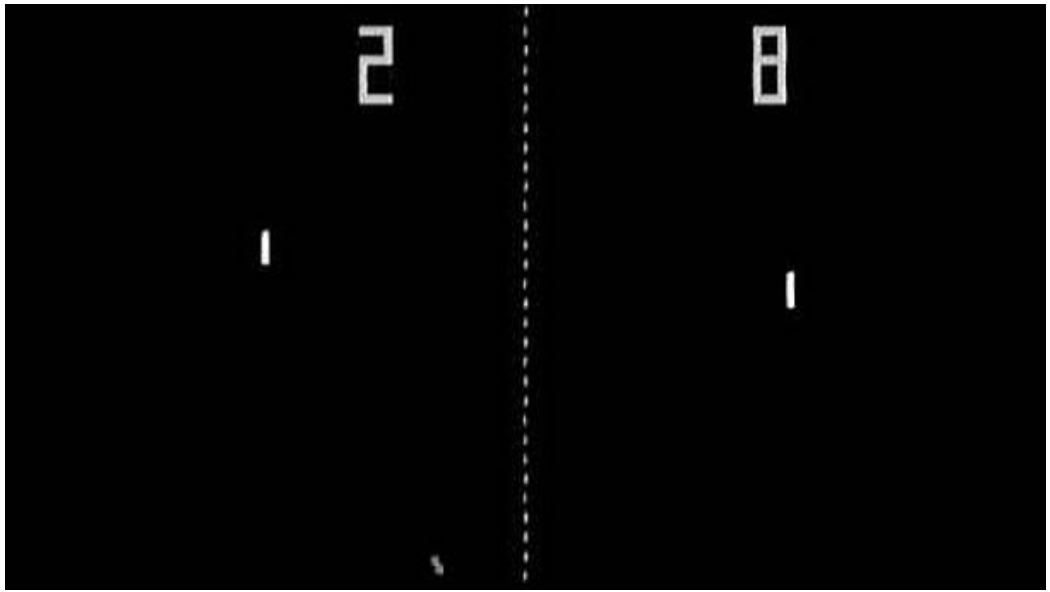
Open Source AI

2nd Oct 2024

What do we mean by AI

- **Ability of machines to perform tasks normally requiring human minds**
- **From a Pong Opponent**



what we mean by AI (cont…)

- To the latest Generative AI (like Sora in the video : link)
- People normally mean this now.

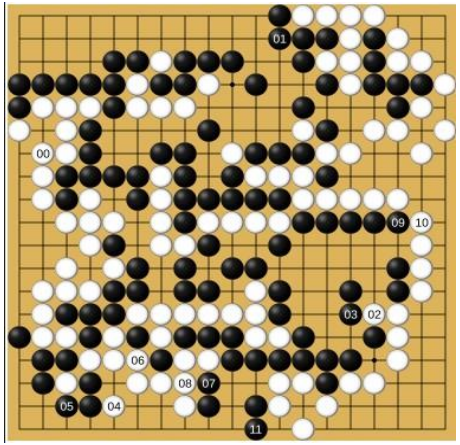
What we mean by Open Source AI

- Usually expect that everything needed to use the software is publicly available.
- License that allows anyone to build, modify and run the software.
- For Machine Learning, not just the code or traditional assets that we need to worry about.
- Weights (huge arrays of numbers that populate the neural network) are needed too.
- ‘Model’ in the mathematical sense, Artificial Neural Network is the ubiquitous implementation. Training is the process of using data to get this model to be as accurate as possible.
- “Model” is the whole system including the code and the weights.

OpenAI, open source AI & cloud services.

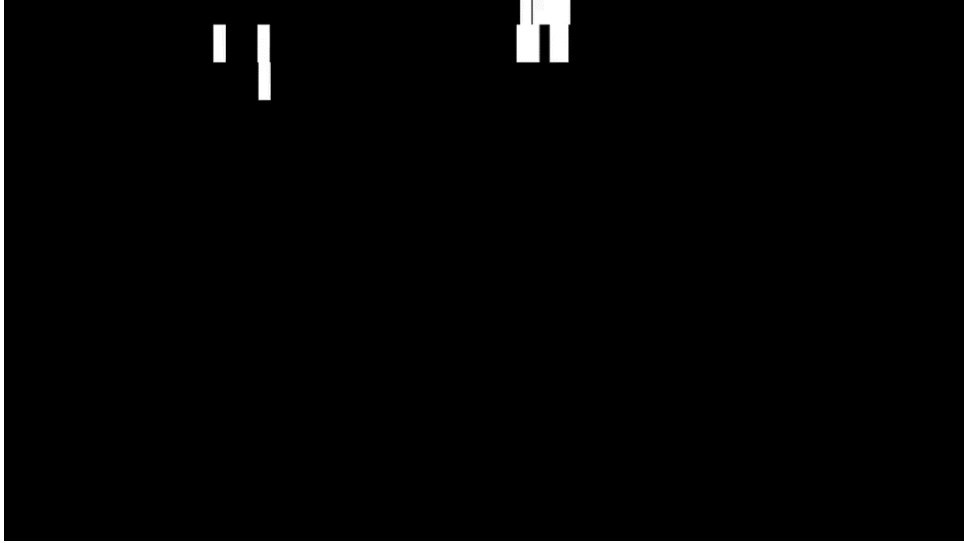
- OpenAI have released some important Open Source AI projects.
- Their most powerful recent models GPT4, Sora, DALL-E-3 very much closed source.
- Some controversy given their original mission statement.
- Closed source models (from different providers) are frequently available as a service.
- Reasons given are usually some combination of protecting profits and concerns over misuse.
- Debate around safety of openly releasing models.

Leela Zero - 2017



- Go proved far harder to create computer players for than chess.
- DeepMind's AlphaGo defeating the World Champion has been referred to as China's "Sputnik moment" for AI.
- DeepMind published papers detailing the code and process, open source project **Leela Zero** recreated it.
- Leela Zero released in 2017, lacked the training required to compete at AlphaGo levels.
- A distributed community effort trained model weights that can defeat the best human players.
- Leela Zero available on GitHub with links to weights and GUIs if you want a challenge.

Stable Diffusion - 2022



[Link to SDXL video]

- Some previous open source image generators gave great results but not from text promps in the way DALL-E could, for a while there was nothing open source that compared to DALL-E.
- Stable Diffusion was openly released, produced images from text prompts very close in quality to DALL-E.
- XL variant even better.
- Many variants for all kinds of images, filling in the blanks, turning sketches photorealistic, replacing content.
- From version 3 the license has completely changed, now less open.

SDXL Lightning - 2024

From ByteDance, based on XL version of Stable Diffusion.
Similar output quality while requiring far less compute time.
Same license as the earlier versions of Stable Diffusion.

Flux Schnell & Flux Dev - 2024

Larger Flux-Pro remains closed source.
Flux-Dev does have license restrictions limiting commercial uses.
Smallest is Flux Schnell, highly regarded, permissive license, new focal point of open source image generation.

Open source matters

- Unlike with a model on your own hardware, access to a service can disappear and you need to trust the service provider with your data.
- Open-source models can be customised and innovated upon by fine tuning which is an extra bit of training.
- Community fine tuned variants of original models are widely available, including on Hugging Face hub.
- Pictures on the left and right are from different fine tunes, middle is the original model, all using the same prompt.

Whisper - 2022

OpenAI still contributed to the open source scene even recently.

Whisper is excellent at transcribing voice audio.

Better than real time performance on consumer hardware with minimal errors, 30 minute podcast transcribed in 5 minutes on a laptop.

BERT and GPT1 - 2018

- Landmark academic paper “Attention is All You Need” by research team at Google – introduced Transformer Architecture.
- BERT model from Google soon followed, released on GitHub.
- BERT used the context of the surrounding words and even previous sentences in a way that was unprecedented to turn text into numerical representation of meaning.
- Models like BERT are used everywhere now that software needs to act on the meaning of text, began powering Google Search in 2019.
- OpenAI published paper with GPT1, outperformed all previous AI text generation.

Generative Text

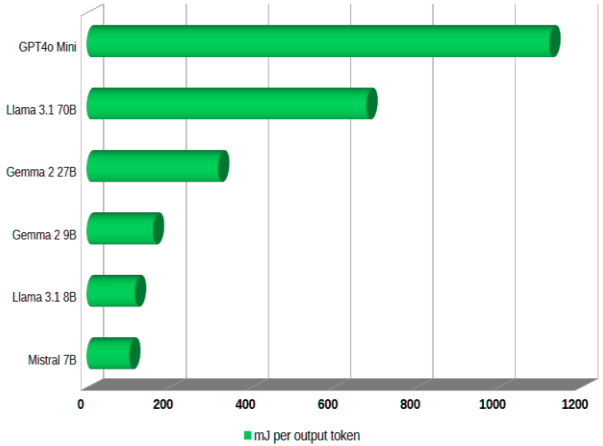
- After GPT3’s release, open source generative text was a long way behind.
- GPT-Neo series from EleutherAI led in period 2021-2022, beyond GPT2 performance, not really GPT3 level.
- Chat models are text generation models too, just fine-tuned on conversations and tasked with generating the next answer.

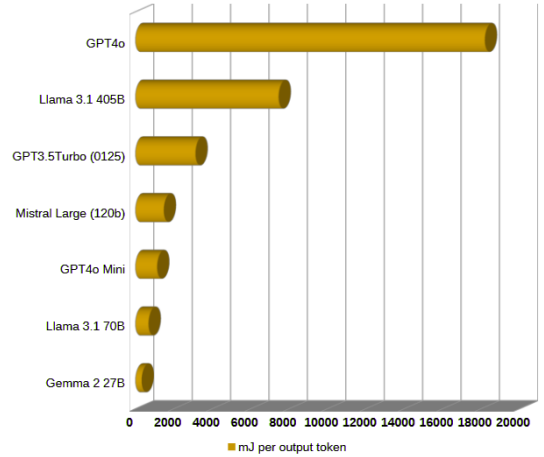
Llama 2 & other generative text - 2023

- In 2023 the gap narrowed.
- Meta released Llama 2 openly, this surprised the community.
- Tii-UAE released the Falcon series, permissive license.
- Different sizes and chat variants of Llama2 and Falcon.
- MistralAI 7b model, excellent outputs given its size.
- Recently, Llama 3.1 & 3.2, along with Falcon 2 have been released, retain their predecessors' permissive licenses.
- Largest Llama (405b) is very competitive with closed source services, however challenges just running it due to its size.
- Google released Gemma, 'baby' versions of Gemini.
- Multimodal variants, take image inputs, not only text : Llava, Llama 3.2

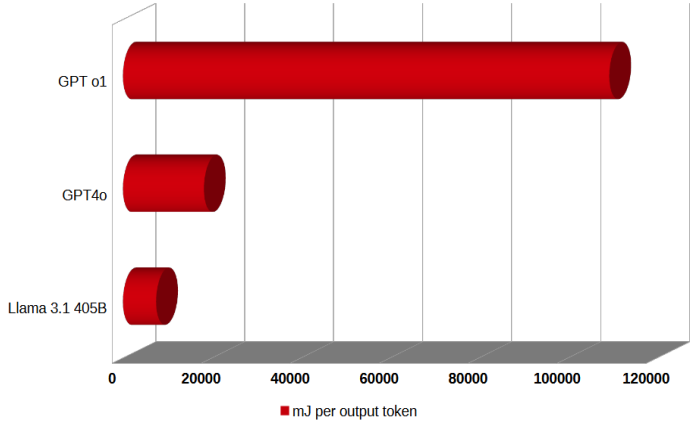
A note about energy

- Energy used varies massively with model size.
- GPT o1 uses a 1000x more than small text generation models..





- Generating a page of text with GPT o1 : 110kJ, energy in a small biscuit or used driving 200m in Nissan Leaf.
- Page of text with Llama 3.1 8b : 0.12kJ, less than in a crumb of biscuit or driving 20**cm** in Nissan Leaf.
- Bigger models do usually mean better output in a wider range of scenarios.
- Fine tuning : smaller models can give better output in specific use cases than a larger general purpose model.
- Open source community sharing fine tuned variants, many to choose from.
- Fine tuning itself costs energy but savings much larger if it's used by many users.



Code Generation

- There are open source alternatives for generating and working with code.
- **CodeLlama** : official variants of Llama 2 from Meta. Excellent at code completion. Instruction following versions haven’t stood the test of time.
- **DeepSeek-Coder-V2 :** now tops open source leaderboards for code instruction following.
- StarCoder & CodeGemma : good reputations given relatively small size.
- VSCode plugin to use the above for code completion in a similar way to copilot.

Audio Generation

- MusicGen family from Meta AI can produce music clips from a text prompt
- MusicGen’s license : only for non-commercial use.
- Others with permissive licenses not as good.
- Mustango the pick of these.

Frameworks

- Open source frameworks underpin all of this.
- Tensorflow from Google, PyTorch from Meta : abstract neural network code and implementations for different hardware.
- Both are written in C++, well supported APIs in Python and C++, other languages less well supported & less complete.
- GGML : lightweight way to run LLMs on CPUs / consumer GPUs, better CPU performance, quantises model weights so more can fit in memory.
- Llama.cpp, Ollama & KoboldCPP are built on top of GGML, facilitate running Llama & Gemma without touching the code.
- Hugging Face Transformers & Diffusers are higher level Python frameworks that sit on top of PyTorch or Tensorflow, good starting points for training or fine tuning.

User Interfaces

There are a number of friendly user interfaces for generative text and images.

ComfyUI, Swarm UI & Forge all allow you to use generate or modify images using SD, Flux and many of their derivatives within a graphical user interface.

KoboldCPP allows you use Llama, Falcon, Gemma, their derivatives and some older things like GPT2 within an easy to use browser based UI (served locally from your own machine). It has modes for chat and generative writing, along with a ‘role playing’ mode that was its original focus. It is becoming something of a one for all, with UI for Image Generation (SD) and speech<->text too.

A few other text generation UIs that are worth a mention are Oobabooga, GPT4All and SillyTavern.

Computer Vision

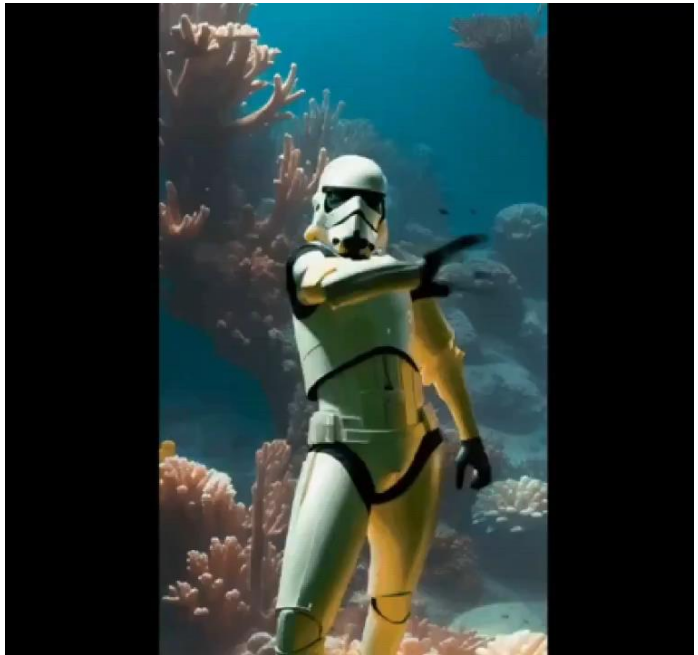
- Vast number of open source computer vision projects
- Getting attention recently is Segment Anything 2, yet again from Meta AI.
- Will pick out all the objects in a complex scene

The Bad Stuff

- Misinformation and worse being enabled by AI.
- Deep Fakes
- Generative text - spreading misinformation and swaying public opinion.
- Harder to stop with open source AI compared to services.
- GPT4Chan fine tuned on messages from 4Chan, generated it’s own novel hate speech and abusive messages.
- LLMs being fine tuned to push a certain ideologies, agendas, misinformation campaigns.
- Digital plagiarism, additional issue with open source, can target plagiarism very specifically with fine tuning.
- License terms forbidding it are no barrier to bad actors.

Video Generation

- Sora is not available as a service yet.
- Runway’s Gen3alpha is the leading (available) closed source model – well ahead of open source right now.
- **AnimateLCM** (2024) : Open model with good quality clips but very short (link)



Video Generation (cont)

Animate-Diff (2023) also very short in duration.

Optimised AnimateDiff-Lightning (2024) from ByteDance. Video to video variant can potentially make longer clips. Example video is very ‘TikTok’ (video link)

Video Generation (cont)

ModelScope, the derived **ZeroScope** series (2023) & **VideoCrafter** (2023) : longer clips but quality lower. (video link)

Stable Diffusion Video (2024) better quality than ZeroScope, slightly longer than AnimateLCM, no commercial use.

What you need to run these

Much will work on a modern mid-range laptop, images may take a minute on CPU. As the size goes up you do start to need different hardware and lots of memory.

NVidia GPUs still dominate this whole space

AMD GPUs are supported but not as well as NVidia.

Many modern integrated processors have specialist cores (Apple led the way on this), still big gap to NVidia.

Cloud options – lose the privacy benefits but at least you can always change hosting provider.

Thank you