How LLMs are publicly released?

BLOOM-176B, LINE-3.6B, MPT-30B, Llama 2, BLOOM-3b-zh, and Taiwan-LLaMa2

2023-08-31

summarized by trc

BLOOM-176B (2022-07)

- 1.5TB text in 46 + 13 languages; 350B tokens; trained on Megatron-LM GPT-2
- "whoever agrees to the terms of the model's Responsible AI License (RAIL) can use and build upon the model on a local machine or on a cloud provider"
- RAIL: “a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable copyright license to reproduce, prepare, publicly display, publicly perform, sublicense, and distribute the **Complementary Material**, the **Model**, and **Derivatives of the Model**”
- **on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND**
- Use Restrictions: “You agree not to use the Model or Derivatives of the Model: ...”
- Corpus: “Each constituent subset of the dataset will be released under the license that applies to it”; Tooling code: Apache 2.0
- (blog) https://bigscience.huggingface.co/blog/bloom
  (license) https://huggingface.co/spaces/bigscience/license
  (details) https://github.com/bigscience-workshop/bigscience/tree/master/train/tr11-176B-ml
  (paper) https://arxiv.org/abs/2211.05100
  (corpus) https://openreview.net/forum?id=UoEw6KigkUn

RAIL Use Restrictions

You agree not to use the Model or Derivatives of the Model:

- In any way that violates any applicable … law or regulation;
- For the purpose of exploiting, harming or attempting to exploit or harm minors in any way;
- To generate or disseminate verifiably false information with the purpose of harming others;
- To generate or disseminate personal identifiable information that can be used to harm an individual;
- To generate or disseminate information or content, in any context … without expressly and intelligibly disclaiming that the text is machine generated;
- To defame, disparage or otherwise harass others;
- To impersonate or attempt to impersonate others;
- For fully automated decision making that adversely impacts an individual's legal rights ...
- For any use intended to or which has the effect of discriminating against or harming individuals or groups ...
- To exploit any of the vulnerabilities of a specific group of persons based on their age, social, physical or mental characteristics ...
- For any use intended to or which has the effect of discriminating against individuals or groups based on legally protected characteristics or categories;
- To provide medical advice and medical results interpretation;
- To generate or disseminate information for the purpose to be used for administration of justice, law enforcement, immigration or asylum processes, such as predicting an individual will commit fraud/crime commitment ...

LINE-3.6B (2023-08)

- 650 GB text (“Japanese portions of publicly available corpus such as C4, CC-100, and Oscar”); trained on GPTNeoX
- Apache 2.0 license: “a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable copyright license to reproduce, prepare Derivative Works of, publicly display, publicly perform, sublicense, and distribute the Work and such Derivative Works in Source or Object form”
- Also “patent license to make, have made, use, offer to sell, sell, import, and otherwise transfer the Work”
- **on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND**
- (blog) https://engineering.linecorp.com/ja/blog/3.6-billion-parameter-japanese-language-model
  (license) https://www.apache.org/licenses/LICENSE-2.0
  (details) https://huggingface.co/line-corporation/japanese-large-lm-3.6b
  (datasets) https://huggingface.co/datasets/oscar
  https://huggingface.co/datasets/wikipedia
  https://huggingface.co/datasets/mc4

MPT-30B etc. (2023-06)

- Trained on 1T tokens
- MPT-30B: Apache 2.0 licensed
- MPT-30B-Instruct: a model for long-form instruction following; CC BY-SA 3.0 licensed
- MPT-30B-Chat: a chatbot-like model for dialogue generation; CC BY-NC-SA 3.0 licensed
- (blog) https://www.mosaicml.com/blog/mpt-30b
  (license)
  https://www.apache.org/licenses/LICENSE-2.0
  https://creativecommons.org/licenses/by-sa/3.0/
  https://creativecommons.org/licenses/by-nc-sa/3.0/
  (details)
  https://huggingface.co/mosaicml/mpt-30b
  https://huggingface.co/mosaicml/mpt-30b-instruct
  https://huggingface.co/mosaicml/mpt-30b-chat

mosaicML MPT-30B Training Data

Data Source	Number of Tokens in Source (Billion)	Proportion	Effective Number of Tokens (Billion)	Epochs
mC4 3.1.0 - English (200+ words)	2417.99	33.5%	335	0.14
c4 - English - SemDedup 80%	100.42	29.9%	299	2.98
RedPajama - CommonCrawl	878.45	8.5%	85	0.10
The Stack - Selected Languages	463.78	10.0%	100	0.22
RedPajama - Wikipedia	4.87	4.0%	40	8.21
The Stack - Markdown	107.07	4.5%	45	0.42
Semantic Scholar ORC	48.95	3.3%	33	0.67
RedPajama - Books	26.02	3.0%	30	1.15
RedPajama - arXiv	28.1	1.9%	19	0.68
RedPajama - StackExchange	20.54	1.4%	14	0.68

mosaicML Foundation Series

MPT-30B / MPT-30B-Instruct / MPT-30B-Chat

- Limitations and Biases
  - “MPT-30B (Base) is not intended for deployment without finetuning. It should not be used for human-facing interactions without further guardrails and user consent.”
  - MPT-30B / MPT-30B-Instruct / MPT-30B-Chat “can produce factually incorrect output, and should not be relied on to produce factually accurate information.”
- Disclaimer
  - “The license on this model does not constitute legal advice. We are not responsible for the actions of third parties who use this model. Please consult an attorney before using this model for commercial purposes.”

Llama 2 (2023-07)

- Trained on 2T tokens “of data from publicly available sources”; 40% more than LLaMA
- “You are granted a non-exclusive, worldwide, **non-transferable** and royalty-free limited license under Meta's intellectual property or other rights owned by Meta embodied in the **Llama Materials** to use, reproduce, distribute, copy, create derivative works of, and make modifications to the Llama Materials.”
- “Your use of the Llama Materials must comply with applicable laws and regulations … and adhere to the **Acceptable Use Policy** for the Llama Materials...”
- “You will not use the **Llama Materials or any output or results of the Llama Materials** to improve any other large language model (excluding Llama 2 or derivative works thereof)”
- **Additional Commercial Terms**
- **on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND**
- (blog) https://ai.meta.com/resources/models-and-libraries/llama/
  (license) https://github.com/facebookresearch/llama/blob/main/LICENSE
  (acceptable use policy) https://ai.meta.com/llama/use-policy
  (details) https://github.com/facebookresearch/llama/tree/main
  https://huggingface.co/meta-llama
  (paper) https://arxiv.org/abs/2302.13971
  https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/
  (datasets) **None public yet**

BLOOM-3b-zh (2023-03)

- “We trained the 3B parameter model on a total of 13 Billion tokens of mostly high quality Traditional Chinese text”
- License: MEDIATEK RESEARCH License and RAIL License v1.0
  - ‘... you agree that (1) any use of the MR Model is solely for **“non-commercial research purposes”**; and (2) all the use and distribution of this MR Model shall fully comply with the terms and conditions set forth in the BigScience RAIL License 1.0 ...’
- (blog) https://bigscience.huggingface.co/blog/bloom
  (license) https://huggingface.co/ckip-joint/bloom-3b-zh/blob/main/LICENSE_MR.md
  https://huggingface.co/spaces/bigscience/license
  (details) https://huggingface.co/ckip-joint/bloom-3b-zh
  (paper) https://arxiv.org/abs/2303.04715
  (corpus) **listed but not packaged**

	Category	Size (tokens)	Epochs	Sampling Proportion
Gigaword5-CNA	Written (news)	0.8B	2.8	19.4%
ASBC	Written (literature)	0.01B	4.6	0.4%
COCT-books	Written (literature)	0.3B	7.7	20.0%
CC-100-zht	General (web)	2.0B	1.7	28.9%
Wikipedia-zht	Written (knowledge)	0.4B	2.9	10.1%
Theses	Written (knowledge)	0.4B	2.9	10.1%
xP3-zht	Instructions	1.1B	1.2	11%
All		5.2B		100%

Table 2: Data composition of our Traditional Chinese data set. The epochs sum up to 11.5 billion tokens for the 1B model. While for the 3B model, we use the same sample proportion among the subsets but higher numbers of epochs which lead to 13 billion tokens in total.

Taiwan-LLaMa2 (2023-05)

- “Taiwan-LLaMa v1.0 pretrained on over 5 billion tokens and instruction-tuned on over 490k conversations both in traditional mandarin.”
- “Code is licensed under Apache 2.0 License.”
- “Models are licensed under the LLAMA 2 Community License.”
- “The data included in this project were generated using OpenAI's models and are subject to OpenAI's Terms of Use.”
- Dataset: “Sorry, we can't find the page you are looking for.”
- (license) https://github.com/facebookresearch/llama/blob/main/LICENSE

  https://www.apache.org/licenses/LICENSE-2.0https://www.apache.org/licenses/LICENSE-2.0
  (details) https://huggingface.co/spaces/yentinglin/Taiwan-LLaMa2
  https://github.com/MiuLab/Taiwan-LLaMa
  (paper) https://arxiv.org/abs/2305.13711
  (Dataset) **(404)** https://huggingface.co/datasets/yentinglin/traditional_mandarin_instructions
  (404) https://huggingface.co/datasets/yentinglin/zh_TW_c4

大型語言模型的公眾授權使用

- 大型語言模型的「釋出」？
  - 模型、程式、資料集的**公眾**使用、重製、改用與散佈？
- （著作的）公眾授權：以條款文字言明使用條件，若同意不需簽約即可自行使用
  - （通常）永久、全球、非專屬、不會撤回、不收費、無授權金、不限使用目的
  - （通常）授權條款需伴隨著作一起散佈
  - （通常）不包括專利與商標的授權
  - （有時會再要求）相同方式分享 | 姓名標示 | 非商業性
  - 依「現狀」提供，不保證任何可用性（明示或暗示）
- 來自上游（模型、程式、資料集）的釋出方式限制了中游（開發者）可以對下游（使用者）的釋出方式