AI Info

General info

Large language models vs small language models:
https://venturebeat.com/ai/why-small-language-models-are-the-next-big-thing-in-ai/

https://www.toptal.com/machine-learning/machine-learning-theory-an-introductory-primer
https://www.databricks.com/glossary/machine-learning-models
https://www.finextra.com/blogposting/20028/how-is-ai-revolutionizing-fx-market-in-a-way-we-didnt-even-realize

Transformer Model file types: GGUF: binary model format optimised for quick loading and saving. Uses GGML as executor. used by llama.cpp framework. successor file format to GGML, GGMF and GGJT PyTorch: Models usually trained in PyTorch. Format can be converted into GGUF format. (https://huggingface.co/docs/hub/gguf) Safetensors: Safetensors is a format for storing tensors safely (as opposed to pickle) and that is still fast (zero-copy). 2.1x faster than pytorch on gpu, 76.6x faster on cpu

Instruct/chat models: These models 'expect' to be involved in a conversation with different actors. In contrast non-instruct tuned models will simply generate an output that follows on from the prompt. If you are making a chatbot, implementing RAG or using agents, use instruct or chat models.

Embedding models: Embedding models are used to represent your documents using a sophisticated numerical representation. Embedding models take text as input, and return a long list of numbers used to capture the semantics of the text. These embedding models have been trained to represent text this way, and help enable many applications, including search! NV Embed model

Frameworks/GUIs: llama.cpp: https://github.com/ggerganov/llama.cpp GGML: Tensor library for machine learning https://github.com/ggerganov/ggml GPT4All: https://gpt4all.io/

Terminology

- Inference: the process that a trained machine learning model uses to draw conclusions from brand new data. (coming up with a solution to a question) - Q2/Q5/etc: These are quantisation standards used by llama.cpp/ggml, q4 means quantize to 4 bit, and q5 to 5 bit. An offset might be added like q4_1 meaning improved precision. This can be compared to lossy image compression. 2bit is highest compression, smallest filesize, but a lot less accurate than the original model. see GGML Quantization diff q4/q5 Guide to quantization - SOTA: SOTA is an acronym for State-Of-The-Art. It refers to the best models that can be used for achieving the results in an AI-specific task. - GPT: Generative Pre Trained - Parameters: 7B/14B/130B/etc. This means the amount of parameters of pre-trained models in billions. It represents the learned variables acquired during the model's training process. They enable the model to capture language patterns, resulting in context-based predictions in response to input. (i.e. amount of brain cells of the AI) - LoRA (Low-Rank Adaptation) is a fine-tuning method developed by MS researchers in 2021. LoRA is a type of Parameter-efficient Fine-tuning (PEFT). RAG: Retrieval-augmented generation (RAG) is a technique for enhancing the accuracy and reliability of generative AI models with facts fetched from external sources. (using databases, files etc) - Tensors: A tensor is

an algebraic object that describes a multilinear relationship between sets of algebraic objects related to a vector space. Tensors are a generalisation of scalars and vectors; a scalar is a zero rank tensor, and a vector is a first rank tensor. The rank (or order) of a tensor is defined by the number of directions (and hence the dimensionality of the array) required to describe it. It can be thought of as a multidimensional numerical array. An example of a tensor would be 1000 video frames of 640×480 size. - Vector: Vectors are used to represent both the input data (features) and the output data (labels or predictions). Each data point is represented as a feature vector, where each component of the vector corresponds to a specific feature or attribute of the data. - Temperature: Temperature in AI settings control the diversity and creativity of the generated text. Value ranges between 0 and 2 depending on model. Higher temperature values result in more diverse outputs but less logically coherent, while lower values lead to more focused and deterministic text. - GraphRAG RAG using a graph database instead or in addition to a vector database to return additional knowledge context to an LLM to answer a question or request.

Evolution of AI

Traditional AI: focuses on analyzing historical data and making future numeric predictions ('standard machine learning') example is auto complete suggesting the next word by likelihood and frequency of previously used words.

Generative AI: allows computers to produce brand-new outputs that are often indistinguishable from human-generated content. Can become toxic or show harmful behaviour (early ChatGPT like systems)

Generative AI with RLHF: Reinforcement Learning from Human Feedback (RLHF) is a method to adjust the training of the AI with human preferences. Feedback collected from humans for tasks/responses of the AI will help the AI understand people's preferences and learn to generate better output. (Thumbs up/thumbs down in chats). Thousands of responses are required. (online ChatGPT like systems with many users) Red-teaming attempts to trigger toxic responses to train an AI to avoid them, but requires teams and time to test and come up with triggers. A new AI trained to be curious to find triggers can automate this. Safe AIs were tested and almost 200 toxic responses were triggered quickly. https://techxplore.com/news/2024-04-faster-ai-chatbot-toxic-responses.amp

Constitutional AI: provides a transparent method of reducing the toxicity and harmful behaviour exhibited by generative language models. It uses a set of rules or principles that act as a “constitution” for the AI system. RLHF is scaled and evaluated with principles to improve if they exist. (Claude / others)

https://www.linkedin.com/pulse/navigating-responsible-ai-constitutional-human-loops-vijay-chaudhary

Possible next step/future is structured approach: Symbolica is trying to remove the unknowable black-box in Generative AI's decision making with more rigorous, scientific foundation. They're using the mathematics branch of 'category theory' to formalise mathematical structures and their relationships. With that they believe they can create models that have reasoning as an inherent capability, rather than an emergent side effect of training on huge datasets. This avoids training with massive amounts of data and using very large models.
https://venturebeat.com/ai/move-over-deep-learning-symbolicas-structured-approach-could-transform-ai/

Objective driven AI:
https://bernardmarr.com/generative-ai-sucks-metas-chief-ai-scientist-calls-for-a-shift-to-objective-driv

en-ai/

The Big Players

Company	Source	Model	Type/Notes
OpenAI	Closed	ChatGPT3.5/4	text AI
OpenAI	Closed	DALL-E/2/3	image generator
OpenAI	Open	Whisper	speech recognition/transcribe/translate
Anthropic	Closed	Claude/2/3 Haiku,Sonnet,Opus	text Constitutional AI
Stability.AI	Open	Stable Diffusion/2/3/XL	image generator
Stability.AI	Open	Stable LM/2	text chat
Microsoft	Closed	Microsoft Copilot	Uses ChatGPT4 with Dall-E (replaces Cortana, formerly Bing Chat or Bing AI)
Microsoft	Open	WizardLM/2	chat AI
Microsoft	Open	Phi-2/3	chat AI
Google	Free	Gemini (formerly Bard)	Chat AI
Google	Free	CodeGemma	Coding AI
Meta	Open	Llama/2/3/Code Llama	Chat AI
Meta	Closed	Meta AI	Imagine (image) and chat (Llama)
MidJourney	Closed	MidJourney	Image generator AI
Falcon AI	Open	FalconLLM	chat AI https://falconllm.tii.ae/
Mistral	Open	Mixtral and others	trained on Llama2 70B and outperforming it as well as ChatGPT 3.5 with faster inference
Mistral	Open	MistralAI	text chat
Reka	Closed	Reka Core	on par with GPT-4 and Claude 3 Opus
Alibaba	Open	Qwen/1.5	Text chat with vision and audio understanding
HuggingFace	Open	Idefics/2	Vision text model

https://www.marktechpost.com/2024/04/16/wizardlm-2-an-open-source-ai-model-that-claims-to-outperform-gpt-4-in-the-mt-bench-benchmark/?amp
https://venturebeat.com/ai/hugging-face-introduces-idefics2-an-8b-open-source-visual-language-model/ https://huggingface.co/blog/codellama
https://ai.meta.com/blog/code-llama-large-language-model-coding/
https://venturebeat.com/data-infrastructure/snowflake-copilot-a-mistral-large-powered-ai-assistant-launches-in-public-preview/

Open=Open Source, Free=Closed source with local use, Closed=Closed source and only free or paid cloud usage.

Not confirmed info about copilot (enterprise?free?) vs chatgpt: Microsoft 365 Copilot is grounded against your tenant data, while ChatGPT (including the Pro version) is not. Grounded means - having access to the data and using it as its “source of truth”. In addition, Microsoft 365 Copilot is not getting trained on your company data and information. Whenever you prompt/send a query, it has to look for the information. Once the response was given, Copilot forgets about what it had just found to ensure customer data stays within the customer's premises.

Features / Use

Free versions:

Claude 3 Sonnet, free, no live internet access, data to Aug 2023, Human writing style, verbose, largest context window (memory of chat) of 200000 default, up to 1m. Needs google account and phone number verification. Images can be uploaded. No image generation.

Gemini 1, free, needs google account. Live interner access. Good for programming, integration google docs. Very moderated. Image generation included.

Chatgpt 3.5,free,needs google or ms account. General AI, no live internet access, image generation separate with dall-e 3. Issues with making up data, forgetting beginning of conversation, too confident in answers even when wrong.

Mistral 7B / Orca, on par with llama2 70B model for local use, outperforms GPT-4 in some tasks, good for text analysis.
(https://predibase.com/blog/lora-land-fine-tuned-open-source-llms-that-outperform-gpt-4)

Copilot, based on chatgpt 4 with Dall-E and with MS enhancements and integration with MS infrastructure. Live internet access. Text chat without account, image generation requires personal MS account.

News / Bugs / others

https://www.reuters.com/technology/new-york-city-defends-ai-chatbot-that-advised-entrepreneurs-break-laws-2024-04-04/
https://www.good.is/company-introduces-ai-bot-to-help-workers-with-queries-but-it-starts-glitching-in-weirdly-funny-way

https://www.marktechpost.com/2024/03/27/stability-ai-introduces-stable-code-a-general-purpose-base-code-language-model/ https://github.com/PromtEngineer/localGPT?tab=readme-ov-file

https://www.nvidia.com/en-gb/ai-on-rtx/chatrtx/
https://www.tomshardware.com/tech-industry/artificial-intelligence/nvidias-new-ai-chatbot-falls-victim-to-high-severity-security-vulnerabilities-urgent-chatrtx-patch-issued#xenforo-comments-3840821

https://www.marktechpost.com/2024/03/25/meet-devika-an-open-source-ai-software-engineer-that-aims-to-be-a-competitive-alternative-to-devin-by-cognition-ai/
https://github.com/stitionai/devika/issues/204 https://github.com/mudler/LocalAI
https://localai.io/basics/getting_started/index.html https://localai.io/docs/reference/aio-images/

Image/Video AIs

https://decrypt.co/219776/ideogram-is-a-new-ai-image-generator-that-obliterates-the-competition-outperforming-midjourney-and-dall-e-3

https://news.mit.edu/2024/ai-generates-high-quality-images-30-times-faster-single-step-0321

https://decrypt.co/218577/stable-diffusion-3-review-comparison-midjourney-dall-e-imagefx

https://www.theguardian.com/commentisfree/2024/feb/24/openai-video-generation-tool-sora-babies-ai-artificial-intelligence

Cerule: https://huggingface.co/Tensoic/Cerule-v0.1

Redbox / Claude

Redbox is a framework in development by the UK government to assist civil servants. It's meant to be used primarily with Anthropic's Claude which is a constitutional AI and as such designed to prioritise accuracy and correctness over proving answers at any cost and to explain its reasoning and sources.

https://ai.gov.uk/projects/redbox-copilot/ https://github.com/i-dot-ai/redbox-copilot
https://claudeai.uk/claude-ai-model-download/#:~:text=Download%20and%20Install%20Claude%20Locally,yourself%20on%20a%20local%20device. https://www.bbc.com/news/uk-politics-68724499.amp

text summarisers

https://www.reddit.com/r/ChatGPTPro/comments/14eprse/are_there_any_good_free_gptpowered_ai_summarizer/

https://github.com/callstack/ai-summarization → .env file support fork
https://github.com/matkoson/ai-summarization
https://huggingface.co/TheBloke/Mistral-7B-OpenOrca-GGUF

https://www.linkedin.com/pulse/deploying-llama2-locally-docker-ocr-text-hashir-khan-vvwvf

llama/ollama/llama.cpp

ollama is a compatibility framework for llama.cpp translating chat requests for the models. llama.cpp is a framework running llama models locally. llama / llama2 are models from Meta

https://ollama.com/ https://github.com/ollama/ollama

QAnything

https://github.com/netease-youdao/QAnything

PrivateGPT

see PrivateGPT

single file AI llamafile

llamafiles are single file bundles of an AI model and web based GUI as well as OpenAI compatible API working on windows/linux/mac at the same time.

https://simonwillison.net/2023/Nov/29/llamafile/ https://news.ycombinator.com/item?id=37786525
https://huggingface.co/jartine https://huggingface.co/jartine/llava-v1.5-7B-GGUF/tree/main
https://huggingface.co/jartine/WizardCoder-Python-34B-V1.0-llamafile
https://huggingface.co/WizardLM
https://huggingface.co/jartine/WizardCoder-Python-34B-V1.0-llamafile/blob/main/wizardcoder-python-34b-v1.0.Q2_K.llamafile

Coding AI

https://huggingface.co/blog/codegemma

Hardware

Nvidia Tesla H100 'Hopper': flagship card with 80GB RAM, cost about GBP30,000 Nvidia B100 and B200 announced to succeed the H100 Nvidia Tesla A100 'Tensor core' 80GB cost about GBP 16,000-22,000 Nvidia Tesla A100 'Tensor core' 40GB cost about GBP 8,000-10,000 (about 2.5x more powerful than V100) Nvidia Tesla V100 'Tensor core' 16GB cost about GBP 2,500 AMD MI300X : cost around USD10,000-15,000 Intel Gaudi 3 announced for Q3/24 claiming 1.5% performance of Nvidia H100

Ollama reqs: Any modern CPU with at least 4 cores is recommended. For running 13B models, a CPU with at least 8 cores is recommended. You should have at least 8 GB of RAM available to run the 7B models, 16 GB to run the 13B models, and 32 GB to run the 33B models.

Real world stats: CPU only, mistral-7b-openorca.Q5_K_M model, 32GB RAM, Intel Core i5-1135G7 @ 2.40GHz (NUC11) Prompt eval 7.5-8 tokens/second, eval ~2 tokens/second. 58 seconds model load time.

Examples of computing power required to generate/train a transformer model: Reka used 1000s of Nvidia H100 GPUs to train the closed source “Reka Code” in several months to rival GPT-4 and Claude 3 Opus.

GPT-3 175B model: Microsoft built a supercomputer with 285,000 CPU codes and 10,000 Nvidia V100 GPUs exclusively for OpenAI, hosted in Azure. Researchers calculated that to train GPT-3 OpenAI could have needed 34 days using 1024 A100 GPUs costing $5M just in compute time based on 175 billion parameters if the A100 had been available at the time.

Llama 3.1 used 16,000 Nvidia H100 GPUs to train the 405B model.

Evaluation

https://www.philschmid.de/llm-evaluation