Ubuntu で生成 AI 入門

Ubuntu Japanese Team
あわしろいくや

概要

- 生成 AI の基礎知識
- AMD の CPU で llama.cpp を使う方法
  - NVIDIA とか Intel は時間的に無理でした……
- 生成 AI の使用例
- オチ
- 今回の資料は後日公開予定です
  - 自己紹介ページと動画を除く

生成 AI って何？

- GPU
- LLM(Large Language Model)
- LLM を機能させるためのインターフェース
- 以上を組み合わせたもの
  - あくまでここでの定義

GPGPU

- General Purpose GPU
- ゲーム専用だった GPU を汎用的（General Purpose）にするためのフレームワーク
- GPU メーカーごとに異なる
  - AMD…ROCm
  - Intel…OneAPI/OpenVINO
  - NVIDIA…CUDA
- 最近は汎用な Vulkan でも動作
  - ゲームグラフィックス用 API
- CPU でも別に構わない
  - 処理の種類的に GPU のほうが得意

LLM

- 入力された問いに対して、確率を用いてそれっぽい回答を生成するもの
- パラメータ数
- Reasoning モデルかそうでないか
- Dense モデル /MoE モデル
- 量子化
- 例：gpt-oss-120b
  - OpenAI 製
  - パラメータ数は 117b（billion）
  - Reasoning モデル
  - MoE モデル。アクティブパラメータは 5.1b

量子化①

- モデルを圧縮する手法
  - 非可逆圧縮
- gpt-oss-120b の場合
  - オリジナルは Hugging Face 形式で配布
  - llama.cpp/Ollama で使えるように GGUF 形式に変換
  - それを量子化
  - たいていの場合、めんどくさいので誰かが量子化したモデルを使わせてもらう
    - https://huggingface.co/ggml-org/gpt-oss-120b-GGUF/tree/main

量子化②

- 量子化の手法はさまざま
- 原則としてはサイズが小さくなると高速化するが、精度が下がる
- また量子化の手法によってはバックエンドが対応しない場合もある
- Q4_K_M の使用例が多いものの……
  - モデルが充分に小さい場合は Q8 もいい
  - NVIDIA 用の NVFP4 なんてのもある
    - 今回は汎用的な MXFP4 を使用

代表的なオープンモデル

名称	会社	モデル数	Reasoning	Dense/MoE	リリース時期
gpt-oss	OpenAI	20b/120b	○	MoE	2025/08
Devstral Small2	Mistral	24b	✗	Dense	2025/12
Ministral 3	Mistral	3b/8b/14b	✗	Dense	2025/12
Qwen3-Next	Qwen	80b	○	MoE	2025/09
DeepSeek- R1	DeepSeek	1.5/7/8/14/32/70/671b	○	Dense	2025/01

インターフェース

- Ollama と llama.cpp が代表的
- LLM と PC、人間を仲介するインターフェース
  - LLM との仲介役
  - API 提供
  - Web サーバー機能
  - GPU の違いを吸収
  - LLM の管理(Ollama のみ)
- Alpaca (Ollama バックエンド)、LM Studio (llama.cpp バックエンド)
- 今回は llama.cpp を紹介

llama.cpp

- プリミティブ
  - ビルド / コマンドラインオプションが山ほど
- 開発が活発
  - たまに(？)壊れることも
- バックエンド(GPU サポート)が豊富
- 高速
  - Ollama に比べて。気のせいかも
- バイナリがリリースされているし、ソースからのビルドも簡単
- guide : running gpt-oss with llama.cpp
  - https://github.com/ggml-org/llama.cpp/discussions/15396

Why Ubuntu?

- 各社がドライバーや GPGPU 実行環境を提供している
- AMD…ROCm の Ubuntu 用パッケージを用意している。Ubuntu のリポジトリでも配布予定
- Intel…Ubuntu 用パッケージを PPA で配布している
- NVIDIA…「追加のドライバー」でプロプライエタリなドライバーを簡単インストールできる。CUDA も Ubuntu のリポジトリで提供予定？

実機の例①

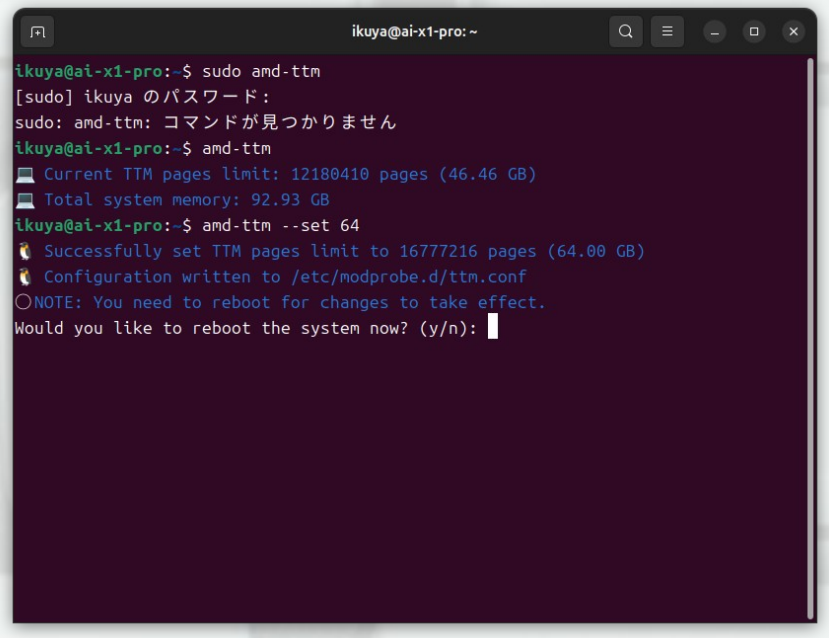
- Minisforum AI X1 PRO
  - AMD Ryzen AI 9 HX 370
  - メモリー 64GB、SSD 1TB モデルあり
  - ベアボーンだと12.5 万円くらい
  - 今回使用しているのはメモリー 96GB、SSD 1TB のベアボーンモデル
- Ubuntu 24.04.3 LTS
- gpt-oss-120b-mxfp4-00001-of-00003.gguf

実機の例②

- sudo apt update && sudo apt install linux-image-6.14.0-1017-oem
- cd ~/Downloads/
- wget https://repo.radeon.com/amdgpu-install/7.2/ubuntu/noble/amdgpu-install_7.2.70200_all.deb
- sudo apt install ./amdgpu-install_7.2.70200-1_all.deb
- sudo amdgpu-install -y --usecase=rocm --no-dkms
- sudo usermod -a -G render,video $LOGNAME
- rocminfo
- reboot
- sudo apt install git cmake libcurlpp-dev
- mkdir ~/git
- git clone https://github.com/ggml-org/llama.cpp.git
- HIPCXX="$(hipconfig -l)/clang" HIP_PATH="$(hipconfig -R)" cmake -S . -B build-rocm -DGGML_HIP=ON -DGPU_TARGETS=gfx1150 -DGGML_HIP_ROCWMMA_FATTN=ON -DCMAKE_BUILD_TYPE=Release
- cmake --build build-rocm --config Release -j 24
- ./build-rocm/bin/llama-server -m ~/Downloads/gpt-oss-120b-mxfp4-00001-of-00003.gguf --threads -1 --ctx-size 65535 --jinja -ngl 99 --flash-attn on --host 0.0.0.0 --port 8080 --chat-template-kwargs '{"reasoning_effort": "medium"}' --temp 1.0 --top-p 1.0 --top-k 0.0 --no-mmap --fit on

実機の例③

- Linux カーネルの Translation Table Manager（TTM）という機能を使うと、メインメモリーを VRAM っぽく使うことができる
- UEFI BIOS の設定で、VRAM を 1GB に
- https://rocm.docs.amd.com/projects/radeon-ryzen/en/latest/docs/install/installryz/native_linux/install-ryzen.html#configure-shared-memory



使用例①

- RSS リーダー
  - 今日日 RSS リーダーなんて山ほどあると思っていたが、意外となかった
  - gpt-oss-120b のリリース前で、ほぼ Devstral Small で作成
    - 不具合が見つかったので gpt-oss-120b で修正

使用例②

- Pukiwiki→Markdown の変換プログラム
- 既製品もあるものの、書き方の癖みたいなものがあり、全くマッチしなかった
- テストが足りなくて苦労しているものの、一応移行は完了した
  - Nextcloud の Collectives アプリが移行先
    - Markdown を Wiki みたいに見えるようにしてくれる

使用例③

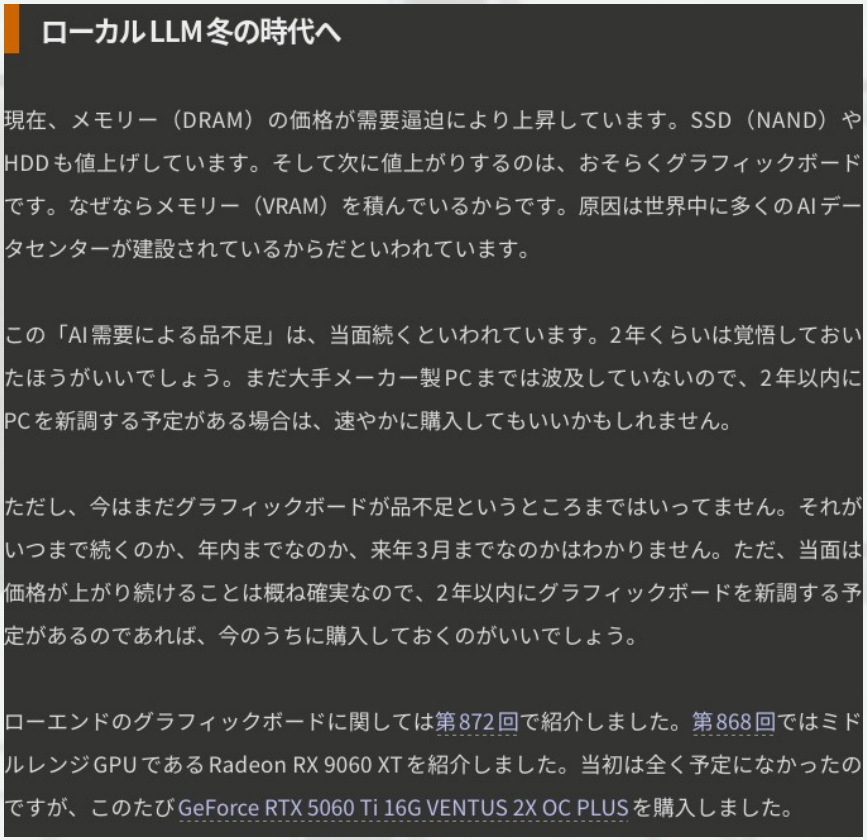
- フォルダーを監視して、ファイルが書かれたらなんかする
- こういうのばっかりなので、テンプレートを作成している
- 例：
  - どうしても古い CUPS でないと印刷できないプリンターがあって、LXD のコンテナーで Ubuntu 20.04 LTS を動作させ、特定のフォルダーに PDF を書き出したら”lxc exec（コンテナ名）-- lp（以下略）”のコマンドを実行する

使用例④

- Qwen3-VLでOCR
- Ubuntu Weekly Recipe 第897回
  - https://gihyo.jp/admin/serial/01/ubuntu-recipe/0897

悲報

- Ubuntu Weekly Recipe 第891回より
  - 2025/12/10 公開



ローカルLLM冬の時代へ

現在、メモリー（DRAM）の価格が需要逼迫により上昇しています。SSD（NAND）やHDDも値上げしています。そして次に値上がりするのは、おそらくグラフィックボードです。なぜならメモリー（VRAM）を積んでいるからです。原因は世界中に多くのAIデータセンターが建設されているからだといわれています。

この「AI需要による品不足」は、当面続くといわれています。2年くらいは覚悟しておいたほうがいいでしょう。まだ大手メーカー製PCまでは波及していないので、2年以内にPCを新調する予定がある場合は、速やかに購入してもいいかもしれません。

ただし、今はまだグラフィックボードが品不足というところまではいってません。それがいつまで続くのか、年内までなのか、来年3月までなのかはわかりません。ただ、当面は価格が上がり続けることは概ね確実なので、2年以内にグラフィックボードを新調する予定があるのであれば、今のうちに購入しておくのがいいでしょう。

ローエンドのグラフィックボードに関しては第872回で紹介しました。第868回ではミドルレンジGPUであるRadeon RX 9060 XTを紹介しました。当初は全く予定になかったのですが、このたびGeForce RTX 5060 Ti 16G VENTUS 2X OC PLUSを購入しました。

GPU の入手状況

- NVIDIA
  - dGPU… 価格高騰中。買えなくなりつつある。次のモデルは 2027Q2 以降とのこと
  - iGPU…DGX Spark のみ。純正モデルは 80 万円超。互換モデルもある
- AMD
  - dGPU… ハイエンドモデルはなし。NVIDIA よりは入手性はマシ。価格も高騰中
  - iGPU… メモリー高騰の影響をもろに受ける
- Intel
  - LLM 用に選択するのはあまりおすすめしない
    - というか Linux 上の Vulkan 遅すぎない？
    - リストラの影響がいろいろと…… RIP IPEX-LLM

展示 PC の価格比較

- Minisforum AI X1 PRO ベアボーンモデル
  - 111,190→124,799 円
  - 96GB/2TB/Win 11 Pro で 268,000 円くらい
- Crucial CT2K48G56C46S5
  - 36,580→155,800 円
- KIOXIA SSD-CK1.0N4P/N
  - 11,480→ (廃番)
- KIOXIA SSD-CK1.0N4PLG3N or J
  - 8,390→25,280 円

AI PC?

- マイクロソフトのマーケティング用語
  - 40TOPS 以上の速度を出せる NPU を搭載している
- Linux でも使用可能ではある
  - Intel…OpenVINO
    - まだ使い道はある。llama.cpp のバックエンドも開発中
  - AMD…XRT
    - 今のところ使い道なし。ビジネス向けの Ryzen AI Softare は存在する
- そもそも 40TOPS 程度では遅すぎる。
  - もう少し技術革新が必要
  - NPU が速くなるか、あきらめるのか、どっちが先か？

まとめ

- オープンな LLM のモデルがたくさん出てきた
  - gpt-oss-120b は頭一つ飛び抜けている
- i/dGPU の高速化により、大きなモデルでも実用的な速度で動作させるようになってきた
- ただし昨今のメモリーの値上げでいろいろ台なし
  - このネタを考えたときはもっとメモリーが安かったんだよ……
  - ~~サム・アルトマンめ……~~

告知

Web 連載

- Ubuntu Weekly Recipe
  - https://gihyo.jp/list/group/Ubuntu-Weekly-Recipe
- Ubuntu 日和
  - https://pc.watch.impress.co.jp/docs/column/ubuntu/

雑誌

- Software Design 2026年1月号
  - 始めるなら今!? 最新情報と活用のポイント
  - デスクトップLinux元年 in 2026
- https://gihyo.jp/dp/ebook/2025/978-4-297-15155-3

書籍

- はじめてのUbuntu
  - 2025/11/18発売
- Ubuntuの入門書