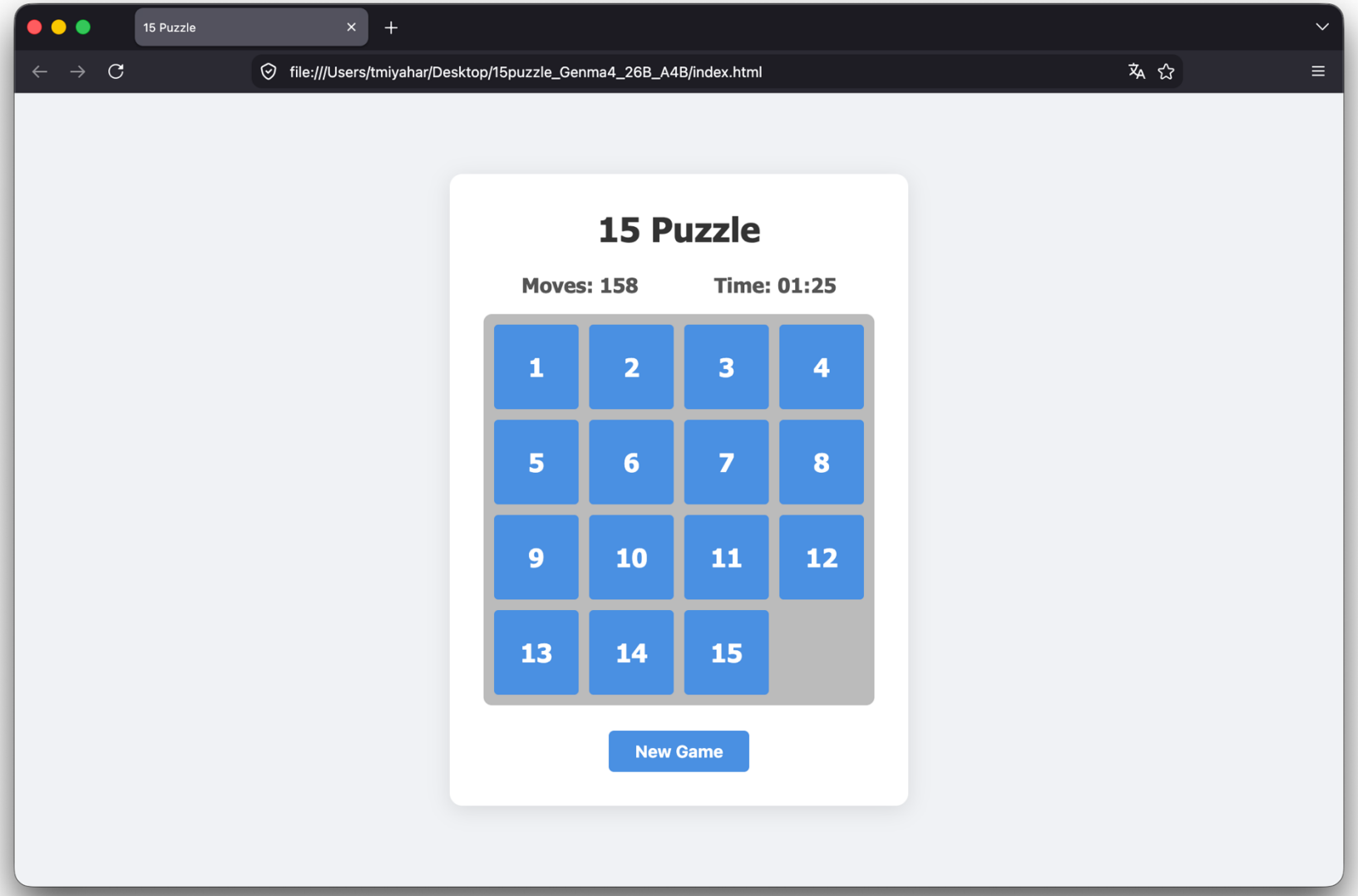
- まあ、動くといえば動くのだが・・？

結果をご覧ください

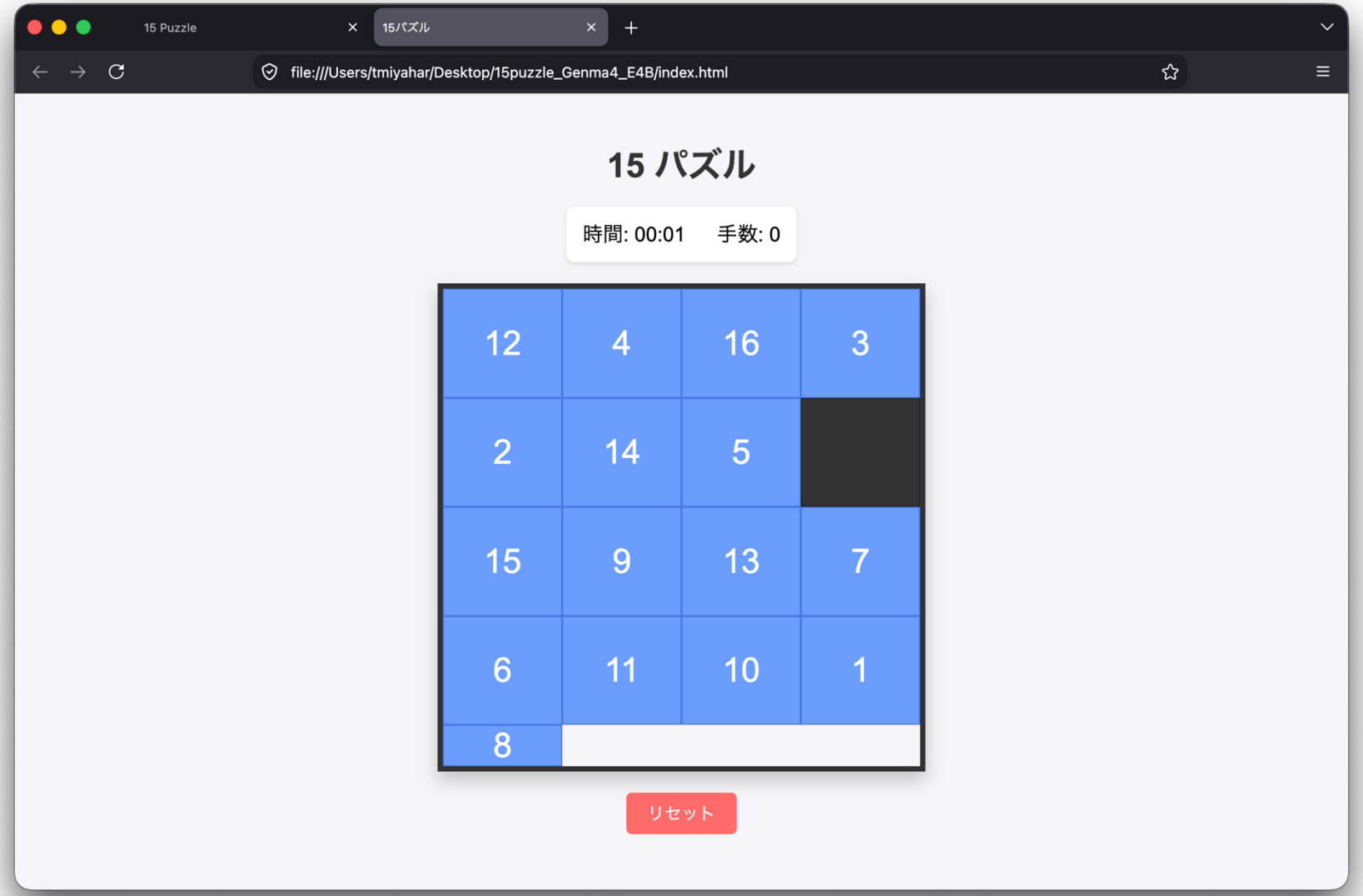
Genma 4 31B Q8_0



Genma 4 26B A4B Q4_K_M



Genma 4 E4B Q4_K_M



15パズルじゃない！

結論

- プロンプトを正しく理解できるかどうかは、ある程度のサイズのモデルでないとまだまだ難しい
- GPUメモリ8GBぐらいでは難しくて、16GBぐらいがいい塩梅？
- 本当にガリガリやりたいならAIサービス利用かもしれないが、課金してもダメなコードしか出来ない可能性もあるので、そこはガチャ感ある