大语言模型发展中的重要问题讨论

CS2916 大语言模型

饮水思源 爱国荣校

https://plms.ai/teaching/index.html

大语言模型构建过程中的“透明性”

杰夫·贝索斯

目前形式的大语言模型并不是**发明**，而是**发现**。望远镜是一项发明，但通过它观察木星，知道它有卫星，是一项发现。而大语言模型更像是发现，它们的能力不断让我们感到惊讶

人生中让我印象深刻的两次**技术革命**演示，一次是现在操作系统的先驱“图形用户界面”，另一个就是以ChatGPT为代表的**生成式人工智能**技术

比尔盖茨

黄仁勋

ChatGPT相当于**AI界的iPhone**问世，它使**每一个人**都可以成为程序员

马斯克悄悄成立大模型公司xAI

大语言模型构建过程中的“透明性”

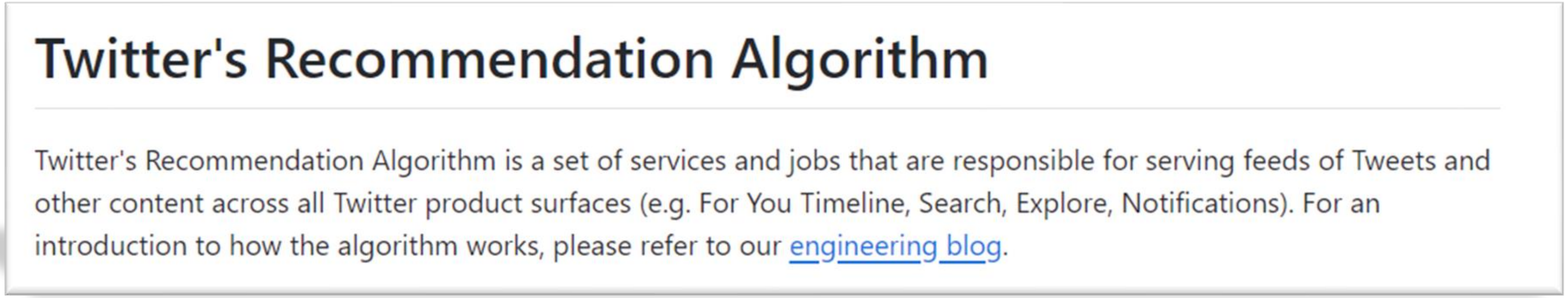
马斯克

This week, @xAI will open source Grok

大语言模型构建过程中的“透明性”

马斯克

This week, @xAI will open source Grok



Twitter's Recommendation Algorithm

Twitter's Recommendation Algorithm is a set of services and jobs that are responsible for serving feeds of Tweets and other content across all Twitter product surfaces (e.g. For You Timeline, Search, Explore, Notifications). For an introduction to how the algorithm works, please refer to our engineering blog.

大语言模型领域的“怪圈文化”

2.2 Training Dataset

Datasets for language models have rapidly expanded, culminating in the Common Crawl dataset[2] [RSR+19] constituting nearly a trillion words. This size of dataset is sufficient to train our largest models without ever updating on the same sequence twice. However, we have found that unfiltered or lightly filtered versions of Common Crawl tend to have lower quality than more curated datasets. Therefore, we took 3 steps to improve the average quality of our datasets (1) we downloaded and filtered a version of CommonCrawl based on similarity to a range of high-quality reference corpora, (2) we performed fuzzy deduplication at the document level, within and across datasets, to prevent redundancy and preserve the integrity of our held-out validation set as an accurate measure of overfitting, and (3) we also added known high-quality reference corpora to the training mix to augment CommonCrawl and increase its diversity.

Details of the first two points (processing of Common Crawl) are described in Appendix A. For the third, we added several curated high-quality datasets, including an expanded version of the WebText dataset [RWC+19], collected by scraping links over a longer period of time, and first described in [KMH+20], two internet-based books corpora (Books1 and Books2) and English-language Wikipedia.

Table 2.2 shows the final mixture of datasets that we used in training. The CommonCrawl data was downloaded from 41 shards of monthly CommonCrawl covering 2016 to 2019, constituting 45TB of compressed plaintext before filtering and 570GB after filtering, roughly equivalent to 400 billion byte-pair-encoded tokens. Note that during training, datasets are not sampled in proportion to their size, but rather datasets we view as higher-quality are sampled more frequently such that CommonCrawl and Books2 datasets are sampled less than once during training, but the other datasets are sampled 2-3 times. This essentially accepts a small amount of overfitting in exchange for higher quality training data.

OpenAI GPT3

细致地描述使用的预训练语料，包括组成、大小、过滤方法

大语言模型领域的“怪圈文化”

2.2 Training Dataset

Datasets for language models have rapidly expanded, culminating in the Common Crawl dataset[2] [RSR+19] constituting nearly a trillion words. This size of dataset is sufficient to train our largest models without ever updating on the same sequence twice. However, we have found that unfiltered or lightly filtered versions of Common Crawl tend to have lower quality than more curated datasets. Therefore, we took 3 steps to improve the average quality of our datasets (1) we downloaded and filtered a version of CommonCrawl based on similarity to a range of high-quality reference corpora, (2) we performed fuzzy deduplication at the document level, within and across datasets, to prevent redundancy and preserve the integrity of our held-out validation set as an accurate measure of overfitting, and (3) we also added known high-quality reference corpora to the training mix to augment CommonCrawl and increase its diversity.

Details of the first two points (processing of Common Crawl) are described in Appendix A. For the third, we added several curated high-quality datasets, including an expanded version of the WebText dataset [RWC+19], collected by scraping links over a longer period of time, and first described in [KMH+20], two internet-based books corpora (Books1 and Books2) and English-language Wikipedia.

Table 2.2 shows the final mixture of datasets that we used in training. The CommonCrawl data was downloaded from 41 shards of monthly CommonCrawl covering 2016 to 2019, constituting 45TB of compressed plaintext before filtering and 570GB after filtering, roughly equivalent to 400 billion byte-pair-encoded tokens. Note that during training, datasets are not sampled in proportion to their size, but rather datasets we view as higher-quality are sampled more frequently such that CommonCrawl and Books2 datasets are sampled less than once during training, but the other datasets are sampled 2-3 times. This essentially accepts a small amount of overfitting in exchange for higher quality training data.

OpenAI GPT3

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

We are committed to independent auditing of our technologies, and shared some initial steps and ideas in this area in the system card accompanying this release.[2] We plan to make further technical details available to additional third parties who can advise us on how to weigh the competitive and safety considerations above against the scientific value of further transparency.

OpenAI GPT4

一笔概括：使用了公开的互联网数据

大语言模型领域的“怪圈文化”

2.2 Training Dataset

Datasets for language models have rapidly expanded, culminating in the Common Crawl dataset[2] [RSR+19] constituting nearly a trillion words. This size of dataset is sufficient to train our largest models without ever updating on the same sequence twice. However, we have found that unfiltered or lightly filtered versions of Common Crawl tend to have lower quality than more curated datasets. Therefore, we took 3 steps to improve the average quality of our datasets (1) we downloaded and filtered a version of CommonCrawl based on similarity to a range of high-quality reference corpora, (2) we performed fuzzy deduplication at the document level, within and across datasets, to prevent redundancy and preserve the integrity of our held-out validation set as an accurate measure of overfitting, and (3) we also added known high-quality reference corpora to the training mix to augment CommonCrawl and increase its diversity.

Details of the first two points (processing of Common Crawl) are described in Appendix A. For the third, we added several curated high-quality datasets, including an expanded version of the WebText dataset [RWC+19], collected by scraping links over a longer period of time, and first described in [KMH+20], two internet-based books corpora (Books1 and Books2) and English-language Wikipedia.

Table 2.2 shows the final mixture of datasets that we used in training. The CommonCrawl data was downloaded from 41 shards of monthly CommonCrawl covering 2016 to 2019, constituting 45TB of compressed plaintext before filtering and 570GB after filtering, roughly equivalent to 400 billion byte-pair-encoded tokens. Note that during training, datasets are not sampled in proportion to their size, but rather datasets we view as higher-quality are sampled more frequently such that CommonCrawl and Books2 datasets are sampled less than once during training, but the other datasets are sampled 2-3 times. This essentially accepts a small amount of overfitting in exchange for higher quality training data.

OpenAI GPT3

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

We are committed to independent auditing of our technologies, and shared some initial steps and ideas in this area in the system card accompanying this release.[2] We plan to make further technical details available to additional third parties who can advise us on how to weigh the competitive and safety considerations above against the scientific value of further transparency.

OpenAI GPT4

1. **Llama 2**, an updated version of Llama 1, trained on a new mix of publicly available data. We also increased the size of the pretraining corpus by 40%, doubled the context length of the model, and adopted grouped-query attention (Ainslie et al., 2023). We are releasing variants of Llama 2 with 7B, 13B, and 70B parameters. We have also trained 34B variants, which we report on in this paper but are not releasing.[§]
2. **Llama 2-Chat**, a fine-tuned version of Llama 2 that is optimized for dialogue use cases. We release variants of this model with 7B, 13B, and 70B parameters as well.

Meta LLaMa 2

一笔概括：更新了上个版本数据且引入了新的数据

大语言模型领域的“怪圈文化”

2.2 Training Dataset

Datasets for language models have rapidly expanded, culminating in the Common Crawl dataset[2] [RSR+19] constituting nearly a trillion words. This size of dataset is sufficient to train our largest models without ever updating on the same sequence twice. However, we have found that unfiltered or lightly filtered versions of Common Crawl tend to have lower quality than more curated datasets. Therefore, we took 3 steps to improve the average quality of our datasets (1) we downloaded and filtered a version of CommonCrawl based on similarity to a range of high-quality reference corpora, (2) we performed fuzzy deduplication at the document level, within and across datasets, to prevent redundancy and preserve the integrity of our held-out validation set as an accurate measure of overfitting, and (3) we also added known high-quality reference corpora to the training mix to augment CommonCrawl and increase its diversity.

Details of the first two points (processing of Common Crawl) are described in Appendix A. For the third, we added several curated high-quality datasets, including an expanded version of the WebText dataset [RWC+19], collected by scraping links over a longer period of time, and first described in [KMH+20], two internet-based books corpora (Books1 and Books2) and English-language Wikipedia.

Table 2.2 shows the final mixture of datasets that we used in training. The CommonCrawl data was downloaded from 41 shards of monthly CommonCrawl covering 2016 to 2019, constituting 45TB of compressed plaintext before filtering and 570GB after filtering, roughly equivalent to 400 billion byte-pair-encoded tokens. Note that during training, datasets are not sampled in proportion to their size, but rather datasets we view as higher-quality are sampled more frequently such that CommonCrawl and Books2 datasets are sampled less than once during training, but the other datasets are sampled 2-3 times. This essentially accepts a small amount of overfitting in exchange for higher quality training data.

OpenAI GPT3

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

We are committed to independent auditing of our technologies, and shared some initial steps and ideas in this area in the system card accompanying this release.[2] We plan to make further technical details available to additional third parties who can advise us on how to weigh the competitive and safety considerations above against the scientific value of further transparency.

OpenAI GPT4

简单说了数据组成以及总数据量

1. Llama 2, an updated version of Llama 1, trained on a new mix of publicly available data. We also increased the size of the pretraining corpus by 40%, doubled the context length of the model, and adopted grouped-query attention (Ainslie et al., 2023). We are releasing variants of Llama 2 with 7B, 13B, and 70B parameters. We have also trained 34B variants, which we report on in this paper but are not releasing.[§]
2. Llama 2-Chat, a fine-tuned version of Llama 2 that is optimized for dialogue use cases. We release variants of this model with 7B, 13B, and 70B parameters as well.

Meta LLaMa 2

Pretraining

Training Data

Gemma 2B and 7B are trained on 2T and 6T tokens respectively of primarily-English data from web documents, mathematics, and code. Unlike Gemini, these models are not multimodal, nor are they trained for state-of-the-art performance on multilingual tasks.

We use a subset of the SentencePiece tokenizer (Kudo and Richardson, 2018) of Gemini for compatibility. It splits digits, does not remove extra whitespace, and relies on byte-level encodings for unknown tokens, following the techniques used for both (Chowdhery et al., 2022) and (Gemini Team, 2023). The vocabulary size is 256k tokens.

Google Gemma

大语言模型领域的“怪圈文化”

如何看待微软论文声称 ChatGPT 是 20B (200亿) 参数量的模型?

mo1315：其实单纯大家比参数量是没有多大意义的，人脑的参数量肯定没有大模型AI这么多，但是理解事物和世界的思维、方式显然是远优于AI的，... 阅读全文

Posted by u/AGIbydecember2023 9 months ago

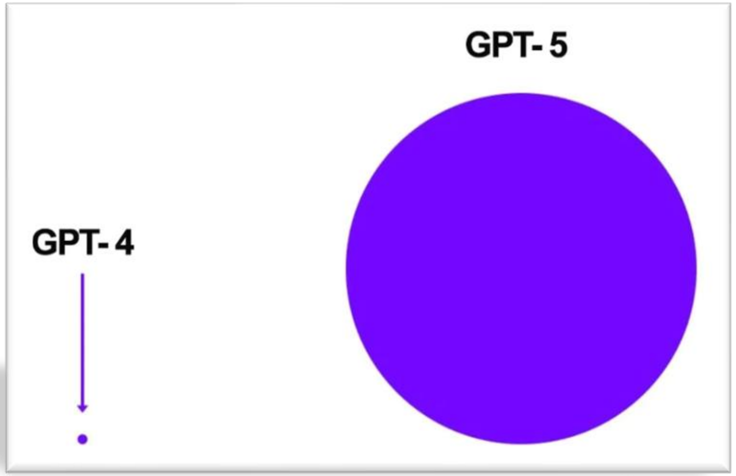
51

GPT-4 has 220billion parameters?

AI

Is this true? I heard George Hotz say this on the Lex podcast. Was he being serious?

37 Comments Share Save ...



大语言模型领域的“怪圈文化”

如何看待微软论文声称 ChatGPT 是 20B (200亿) 参数量的模型?

mo1315：其实单纯大家比参数量是没有多大意义的，人脑的参数量肯定没有大模型AI这么多，但是理解事物和世界的思维、方式显然是远优于AI的，... 阅读全文

Posted by u/AGIbydecember2023 9 months ago

51

GPT-4 has 220billion parameters?

AI

Is this true? I heard George Hotz say this on the Lex podcast. Was he being serious?

37 Comments Share Save ...

```
from openai import OpenAI
client = OpenAI()

completion = client.chat.completions.create(
  model="gpt-3.5-turbo",
  messages=[
    {"role": "system", "content": "You are a poetic assistant, skilled in explai
    {"role": "user", "content": "Compose a poem that explains the concept of rec
  ]
)

print(completion.choices[0].message)
```

GPT-3.5 Turbo

GPT-3.5 Turbo models can understand and generate natural language or code and have been optimized for chat using the Chat Completions API but work well for non-chat tasks as well.

MODEL	DESCRIPTION	CONTEXT WINDOW	TRAINING DATA
gpt-3.5-turbo-0125	New **Updated GPT 3.5 Turbo** The latest GPT-3.5 Turbo model with higher accuracy at responding in requested formats and a fix for a bug which caused a text encoding issue for non-English language function calls. Returns a maximum of 4,096 output tokens. Learn more.	16,385 tokens	Up to Sep 2021
gpt-3.5-turbo	Currently points to gpt-3.5-turbo-0125.	16,385 tokens	Up to Sep 2021
gpt-3.5-turbo-1106	GPT-3.5 Turbo model with improved instruction following, JSON mode, reproducible outputs, parallel function calling, and more. Returns a maximum of 4,096 output tokens. Learn more.	16,385 tokens	Up to Sep 2021
gpt-3.5-turbo-instruct	Similar capabilities as GPT-3 era models. Compatible with legacy Completions endpoint and not Chat Completions.	4,096 tokens	Up to Sep 2021
gpt-3.5-turbo-16k	Legacy Currently points to gpt-3.5-turbo-16k-0613.	16,385 tokens	Up to Sep 2021
gpt-3.5-turbo-0613	Legacy Snapshot of gpt-3.5-turbo from June 13th 2023. Will be deprecated on June 13, 2024.	4,096 tokens	Up to Sep 2021
gpt-3.5-turbo-16k-0613	Legacy Snapshot of gpt-3.5-16k-turbo from June 13th 2023. Will be deprecated on June 13, 2024.	16,385 tokens	Up to Sep 2021

如何培养原创精神？

敏锐捕捉环境变化，敢于定义新问题
（研究的问题不是一成不变的）

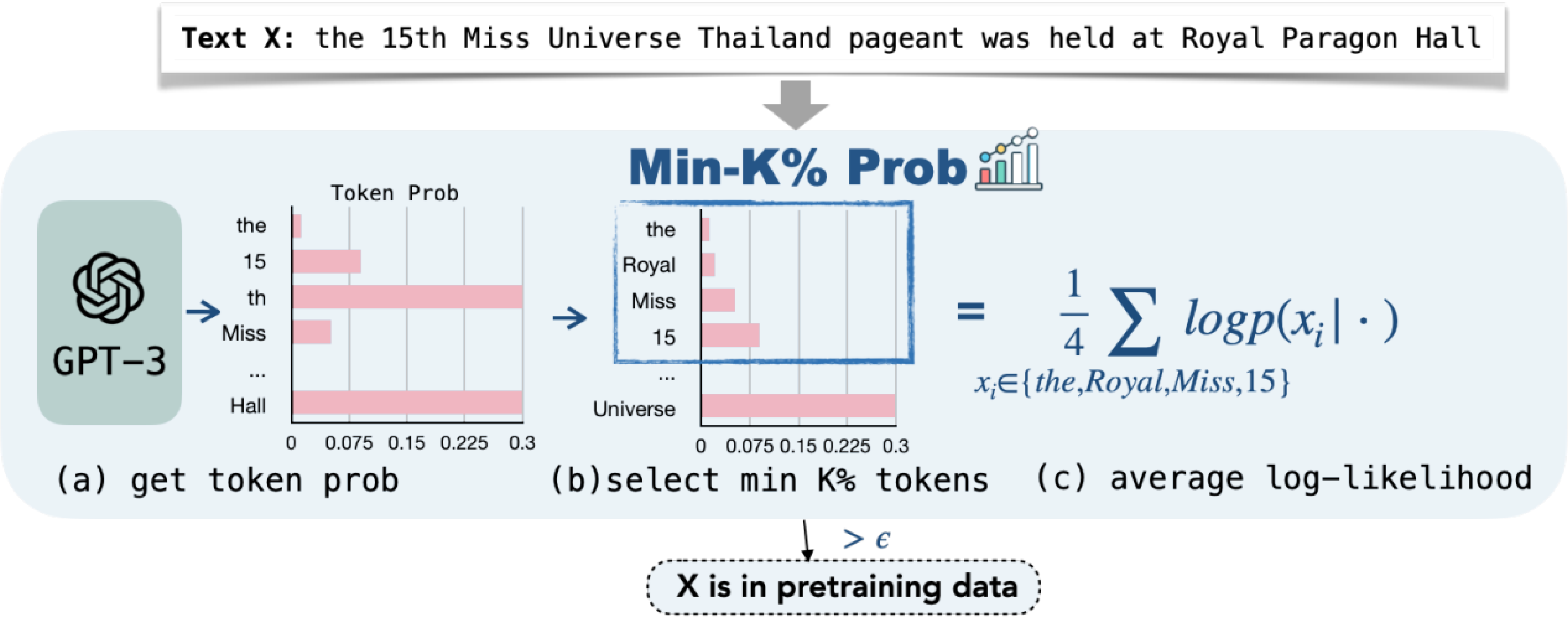
“透明性”驱动的学术研究

Major Dimensions of Transparency	Meta	BigScience	OpenAI	stability.ai	Google	ANTHROP\C	cohere	AI21labs	Inflection	amazon	
	Llama 2	BLOOMZ	GPT-4	Stable Diffusion 2	PaLM 2	Claude 2	Command	Jurassic-2	Inflection-1	Titan Text	**Average**
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	**20%**
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	**17%**
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	**17%**
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	**48%**
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	**63%**
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	**57%**
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	**62%**
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	**24%**
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	**26%**
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	**59%**
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	**44%**
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	**30%**
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	**11%**
Average	**57%**	**52%**	**47%**	**47%**	**41%**	**39%**	**31%**	**20%**	**20%**	**13%**	

10个主要基础模型开发人员在13个主要透明度维度上的得分

The Foundation Model Transparency Index Rishi et al.2023

“透明性”驱动的学术研究



Detecting Pretraining Data from Large Language Models, Shi et al.2024

“透明性”驱动的学术研究

功能

- 20$可以恢复OpenAI “*Babbage*”的embedding projection层
- 2000$可以恢复OpenAI的 “*gpt-3.5-turbo*”

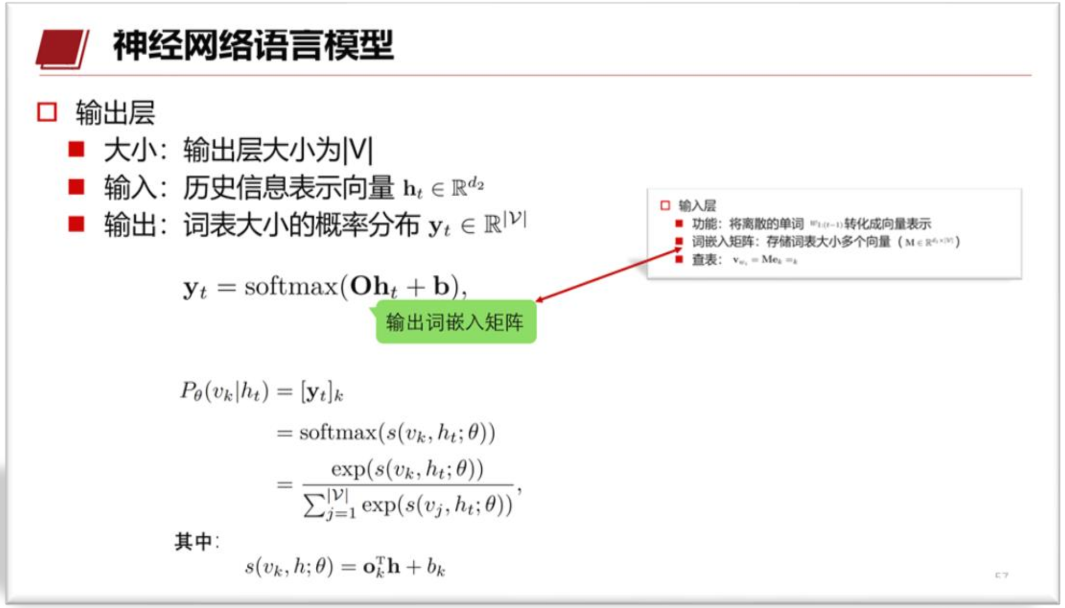
Stealing Part of a Production Language Model

Nicholas Carlini [1] Daniel Paleka [2] Krishnamurthy (Dj) Dvijotham [1] Thomas Steinke [1] Jonathan Hayase [3] A. Feder Cooper [1] Katherine Lee [1] Matthew Jagielski [1] Milad Nasr [1] Arthur Conmy [1] Eric Wallace [4] David Rolnick [5] Florian Tramèr [2]

“透明性”驱动的学术研究

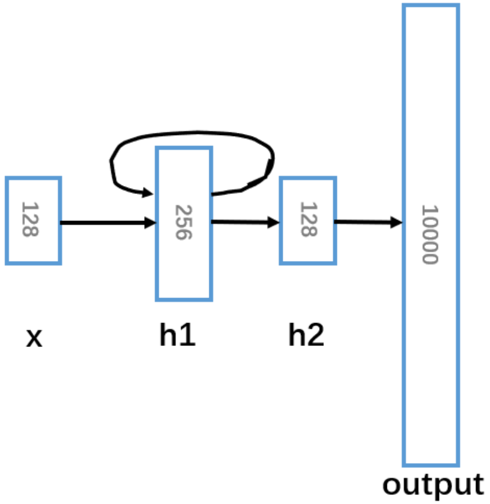
功能

- 20$可以恢复OpenAI “*Babbage*”的embedding projection层
- 2000$可以恢复OpenAI的 “*gpt-3.5-turbo*”



Stealing Part of a Production Language Model

Nicholas Carlini [1] Daniel Paleka [2] Krishnamurthy (Dj) Dvijotham [1] Thomas Steinke [1] Jonathan Hayase [3] A. Feder Cooper [1] Katherine Lee [1] Matthew Jagielski [1] Milad Nasr [1] Arthur Conmy [1] Eric Wallace [4] David Rolnick [5] Florian Tramèr [2]



大语言模型中的“开源”

大语言模型中的“开源”

- 预训练
  - 分布训练架构：高性能分布式训练代码是否公开?
  - 模型架构信息：模型的大小、网络层数等信息是否公开?
  - 训练策略：训练中各种超参数设置?
  - 数据相关信息：预训练数据的组成?
  - 数据的预处理：数据的预处理方法以及处理脚本是否公开?
  - 数据内容：数据本身是否公开?
  - 模型参数：模型完成预训练后的参数是否公开?

大语言模型中的“开源”

- 监督精调
  - 指令数据信息：指令的数据分布、质量、数目等是否公开?
  - 指令数据内容：指令数据本身是否公开?
  - 模型参数：精调后的模型参数是否公开?
- 偏好的对齐
  - 奖励函数：如果是基于奖励函数的对齐，训练方法和模型是否公开?
  - 偏好数据：对齐使用的偏好数据是否公开?
  - 模型参数：偏好对齐后的参数是否公开?

大语言模型中的“开源”

	维度	GPT4	LLaMa2	QWen	Mistral	LLM360	oLMo
预训练	分布式训练架构	x	x	x	x	√	√
	结构信息	x	√	√	√	√	√
	训练策略	x	x	x	x	√	√
	数据信息	x	x	x	x	√	√
	数据处理方式	x	x	x	x	√	√
	数据内容	x	x	x	x	√	√
	模型参数	x	√	√	√	√	√
监督精调	指令数据信息	x	x	x	√	√	√
	指令数据内容	x	x	x	x	√	√
	精调后模型	x	-	-	√	-	√
偏好对齐	奖励函数	x	x	x	-	-	√
	偏好数据	x	x	x	-	x	√
	模型参数	x	√	√	-	√	√

大语言模型中的“开源”

	维度	GPT4	LLaMa2	QWen	Mistral	LLM360	oLMo
预训练	分布式训练架构	x	x	x	x	√	√
	结构信息	x	√	√	√	√	√
	训练策略	x	x	x	x	√	√
	数据信息	x	x	x	x	√	√
	数据处理方式	x	x	x	x	√	√
	数据内容	x	x	x	x	√	√
	模型参数	x	√	√	√	√	√
监督精调	指令数据信息	x	x	x	√	√	√
	指令数据内容	x	x	x	x	√	√
	精调后模型	x	-	-	√	-	√
偏好对齐	奖励函数	x	x	x	-	-	√
	偏好数据	x	x	x	-	x	√
	模型参数	x	√	√	-	√	√

Llama 2: Open Foundation and Fine-Tuned Chat Models, Touvron et al.2023

大语言模型中的“开源”

	维度	GPT4	LLaMa2	QWen	Mistral	LLM360	oLMo
预训练	分布式训练架构	x	x	x	x	√	√
	结构信息	x	√	√	√	√	√
	训练策略	x	x	x	x	√	√
	数据信息	x	x	x	x	√	√
	数据处理方式	x	x	x	x	√	√
	数据内容	x	x	x	x	√	√
	模型参数	x	√	√	√	√	√
监督精调	指令数据信息	x	x	x	√	√	√
	指令数据内容	x	x	x	x	√	√
	精调后模型	x	-	-	√	-	√
偏好对齐	奖励函数	x	x	x	-	-	√
	偏好数据	x	x	x	-	x	√
	模型参数	x	√	√	-	√	√

LLM360: Towards Fully Transparent Open-Source LLMs Liu et al.2023

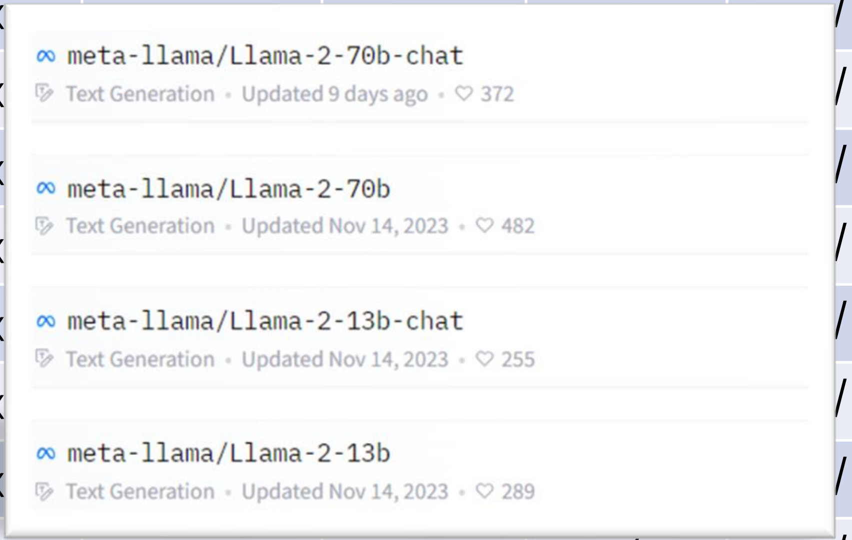
大语言模型中的“开源”

	维度	GPT4	LLaMa2	QWen	Mistral	LLM360	oLMo
预训练	分布式训练架构	x	x	x	x	√	√
	结构信息	x	√	√	√	√	√
	训练策略	x	x	x	x	√	√
	数据信息	x	x	x	x	√	√
	数据处理方式	x	x	x	x	√	√
	数据内容	x	x	x	x	√	√
	模型参数	x	√	√	√	√	√
监督精调	指令数据信息	x	x	x	√	√	√
	指令数据内容	x	x	x	x	√	√
	精调后模型	x	-	-	√	-	√
偏好对齐	奖励函数	x	x	x	-	-	√
	偏好数据	x	x	x	-	x	√
	模型参数	x	√	√	-	√	√

oLMo: Accelerating the Science of Language Models, Groeneveld et al.2024

大语言模型中的“开源”

	维度	GPT4	LLaMa2	QWen	Mistral	LLM360	oLMo
预训练	分布式训练架构	x	[illegible]	[illegible]	[illegible]	[illegible]	√
	结构信息	x	[illegible]	[illegible]	[illegible]	[illegible]	√
	训练策略	x	[illegible]	[illegible]	[illegible]	[illegible]	√
	数据信息	x	[illegible]	[illegible]	[illegible]	[illegible]	√
	数据处理方式	x	[illegible]	[illegible]	[illegible]	[illegible]	√
	数据内容	x	[illegible]	[illegible]	[illegible]	[illegible]	√
	模型参数	x	[illegible]	[illegible]	[illegible]	[illegible]	√
监督精调	指令数据信息	x	x	x	√	√	√
	指令数据内容	x	x	x	x	√	√
	精调后模型	x	-	-	√	-	√
偏好对齐	奖励函数	x	x	x	-	-	√
	偏好数据	x	x	x	-	x	√
	模型参数	x	√	√	-	√	√



oLMo: Accelerating the Science of Language Models, Groeneveld et al.2024

谢谢各位！