FBS-2843: Compare LLMs known performance and parameters.

Insights

- For our use case **MIT and Apache 2.0 license** is relevant.
- **Open Flamingo** has 9B parameter size and dataset is large multimodal dataset c4 – document, images interleaved.
- **GPT-NeoX and GPT-J 20B** and 6B has parameter size and dataset is Pile An 800GB Dataset of Diverse Text for Language Modeling derive from academic or professional sources.
- **OpenLLaMa** has 7B parameter size and dataset is RedPajama dataset contains Common crawl, C4, GitHub, Book, ArXiv, Wikipedia, Stack Exchange
- **StableLM** has 3B parameters size and dataset is higher quality data sources and mixtures; specifically, the use of RefinedWeb (is built through stringent filtering and large-scale deduplication of CommonCrawl) and C4 (is the processed version of Google's C4 dataset).
- **Dolly** has 12B parameter size and dataset is capability domains from the InstructGPT paper, including brainstorming, classification, closed QA, generation, information extraction, open QA and summarization. Dataset is licensed for commercial use.
- **Falcon** has 40B parameter size and dataset was collected from public crawls of the web to build the pretraining dataset of Falcon. Using dumps from CommonCrawl, after significant filtering then extended with a few curated sources such as research papers and conversations from social media.

Criteria

- **Size of model**
  - Larger models requiring more powerful hardware and incurring higher costs.
  - Smaller models are more efficient to train, deploying them incurring lower costs.
  - Choosing the smallest possible model that meets our use case, we can optimize our costs and resources.
- **Performance – I: Use case specific**
  - Trained on Technical dataset for better recommendation.
  - RedPajama, Common crawl, C4, GitHub, Book, ArXiv, Wikipedia, Stack Exchange
- **Performance – II: Behavior as Chatbot**
  - How to Evaluate the Quality of LLM-based Chatbots | by Matt Ambrogi | Towards AI
  - **Chatbot evaluation at a high level**
    - What is a good response?
      - **Strategy – forms an intuitive opinion** done by asking a lot of questions and getting a feel for whether your bot's responses tend to be good or not.
    - How is qualitative evaluation done?
      - **Strategy – thumbs integration:** thumbs up / thumbs down next to a response that allows users to report whether the answer was helpful or not.

Conclusion

- **These models are interesting to look into for our use case.**
  - **GPT-NeoX 20B (MIT)**
  - **GPT-J 6B (Apache 2.0)**
  - **OpenLLaMa 7B (Apache 2.0)**
  - **Falcon 40B (Apache 2.0)**

Below is table showing detailed description of selected model with size, license, dataset used to train them, and GitHub link.

Model	Size	License	Dataset	Description	GitHub Link
Open Flamingo	9B	MIT	Is trained on a large multimodal dataset (e.g. Multimodal C4) arxiv.org/pdf/2304.06939.pdf **is an augmentation of the popular text-only c4 corpus with images interleaved. The corpus contains 103M documents containing 585M images interleaved with 43B English tokens.**	Open source implementation of Deepmind's Flamingo model	GitHub - mlfoundations/open_flamingo: An open-source framework for training large multimodal models.
Alpaca 7B	7B	Apache 2.0	is fine-tuned from a 7B LLaMA model on 52K instruction-following data generated by the techniques in the Self-Instruct GitHub - tatsu-lab/stanford_alpaca: Code and documentation to train Stanford's Alpaca models, and generate the data. **contains 52K instruction-following data we used for fine-tuning the Alpaca model.**	Stanford's Instruction-following LLaMA Model	GitHub - tatsu-lab/stanford_alpaca: Code and documentation to train Stanford's Alpaca models, and generate the data. Stanford CRFM
Falcon	40B	Apache 2.0	data was collected from public crawls of the web to build the pretraining dataset of Falcon. Using dumps from CommonCrawl, after significant filtering [2306.01116] The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only (arxiv.org) **dataset was then extended with a few curated sources such as research papers and conversations from social media.**	is a foundational LLM with 40 billion parameters trained on one trillion tokens.	Falcon LLM - Home (tii.ae)
OpenLLaMa	7B	Apache 2.0	trained on the RedPajama dataset. togethercomputer/RedPajama-Data-1T · Datasets at Hugging Face **DatasetToken Count** **Commoncrawl 878 Billion** **C4 175 Billion** **GitHub 59 Billion** **Books 26 Billion**	Another Open source reproduction of Meta AI's LLaMA 7B trained on the RedPajama dataset.	GitHub - openlm-research/open_llama: OpenLLaMA, a permissively licensed open source reproduction of

			ArXiv 28 billion **Wikipedia 24 Billion** **StackExchange 20 Billion** **Total 1.2 Trillion**		Meta AI's LLaMA 7B trained on the RedPajama dataset
ColossalChat	7B	Apache 2.0	104K bilingual datasets of Chinese and English, and you can find the datasets in this repo InstructionWild and in this file. GitHub - XueFuzhao/InstructionWild **Instruction Tuning is a key component of ChatGPT. OpenAI used their user-based Instruction dataset, but unfortunately, this dataset is not open-sourced.** **project targets on a larger and more diverse instruction dataset. collected (110K in v2 dataset, 429 in v1 dataset) instructions from ChatGPT usage sharing and released both English and Chinese versions.**	An open-source solution for cloning ChatGPT with a complete RLHF pipeline.	ColossalAI/applications/Chat at main · hpcaitech/ColossalAI · GitHub
FastChat-T5	3B	Apache 2.0	125K user-shared conversations gathered from ShareGPT.com with public APIs. To ensure data quality, we convert the HTML back to markdown and filter out some inappropriate or low-quality samples. Additionally, we divide lengthy conversations into smaller segments that fit the model's maximum context length. FastChat/docs/commands/data_cleaning.md at main · lm-sys/FastChat · GitHub **not release the ShareGPT dataset.**	Chatbot trained by fine-tuning Flan-t5-xl on user-shared conversations collected from ShareGPT	GitHub - lm-sys/FastChat: An open platform for training, serving, and evaluating large language models. Release repo for Vicuna and Chatbot Arena.
GPT-J	6B	Apache 2.0	The Pile: An 800GB Dataset of Diverse Text for Language Modeling [2101.00027] The Pile: An 800GB Dataset of Diverse Text for Language Modeling (arxiv.org) **derive from academic or professional sources**	Eluether AI's open source version of GPT with fewer params	GitHub - kingoflolz/mesh-transformer-jax: Model parallel transformers in JAX and Haiku
GPT-NeoX	20B	MIT	The Pile: An 800GB Dataset of Diverse Text for Language Modeling	Eluether AI's open source version of GPT with fewer params	GitHub - EleutherAI/gpt-neox: An implementation of model

			[2101.00027] The Pile: An 800GB Dataset of Diverse Text for Language Modeling (arxiv.org) **derive from academic or professional sources**		parallel autoregressive transformers on GPUs, based on the DeepSpeed library.
Flan-UL2	20B	Apache 2.0	datasets are phrased as instructions which enable generalization across diverse tasks. In Flan2, we released a series of T5 models ranging from 200M to 11B parameters that have been instruction tuned with Flan. A New Open Source Flan 20B with UL2 — Yi Tay [2301.13688] The Flan Collection: Designing Data and Methods for Effective Instruction Tuning (arxiv.org) **has been primarily trained on academic tasks**	Flan has been primarily trained on academic tasks. model that was trained on top of the already open sourced UL2 20B checkpoint.	google-research/ul2 at master · google-research/google-research · GitHub
StableLM	3B	Apache 2.0	usage of higher quality data sources and mixtures; specifically, the use of RefinedWeb and C4 in place of The Pile v2 Common-Crawl scrape as well as sampling web text at a much higher rate (35% -> 71%). tiiuae/falcon-refinedweb · Datasets at Hugging Face **RefinedWeb is built through stringent filtering and large-scale deduplication of Common Crawl; we found models trained on RefinedWeb to achieve performance in-line or better than models trained on curated datasets, while only relying on web data.** allenai/c4 · Datasets at Hugging Face **is the processed version of Google's C4 dataset.**	Stability AI's LLM model series	GitHub - Stability-AI/StableLM: StableLM: Stability AI Language Models
YaLM	100B	Apache 2.0	1.7 TB of online texts, books, and countless other sources in both English and Russian. Yandex Publishes YaLM 100B. It's the Largest GPT-Like Neural Network in Open Source | by Mikhail Khrushchev | Yandex | Medium	Pretrained LLM from Yandex for generating and processing text	GitHub - yandex/YaLM-100B: Pretrained language model with 100B parameters

			25% The Pile — open English dataset by Eleuther AI team 75% Texts in Russian collected by our team (percentages of the whole dataset are given) 12% News from various sources from Yandex Search index 10% Books from the dataset used in Russian Distributional Thesarus 3% Misc texts from the Taiga Dataset 1.5% Dialogues from social media preprocessed in a manner similar to how Reddit is proccessed in The Pile 0.5% Russian portion of Wikipedia **Academic and Russian text.**		
Dolly	12B	Apache 2.0	trained on the Databricks machine learning platform that is licensed for commercial use. databricks/databricks-dolly-15k · Datasets at Hugging Face is trained on ~15k instruction/response fine tuning records databricks-dolly-15k generated by Databricks employees in capability **domains from the InstructGPT paper, including brainstorming, classification, closed QA, generation, information extraction, open QA and summarization.**	Pythia 12B LLM trained on Databricks ML platform	GitHub - databrickslabs/dolly: Databricks' Dolly, a large language model trained on the Databricks Machine Learning Platform