AI Open Source, inevitabile modello di sviluppo?

Francesco Passantino

fpassantino@gmail.com

Linux Day
Teatro Gregotti, Palermo

28 Ottobre 2023

Francesco Passantino

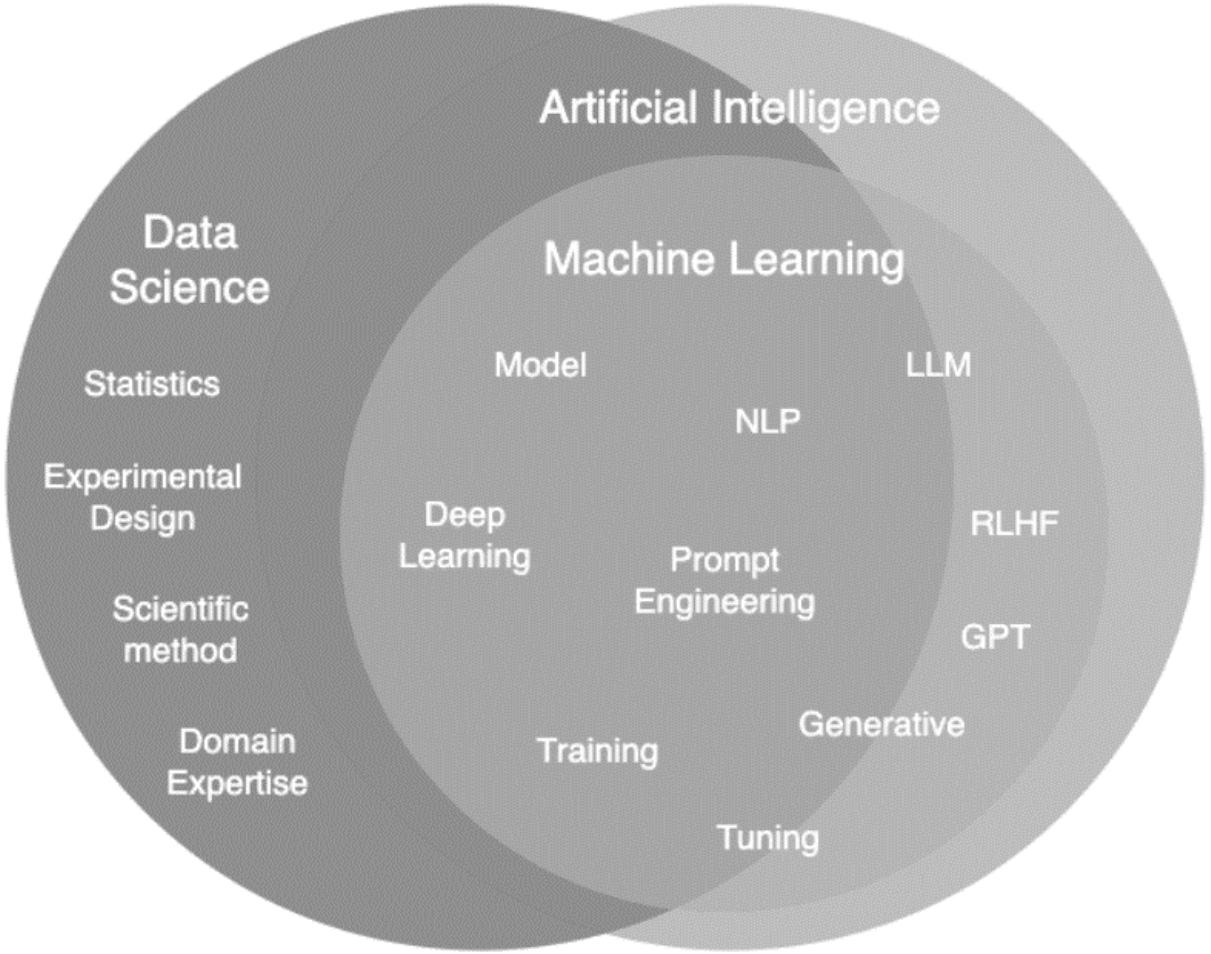
- Consulente nel settore delle Tecnologie dell'Informazione e della Comunicazione per diverse organizzazioni, sia pubbliche che private
- Da oltre trent'anni svolgo attività di formazione ICT, dai bambini ai post-laureati
- Ho fondato e fatto parte di diverse startup ed ho lavorato come valutatore di imprese per banche ed altri enti
- Attualmente, mi dedico all'attività di Digital Strategist, specializzato in Data Science e AI e di Project Manager per servizi mobile e web
- Come volontario, senza fini di lucro, a Palermo ho fondato il Google Developer Group, CoderDojo, co-fondato Sementor ed organizzato sei edizioni di Startup Weekend.

https://it.linkedin.com/in/fpassantino

INTRODUZIONE

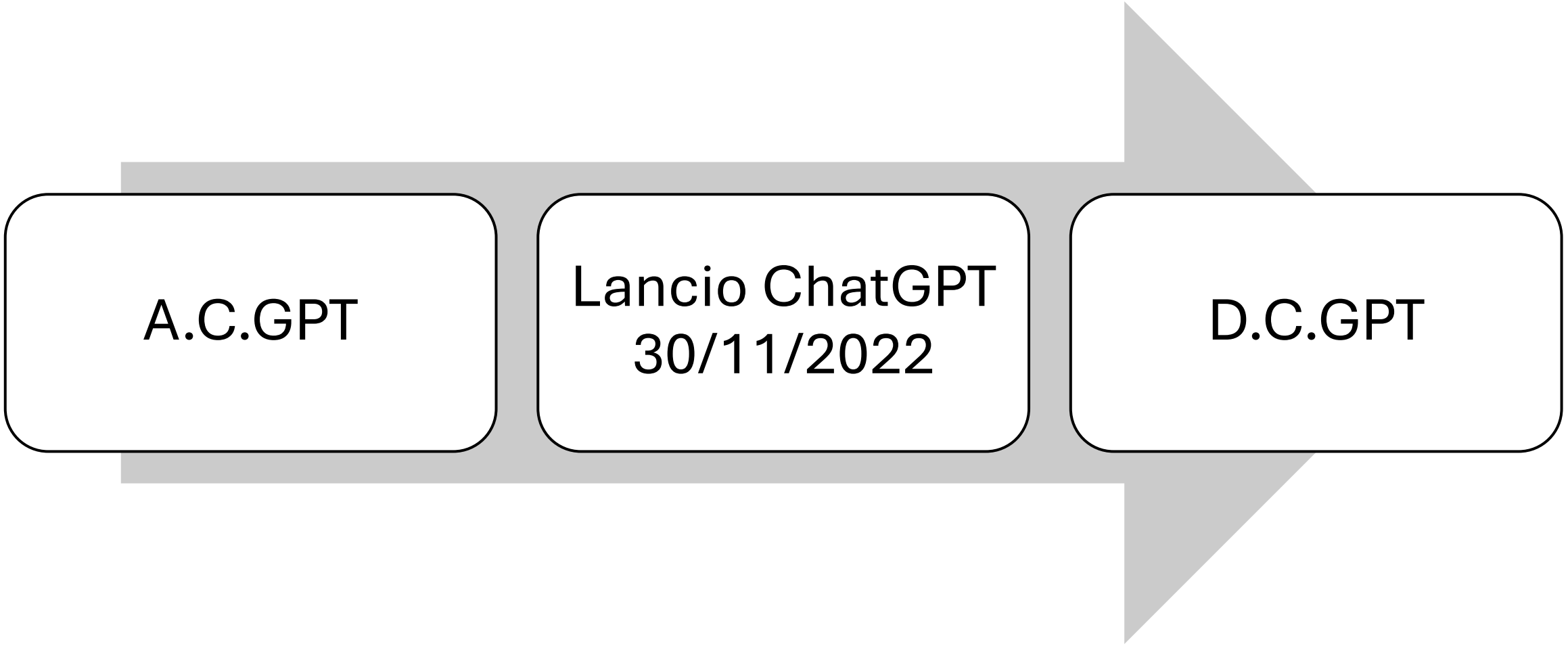
AI Open Source, inevitabile modello di sviluppo?

A Brief History of AI



https://www.phdata.io/blog/taking-the-first-steps-toward-enterprise-ai/

Il punto di svolta



https://openai.com/blog/chatgpt

4/5/23: Google "We Have No Moat, And Neither Does OpenAI"

Documento interno di Google:

- Con LLaMA di Meta, la comunità Open Source ha avuto accesso a un potente strumento, scatenando un'ondata di innovazione con sviluppi significativi
- I modelli Open Source migliorano rapidamente, gli sviluppatori possono iterare sui miglioramenti altrui e sperimentare molto di più
- La dimensione dei dati è meno importante del previsto grazie alla possibilità di ottenere alte prestazioni con «piccoli dataset altamente curati»
- L'articolo propone che Google renda Open Source la sua tecnologia AI.

https://www.semianalysis.com/p/google-we-have-no-moat-and-neither

12/7/23: Regulating Frontier AI: To Open Source or Not?

Negli ultimi mesi, è cresciuto l'interesse su come regolamentare l'Intelligenza Artificiale, in particolare riguardo ai modelli "Frontier AI". Due recenti articoli hanno approfondito la questione.

- Il primo articolo, "Frontier AI Regulation: Managing Risks to Public Safety", è una collaborazione tra vari esperti e organizzazioni, tra cui OpenAI, Google Deep Mind, Microsoft e altri. Propone approcci di governance per i modelli «Frontier AI», che presentano rischi unici, concentrando l'attenzione sulla autoregolamentazione con supervisione governativa.
- Il secondo articolo, "AI Safety and the Age of Dislightenment", scritto da Jeremy Howard, sottolinea i rischi dell'autoregolamentazione. L'autore propone di prendere una strada Open Source che porti a un'innovazione sicura e benefica. Al centro della sua argomentazione vi sono due preoccupazioni:
  - La prima è che, regolamentando i modelli di AI alla fonte e dando ai governi la capacità di concedere "licenze di sviluppo" alle aziende, emergerà una classe di aziende di AI con accesso completo a una delle tecnologie più potenti.
  - La seconda preoccupazione è che, regolamentando i modelli di AI piuttosto che i loro usi, si sposta l'attenzione dal gestire i rischi tangibili alla gestione degli esiti ipotetici di una tecnologia generale.

https://andrewmaynard.substack.com/p/regulating-frontier-ai-models

5/10/23: How AI Can Fight Inequality

- Nell’intervista si discute l'impatto potenziale dell’AI generativa Open Source sulle economie globali, soprattutto nei paesi poveri.
- Evidenzia come l’AI generativa possa migliorare la produttività e la prosperità, riducendo i compiti base e promuovendo l'innovazione.
- L'accesso all’AI potrebbe essere democratizzato attraverso modelli Open Source, rendendo l’AI culturalmente rilevante e accessibile.
- Nonostante le sfide, come i costi elevati di addestramento e la necessità di dati pertinenti, l'approccio Open Source potrebbe replicare il successo del software Open Source nel rendere l’AI ubiqua, in modo simile all'espansione di Internet.

Azeem Azhar, Exponential View
Emad Mostaque, founder and CEO of Stability AI

https://www.youtube.com/watch?v=enp8lM3DwGc

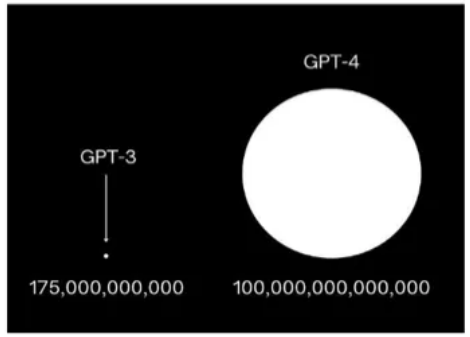
12/10/23: La crescita di Hugging Face

- Hugging Face, da 7 anni punto di riferimento per l'AI Open Source, sta vedendo un notevole slancio mentre la comunità si sforza di mantenere modelli e dataset di AI accessibili a tutti.
- Oltre 1.300 modelli sono stati inviati al loro Open LLM Leaderboard in pochi mesi con più di 600 milioni di download di modelli solo nell'agosto 2023.
- Questi modelli sono esposti su Spaces come applicazioni web costruite con strumenti come Gradio o Streamlit, permettendo una maggiore accessibilità e una rapida prototipazione.
- Gli utenti attivi mensili di Gradio sono cresciuti 5 volte, passando da 120k (gennaio '23) a 580k (agosto '23).

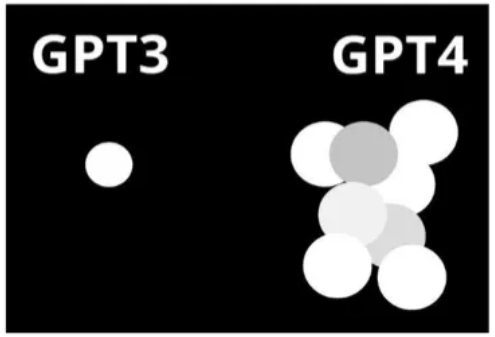
https://huggingface.co/ su https://www.stateof.ai/

23/10/23: AI and Open Source in 2023

- L'articolo mette in luce come la disponibilità di LLM Open Source sia cresciuta grazie alla release di modelli come Llama.
- Sebbene la maggior parte degli LLM Open Source siano ancora modelli di testo puro, vi sono sforzi per renderli multimodali.
- L’articolo menziona anche l'obiettivo della ricerca di eguagliare le prestazioni di GPT-4 con modelli più piccoli, suggerendo che i futuri progressi potrebbero derivare da approcci diversi dalla semplice scala.



meme



truth?

https://magazine.sebastianraschka.com/p/ai-and-open-source-in-2023

26/10/23: Open vs Closed

- Gli LLM Open Source livellano il campo di gioco per la ricerca e le imprese, ma comportano un rischio maggiore di proliferazione e abuso da parte di attori malevoli. Le API a codice chiuso offrono maggiore sicurezza e controllo, ma meno trasparenza.
- Open AI ha creato un team per supportare la sicurezza dei sistemi AI e per valutare, prevedere e proteggere contro i rischi catastrofici.
- L'approccio alla sicurezza Open Source varia tra le aziende senza linee guida standard.
  - Il lancio di Llama2 da parte di Meta è stato accompagnato da una panoramica dettagliata delle misure di sicurezza e da una Guida all'Uso Responsabile per fornire le migliori pratiche per gli sviluppatori.
  - Per scaricare i pesi di Llama2, gli utenti devono firmare un accordo in cui dichiarano di non volerlo utilizzare per scopi malevoli, ma non è chiaro chi lo farà rispettare.
  - I modelli distribuiti tramite Hugging Face hanno licenze che ne limitano l'uso e offrono modelli di moderazione.
  - Al contrario, il lancio del modello Persimmon 8B di Adept ha completamente ignorato la sicurezza

https://openai.com/blog/frontier-risk-and-preparedness

SOTA (State Of The Art)

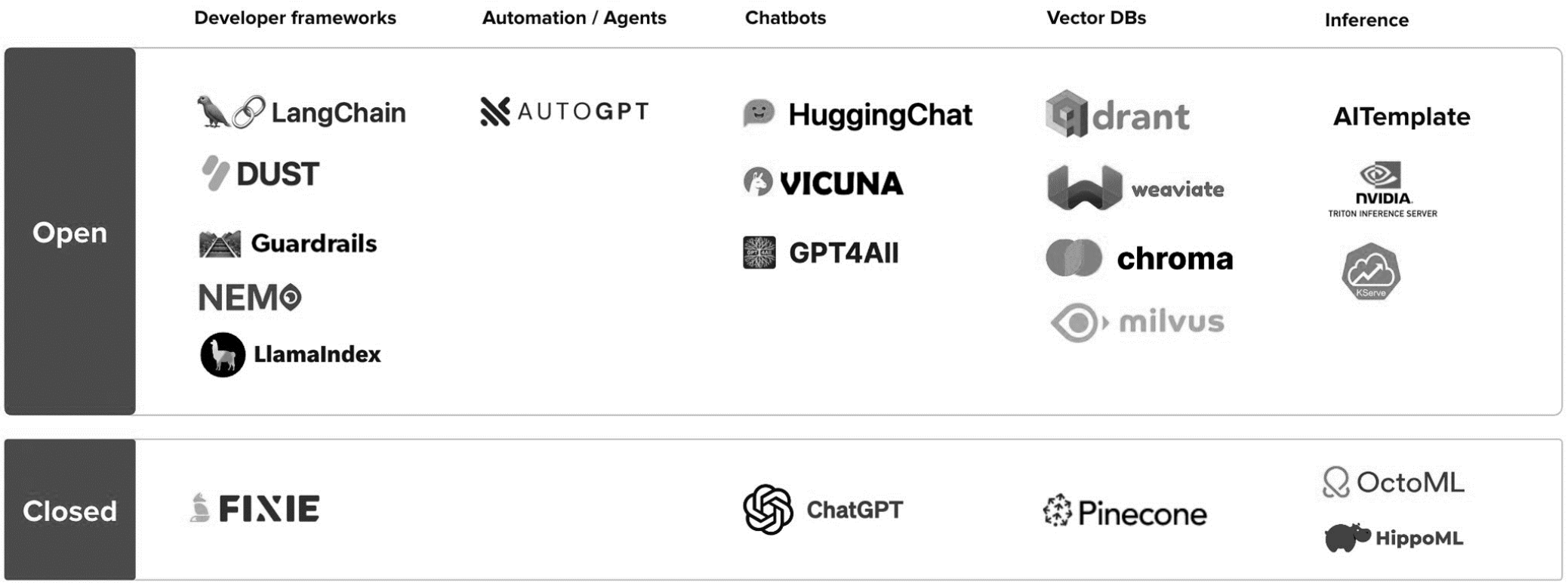
AI Open Source, inevitabile modello di sviluppo?

19/10/23: State of Open Source AI

- Licences: Weights vs Data, Commercial use, Fair use, Pending lawsuits
- Evaluation & Datasets: Leaderboards & Benchmarks for Text/Visual/Audio models
- Models: LLaMA 1 vs 2, Stable Diffusion, DALL-E, Persimmon, ...
- Unaligned Models: FraudGPT, WormGPT, PoisonGPT, WizardLM, Falcon
- Fine-tuning: LLMs, Visual, & Audio models
- Model Formats: ONNX, GGML, TensorRT
- MLOps Engines: vLLM, TGI, Triton, BentoML, ...
- Vector Databases: Weaviate, Qdrant, Milvus, Redis, Chroma, ...
- Software Development toolKits: LangChain, LLaMA Index, LiteLLM
- Desktop Apps: LMStudio, GPT4All, Koboldcpp, ...
- Hardware: NVIDIA CUDA, AMD ROCm, Apple Silicon, Intel, TPUs, ...

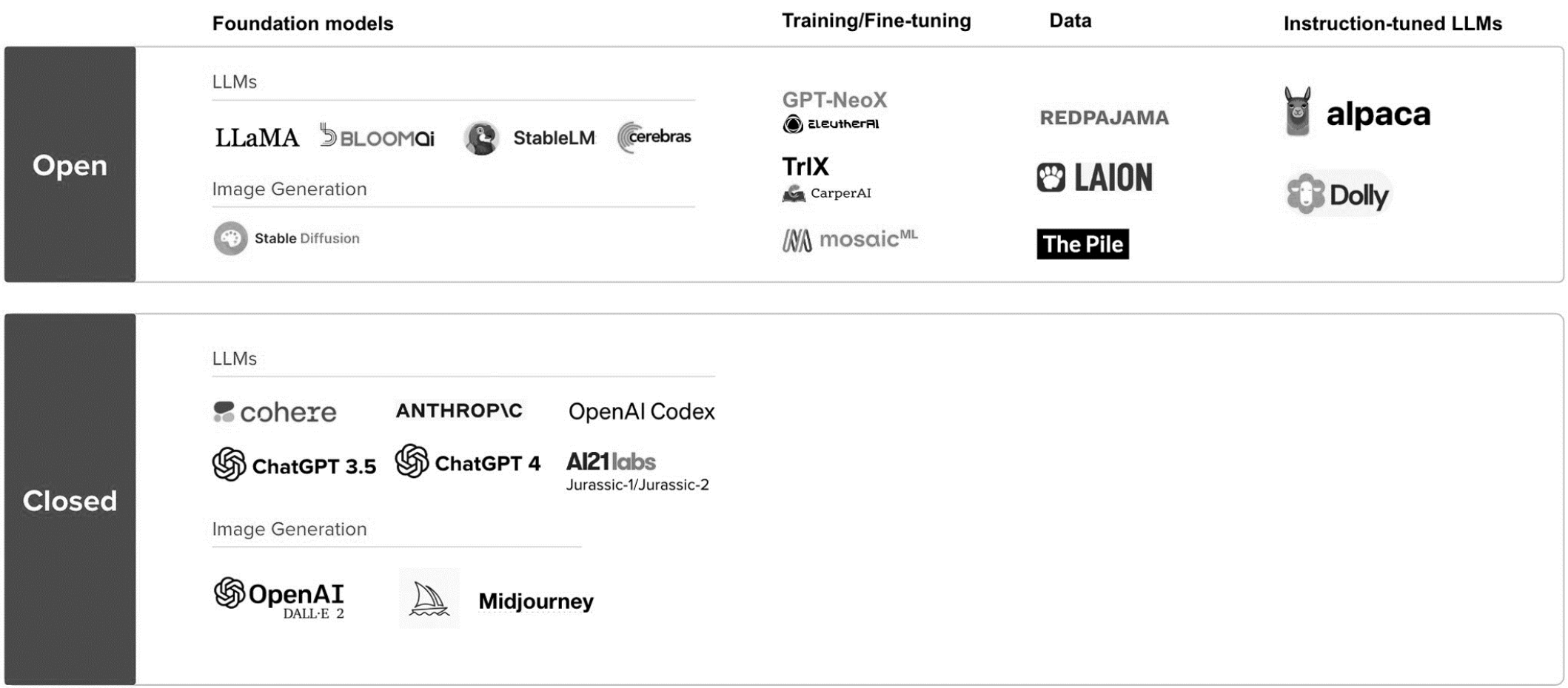
https://book.premai.io/state-of-open-source-ai/

The AI Battle 1/2



https://twitter.com/chiefaioffice/status/1688913452662984708?s=20

The AI Battle 2/2



https://twitter.com/chiefaioffice/status/1688913452662984708?s=20

Restrizioni sui dati di addestramento, sui pesi addestrati e sugli output generati

Model	Weights	Training Data	Output
OpenAI ChatGPT	☒ unavailable	☒ unavailable	☑ user has full ownership
Anthropic Claude	☒ unavailable	☒ unavailable	☐ commercial use permitted
LMSys Vicuna 33B	☑ open source	☒ unavailable	☒ no commercial use
LMSys Vicuna 13B	☑ open source	☒ unavailable	☐ commercial use permitted
MosaicML MPT 30B Chat	☑ open source	☒ unavailable	☒ no commercial use
Meta LLaMA2 13B Chat	☑ open source	☒ unavailable	☐ commercial use permitted
RWKV4 Raven 14B	☑ open source	☑ available	☑ user has full ownership
OpenAssistant SFT4 Pythia 12B	☑ open source	☑ available	☑ user has full ownership
MosaicML MPT 30B Instruct	☑ open source	☒ unavailable	☐ commercial use permitted
MosaicML MPT 30B	☑ open source	☒ unavailable	☑ user has full ownership

https://book.premai.io/state-of-open-source-ai/

Confronto tra Database di Vettori

Vector Database	Open Source	Shard-ing	Supported Distance Metrics	Supported Indices
https://github.com/weaviate/weaviate	☑ Yes	☑ Yes	cosine, dot, L2 squared, ham-ming, manhattan	HNSW, HNSW-PQ
https://github.com/qdrant/qdrant	☑ Yes	☑ Yes	cosine, dot, euclidean	HNSW
https://github.com/milvus-io/milvus	☑ Yes	☑ Yes	cosine, dot, euclidean, jaccard, hamming	HNSW, FLAT, IVF-FLAT, IVF-PQ
https://github.com/RedisVentures/redisvl	☑ Yes	☑ Yes	cosine, inner product, L2	HNSW, FLAT
https://github.com/chroma-core/chroma	☑ Yes	☒ No	cosine, inner product, L2	HNSW
Pinecone	☒ No	☑ Yes	cosine, dot, euclidean	HNSW, FLAT, LSH, PQ

https://book.premai.io/state-of-open-source-ai/

Confronto tra LLM SDK

SDK	Use cases	Vector stores	Embedding model	LLM Model	Languages	Features
LangChain	Chatbots, prompt chaining, document related tasks	Comprehensive list of data sources available to get connected readily	State of art embedding models in the bucket to choose from	A-Z availability of LLMs out there in the market	Python, Javascript, Typescript	Open Source & 1.5k+ contributors strong for active project development
LLaMA Index	Connecting multiple data sources to LLMs, document query interface using retrieval augmented generation, advanced chatbots, structured analytics	Wide options to connect & facility to create a new one	Besides the 3 commonly available models we can use a custom embedding model as well	Set of restricted availability of LLM models besides customised abstractions suited for your custom data	Python, Javascript, Typescript	Tailor-made for high customisations if not happy with the current parameters and integrations
LiteLLM	Integrating multiple LLMs, evaluating LLMs	Not Applicable	Currently supports only text-embedding-ada-002 from OpenAI & Azure	Expanding the list of LLM providers with the most commonly used ones ready for use	Python	Lightweight, streaming model response, consistent output response

https://book.premai.io/state-of-open-source-ai/

Altri progetti

Apache MXNet, Caffe, DeepChem, Detectron2, Fairseq, Fastai, GPT Engineer, Hugging Face Transformers, Keras, Magenta, MindsDB, Open Assistant, OpenAI GPT-2, OpenCV, PyTorch, scikit-learn, Stable Diffusion, Tensor2Tensor, TensorFlow, tflearn, Theano, ...

CONCLUSIONI

AI Open Source, inevitabile modello di sviluppo?

Democratizzare l’AI?

- L'abbondanza di progetti Open Source nell'ecosistema AI, mostra che la barriera all'entrata nel campo dell’AI sta diminuendo, in linea con le affermazioni dell'articolo "Google We Have No Moat, And Neither Does OpenAI"
- La varietà di licenze Open Source, e la vasta gamma di funzionalità offerte da questi progetti, facilitano l'accesso alle risorse di AI e promuovono l'innovazione, riducendo la dipendenza da grandi entità commerciali
- Questo scenario potrebbe alla lunga erodere il vantaggio competitivo di giganti tecnologici come Google e OpenAI, rendendo l'ecosistema AI più democratico e accessibile.

Il mito dell'AI Open Source?

Anche se l'AI Open Source sta avanzando, è evidente che rimane fortemente regolamentata dalle grandi aziende. Queste entità controllano spesso parametri critici, inclusi:

- I chip necessari sono/saranno fabbricati da Nvidia, Intel, AMD, Apple, Meta, Google, Microsoft, OpenAI, Amazon, TMSC ed il nodo di Taiwan, ...
- La potenza di calcolo necessaria per addestrare questi modelli
- I sistemi più avanzati sono costruiti in occidente
- Il controllo dei framework software necessari per costruire tali modelli
- I dati necessari per addestrare questi modelli possono subire restrizioni
- L'addestramento su dati in lingua inglese
  (altre lingue: Falcon, Mistral/Bloom, Aleph Alpha/Jina, Lince Zero, ...)
- Lo «human feedback» dovrebbe essere come Wikipedia.

Link per approfondire

- Democratize machine learning:
  https://huggingface.co/
- State of AI:
  https://www.stateof.ai/
- State of Open Source AI:
  https://book.premai.io/state-of-open-source-ai/
- Impact of chatGPT:
  https://www.youtube.com/playlist?list=PLKemzYMx2_Ot1MZ_er2vFiINdJEgDO8Hg
- Time100/AI:
  https://time.com/collection/time100-ai/
- Le Voci dell'AI:
  https://www.01net.it/voci-ai/
- Etica & AI:
  https://www.youtube.com/playlist?list=PL24C9VeBHCPWMp6av3nkX8iYWz4my3iR5

Working Progress

bitRocket

https://www.linkedin.com/company/bitrocket-ai/

Via Vittorio Emanuele, 188 - Palermo

Progetto AI per professionisti, startup e imprese