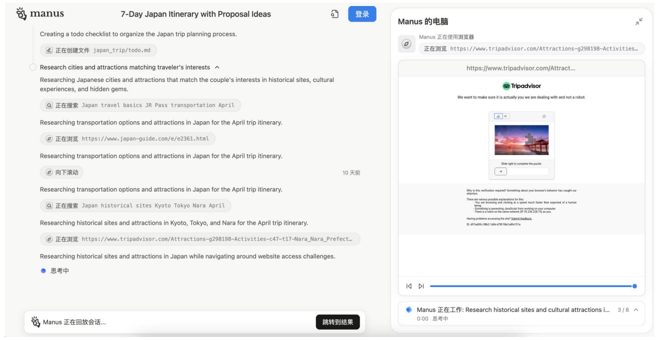
# 一周AI大事

## Manus一夜走红，引巨大争议；OpenManus被5个人3小时开源。



manus　7-Day Japan Itinerary with Proposal Ideas　登录

Creating a todo checklist to organize the Japan trip planning process.

正在创建文件 `japan_trip/todo.md`

Research cities and attractions matching traveler's interests

Researching Japanese cities and attractions that match the couple's interests in historical sites, cultural experiences, and hidden gems.

正在搜索 `Japan travel basics JR Pass transportation April`

Researching transportation options and attractions in Japan for the April trip itinerary.

正在浏览 `https://www.japan-guide.com/e/e2361.html`

Researching transportation options and attractions in Japan for the April trip itinerary.

向下滚动　10 天前

Researching transportation options and attractions in Japan for the April trip itinerary.

正在搜索 `Japan historical sites Kyoto Tokyo Nara April`

Researching historical sites and attractions in Kyoto, Tokyo, and Nara for the April trip itinerary.

正在浏览 `https://www.tripadvisor.com/Attractions-g298198-Activities-c47-t17-Nara_Nara_Prefect…`

Researching historical sites and attractions in Japan while navigating around website access challenges.

思考中

Manus 正在回放会话...　跳转到结果

Manus 的电脑

Manus 正在使用浏览器

正在浏览 `https://www.tripadvisor.com/Attractions-g298198-Activities…`

https://www.tripadvisor.com/Attract…

Tripadvisor

We want to make sure it is actually you we are dealing with and not a robot.

Slide right to complete the puzzle.

Why is this verification required? Something about your browser's behavior has caught our attention.

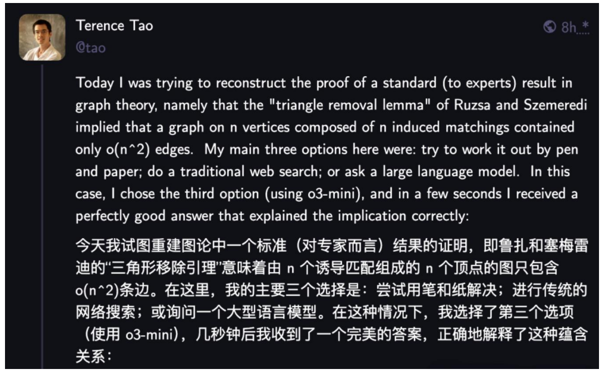
There are various possible explanations for this:
- You are browsing and clicking at a speed much faster than expected of a human being.
- Something is preventing JavaScript from working on your computer.
- There is a robot on the same network (IP 35.236.228.75) as you.

Having problems accessing the site? Submit feedback.

ID: d97ad00c-58b2-1d6b-d78f-58a1a8fe151a

Manus 正在工作: Research historical sites and cultural attractions i…　3 / 8
0:00　思考中

## 陶哲轩亲测点赞o3-mini，在专家级证明中收到了完美的答案。

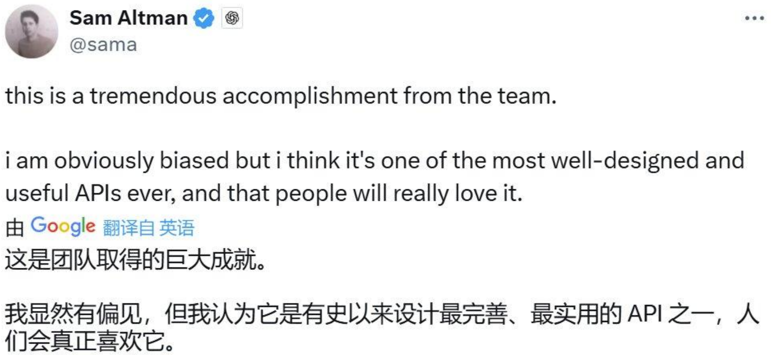


Terence Tao
@tao
8h

Today I was trying to reconstruct the proof of a standard (to experts) result in graph theory, namely that the "triangle removal lemma" of Ruzsa and Szemeredi implied that a graph on n vertices composed of n induced matchings contained only o(n^2) edges. My main three options here were: try to work it out by pen and paper; do a traditional web search; or ask a large language model. In this case, I chose the third option (using o3-mini), and in a few seconds I received a perfectly good answer that explained the implication correctly:

今天我试图重建图论中一个标准（对专家而言）结果的证明，即鲁扎和塞梅雷迪的“三角形移除引理”意味着由 n 个诱导匹配组成的 n 个顶点的图只包含 o(n^2)条边。在这里，我的主要三个选择是：尝试用笔和纸解决；进行传统的网络搜索；或询问一个大型语言模型。在这种情况下，我选择了第三个选项（使用 o3-mini），几秒钟后我收到了一个完美的答案，正确地解释了这种蕴含关系：

## OpenAI发布智能体API，支持网络和文件搜索以及computer use



Sam Altman
@sama

this is a tremendous accomplishment from the team.

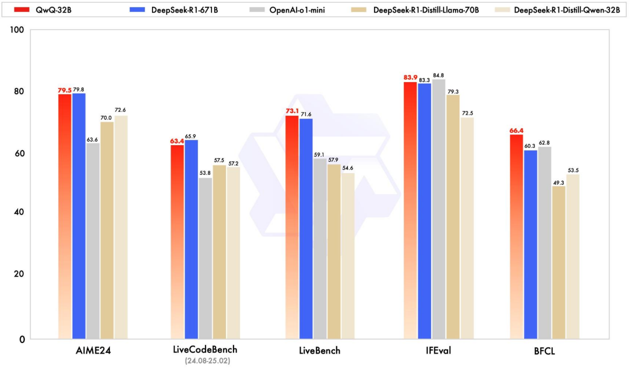
i am obviously biased but i think it's one of the most well-designed and useful APIs ever, and that people will really love it.

由 Google 翻译自 英语

这是团队取得的巨大成就。

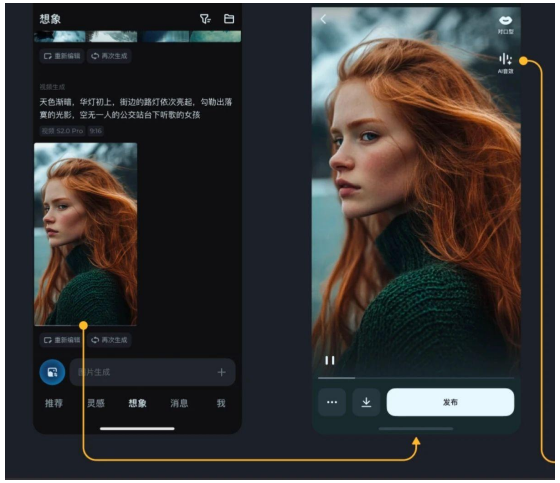
我显然有偏见，但我认为它是有史以来设计最完善、最实用的 API 之一，人们会真正喜欢它。

## 拥抱RL：通义千问发布QwQ-32B



2025-03-07 ~ 2025.03.13

## 字节发布音效生成模型，一键生成大片感音效。



视频生成

天色渐暗，华灯初上，街边的路灯依次亮起，勾勒出落寞的光影，空无一人的公交站台下听歌的女孩

## 通用具身大模型GO-1以及ViLLA架构

“让机器人通过看视频能学会如何动作”

# 预训练

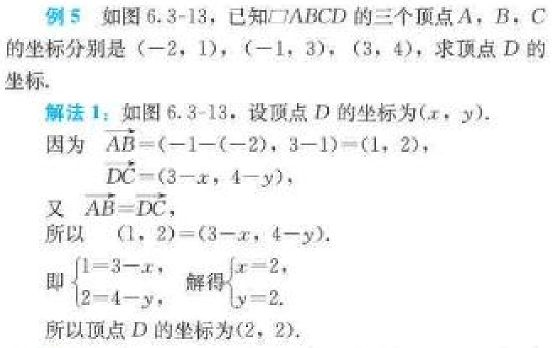
CS2916 大语言模型

飲水思源 愛國榮校

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

# 预训练

□ **以语言模型（类语言模型）**为优化准则在大规模文本语料上进行无监督学习



**例 5**　如图 6.3-13，已知▱$ABCD$ 的三个顶点 $A$，$B$，$C$ 的坐标分别是（−2，1），（−1，3），（3，4），求顶点 $D$ 的坐标.

**解法 1：** 如图 6.3-13，设顶点 $D$ 的坐标为$(x,y)$.

因为　$\overrightarrow{AB}=(-1-(-2),3-1)=(1,2)$，

$\overrightarrow{DC}=(3-x,4-y)$，

又　$\overrightarrow{AB}=\overrightarrow{DC}$，

所以　$(1,2)=(3-x,4-y)$.

即 $\begin{cases}1=3-x,\\2=4-y,\end{cases}$ 解得 $\begin{cases}x=2,\\y=2.\end{cases}$

所以顶点 $D$ 的坐标为$(2,2)$.

# 准确预测出下一个词是一件并不容易的事情

中国的首都是__

# 准确预测出下一个词是一件并不容易的事情

中国的首都是__

小芬对小芳说：“后天的大前天的后天，也就是昨天的昨天的大后天是我的生日，请来参加我的生日会。”小芳应该什么时候赴约呢？ __

# 准确预测出下一个词是一件并不容易的事情

中国的首都是__

小芬对小芳说：“后天的大前天的后天，也就是昨天的昨天的大后天是我的生日，请来参加我的生日会。”小芳应该什么时候赴约呢？ __

这天，柯南收到了一封来自大版的信…(此处省略数千字)…凶手是__

# 准确预测出下一个词是一件并不容易的事情

中国的首都是__

小芬对小芳说：“后天的大前天的后天，也就是昨天的昨天的大后天是我的生日，请来参加我的生日会。”小芳应该什么时候赴约呢？ __

这天，柯南收到了一封来自大版的信…(此处省略数千字)…凶手是__

**1234567 * 54321 + 1234567 / 2 = __**

# 建模世界所有的文本

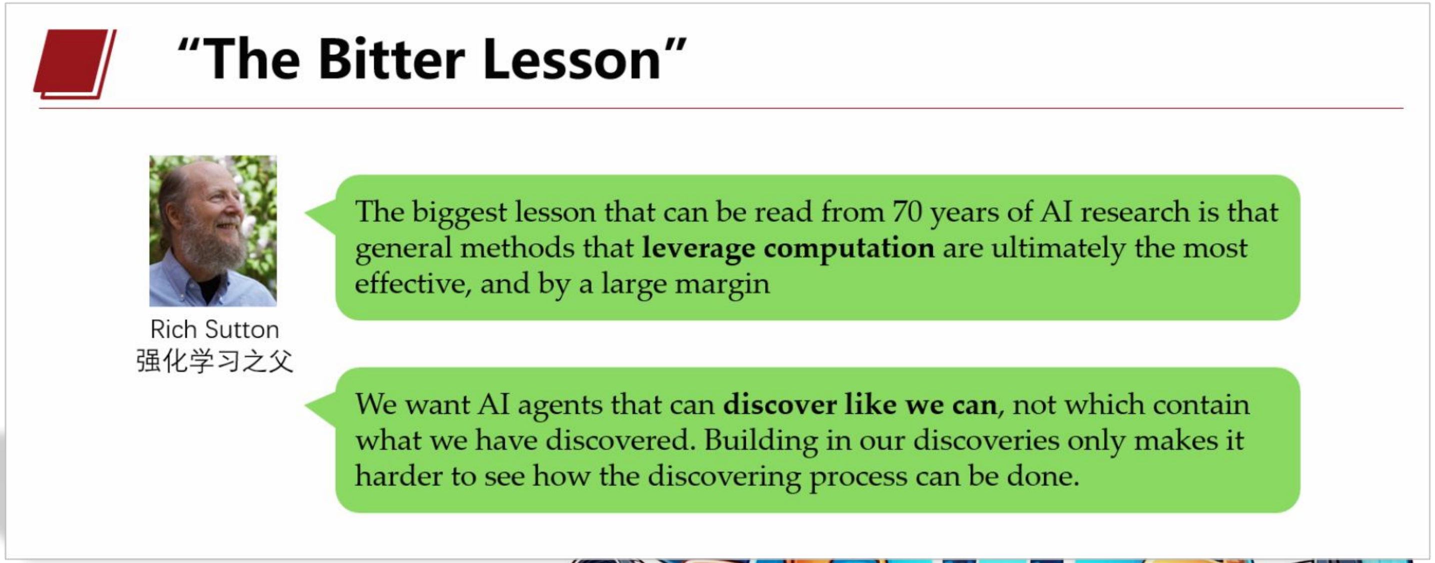
$$P(w_1, \cdots, w_T)$$

# 建模世界

$$P(w_1, \cdots, w_T)$$

- Humanoid Locomotion as Next Token Prediction arXiv 2024
- Genie: Generative Interactive Environments, arXiv 2024

# 建模世界

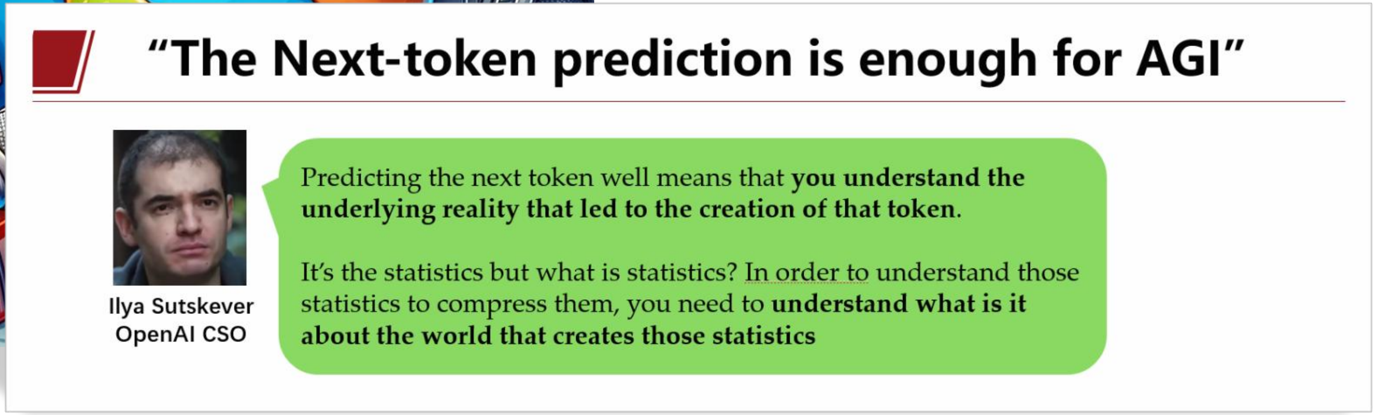


## “The Bitter Lesson”

Rich Sutton
强化学习之父

The biggest lesson that can be read from 70 years of AI research is that general methods that **leverage computation** are ultimately the most effective, and by a large margin

We want AI agents that can **discover like we can**, not which contain what we have discovered. Building in our discoveries only makes it harder to see how the discovering process can be done.



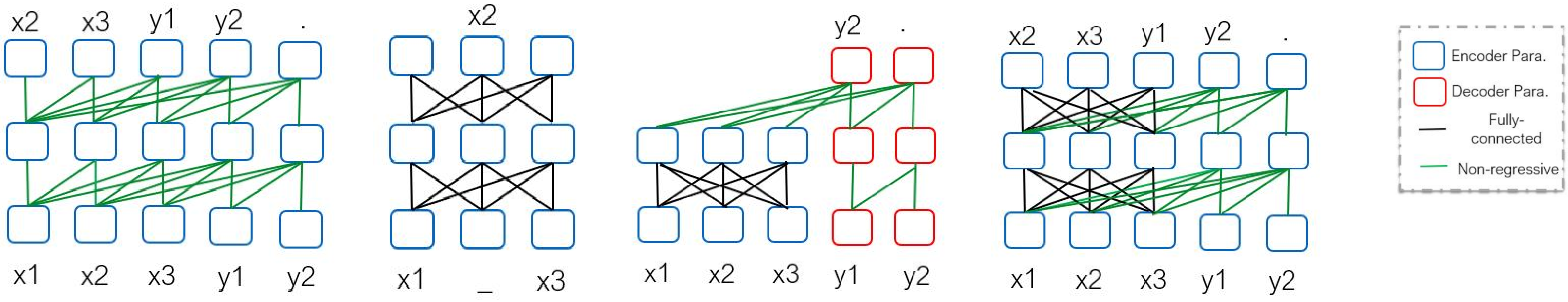
## “The Next-token prediction is enough for AGI”

Ilya Sutskever
OpenAI CSO

Predicting the next token well means that **you understand the underlying reality that led to the creation of that token**.

It’s the statistics but what is statistics? In order to understand those statistics to compress them, you need to **understand what is it about the world that creates those statistics**

# “Big Four” Pretraining Framework



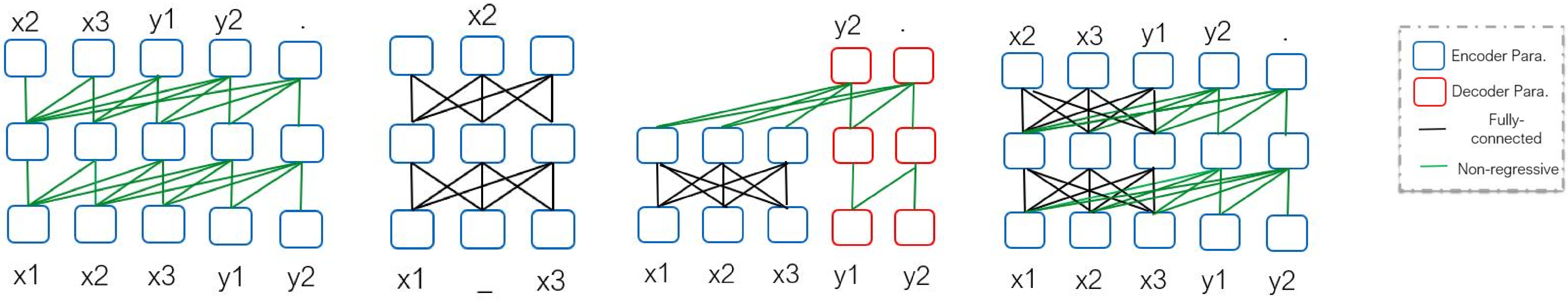
Left-to-right

Masked LM

Encode-decoder

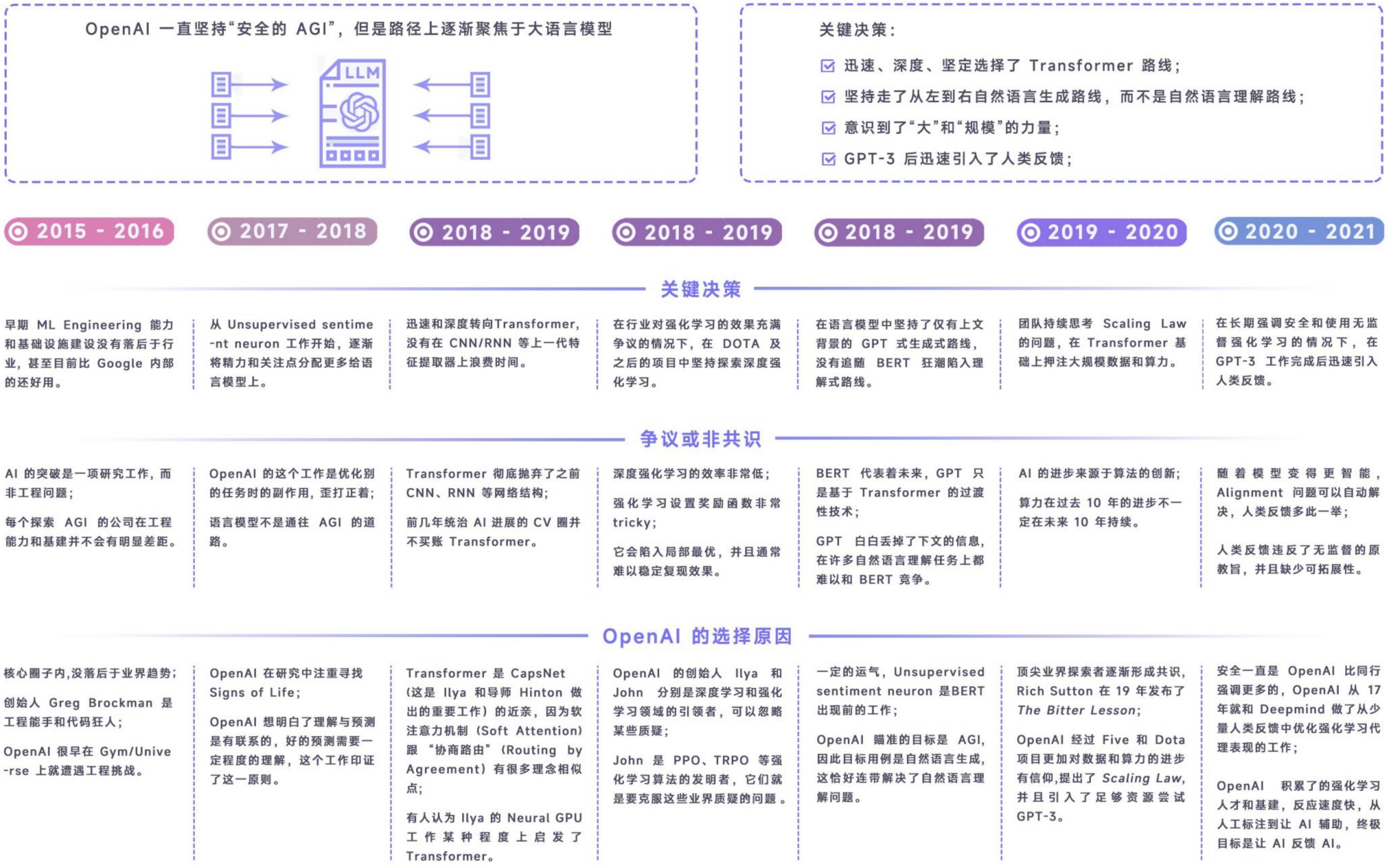
Prefixed LM

# “Big Four” Pretraining Framework



| | | | |
|---|---|---|---|
| Left-to-right | Masked LM | Encode-decoder | Prefixed LM |
| **unidirectional** | **no decoder** | **more params** | **limited capacity** |
| GPT1/2/3 | BERT | MASS/T5/BART | UNiLM/T5 |

# 大语言模型历史梳理 OpenAI视角



OpenAI 一直坚持“安全的 AGI”，但是路径上逐渐聚焦于大语言模型

关键决策：

☑ 迅速、深度、坚定选择了 Transformer 路线；

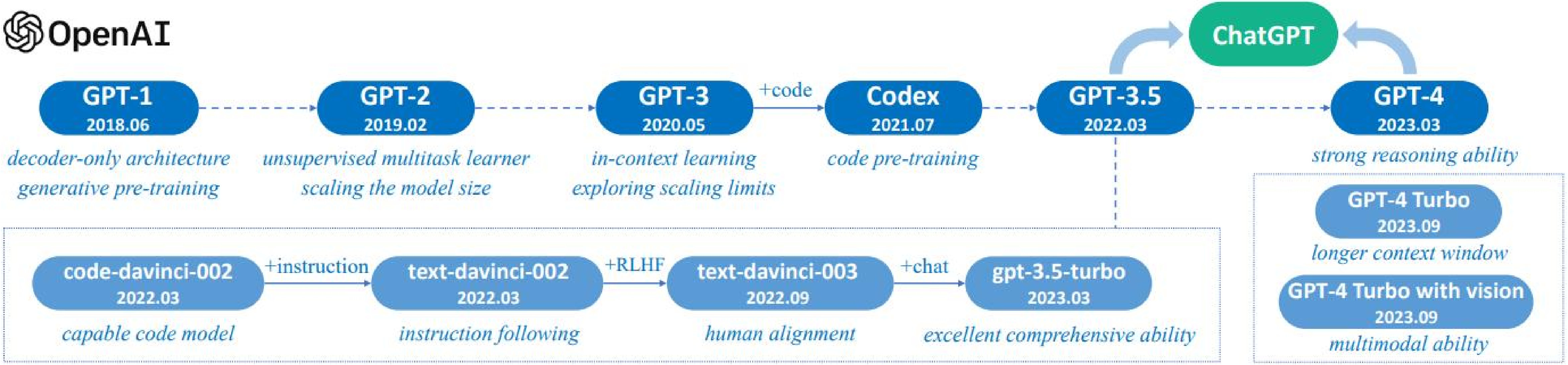
☑ 坚持走了从左到右自然语言生成路线，而不是自然语言理解路线；

☑ 意识到了“大”和“规模”的力量；

☑ GPT-3 后迅速引入了人类反馈；

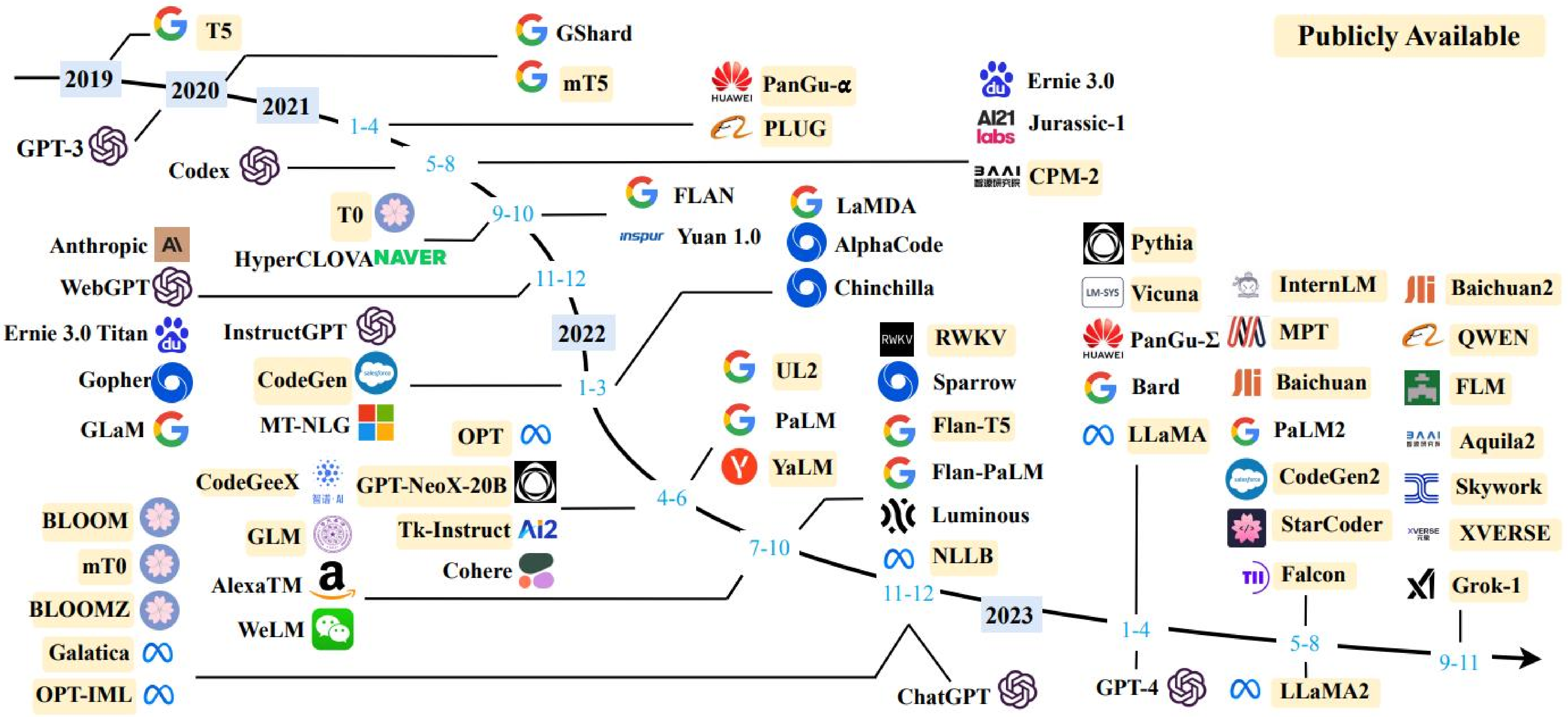
| | ◎ 2015 - 2016 | ◎ 2017 - 2018 | ◎ 2018 - 2019 | ◎ 2018 - 2019 | ◎ 2018 - 2019 | ◎ 2019 - 2020 | ◎ 2020 - 2021 |
|---|---|---|---|---|---|---|---|
| 关键决策 | 早期 ML Engineering 能力和基础设施建设没有落后于行业，甚至目前比 Google 内部的还好用。 | 从 Unsupervised sentiment neuron 工作开始，逐渐将精力和关注点分配更多给语言模型上。 | 迅速和深度转向Transformer，没有在 CNN/RNN 等上一代特征提取器上浪费时间。 | 在行业对强化学习的效果充满争议的情况下，在 DOTA 及之后的项目中坚持探索深度强化学习。 | 在语言模型中坚持了仅有上文背景的 GPT 式生成式路线，没有追随 BERT 狂潮陷入理解式路线。 | 团队持续思考 Scaling Law 的问题，在 Transformer 基础上押注大规模数据和算力。 | 在长期强调安全和使用无监督强化学习的情况下，在 GPT-3 工作完成后迅速引入人类反馈。 |
| 争议或非共识 | AI 的突破是一项研究工作，而非工程问题；<br>每个探索 AGI 的公司在工程能力和基建并不会有明显差距。 | OpenAI 的这个工作是优化别的任务时的副作用，歪打正着；<br>语言模型不是通往 AGI 的道路。 | Transformer 彻底抛弃了之前 CNN、RNN 等网络结构；<br>前几年统治 AI 进展的 CV 圈并不买账 Transformer。 | 深度强化学习的效率非常低；<br>强化学习设置奖励函数非常 tricky；<br>它会陷入局部最优，并且通常难以稳定复现效果。 | BERT 代表着未来，GPT 只是基于 Transformer 的过渡性技术；<br>GPT 白白丢掉了下文的信息，在许多自然语言理解任务上都难以和 BERT 竞争。 | AI 的进步来源于算法的创新；<br>算力在过去 10 年的进步不一定在未来 10 年持续。 | 随着模型变得更智能，Alignment 问题可以自动解决，人类反馈多此一举；<br>人类反馈违反了无监督的原教旨，并且缺少可拓展性。 |
| OpenAI 的选择原因 | 核心圈子内,没落后于业界趋势；<br>创始人 Greg Brockman 是工程能手和代码狂人；<br>OpenAI 很早在 Gym/Universe 上就遭遇工程挑战。 | OpenAI 在研究中注重寻找 Signs of Life；<br>OpenAI 想明白了理解与预测是有联系的，好的预测需要一定程度的理解，这个工作印证了这一原则。 | Transformer 是 CapsNet (这是 Ilya 和导师 Hinton 做出的重要工作）的近亲，因为软注意力机制 (Soft Attention) 跟 “协商路由” (Routing by Agreement) 有很多理念相似点；<br>有人认为 Ilya 的 Neural GPU 工作某种程度上启发了 Transformer。 | OpenAI 的创始人 Ilya 和 John 分别是深度学习和强化学习领域的引领者，可以忽略某些质疑；<br>John 是 PPO、TRPO 等强化学习算法的发明者，它们就是要克服这些业界质疑的问题。 | 一定的运气，Unsupervised sentiment neuron 是BERT 出现前的工作；<br>OpenAI 瞄准的目标是 AGI，因此目标用例是自然语言生成，这恰好连带解决了自然语言理解问题。 | 顶尖业界探索者逐渐形成共识，Rich Sutton 在 19 年发布了 *The Bitter Lesson*；<br>OpenAI 经过 Five 和 Dota 项目更加对数据和算力的进步有信仰,提出了 *Scaling Law*，并且引入了足够资源尝试 GPT-3。 | 安全一直是 OpenAI 比同行强调更多的，OpenAI 从 17 年就和 Deepmind 做了从少量人类反馈中优化强化学习代理表现的工作；<br>OpenAI 积累了的强化学习人才和基建，反应速度快，从人工标注到让 AI 辅助，终极目标是让 AI 反馈 AI。 |

# 大语言模型历史梳理 OpenAI视角



图来自于：https://arxiv.org/pdf/2303.18223.pdf

# 大语言模型历史梳理 全局视角



图来自于：https://arxiv.org/pdf/2303.18223.pdf

# 大语言模型相关资源：文献

- □ A Survey of Large Language Models, Zhao et al.2023
- □ Pre-trained models for natural language processing: A survey
- □ Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing

# 大语言模型相关资源：周边

PLM Emoji (English)

PLM Emoji (Chinese)

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

# 大语言模型发展中的重要问题讨论

CS2916 大语言模型

飲水思源 爱國榮校

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

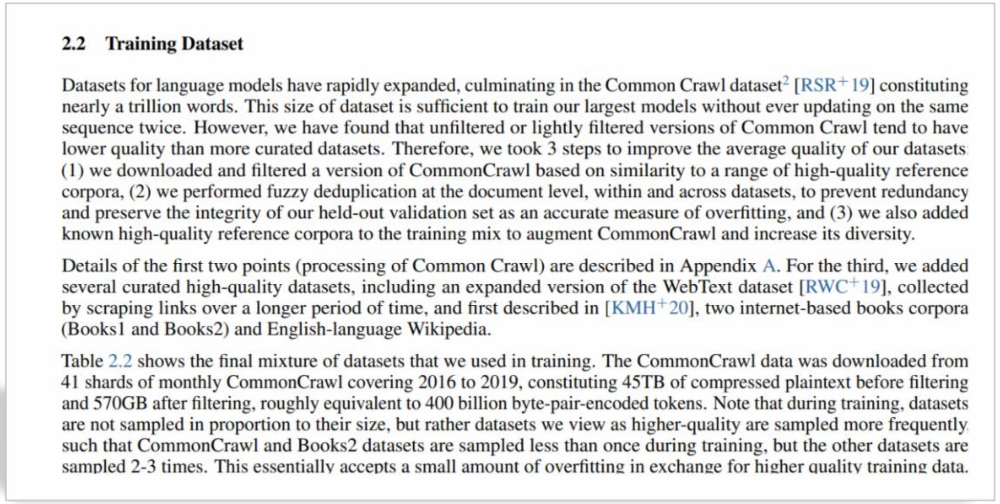
比尔盖茨

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.

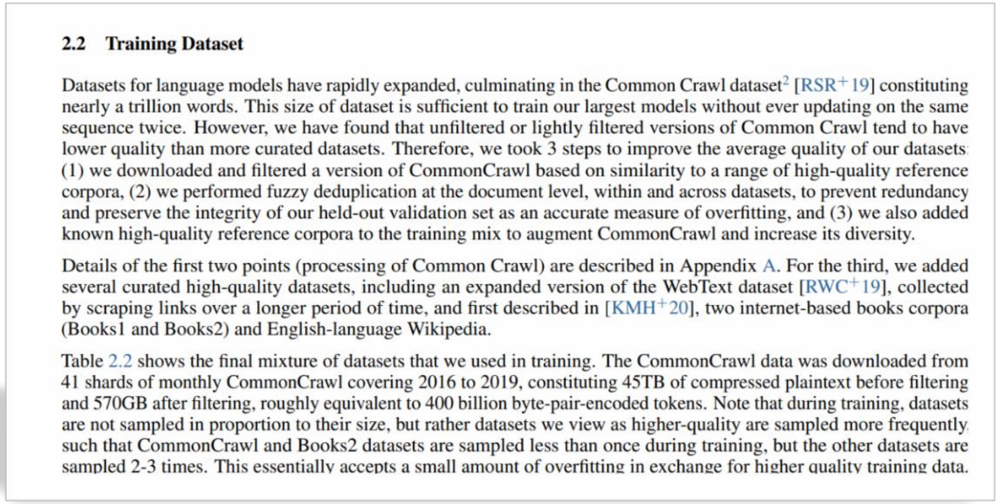
Details of the first two points (processing of Common Crawl) are described in Appendix A. For the third, we added several curated high-quality datasets, including an expanded version of the WebText dataset [RWC$^{+}$19], collected by scraping links over a longer period of time, and first described in [KMH$^{+}$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.

细致地描述使用的预训练语料，包括组成、大小、过滤方法

OpenAI GPT3

# 大语言模型领域的“怪圈文化”



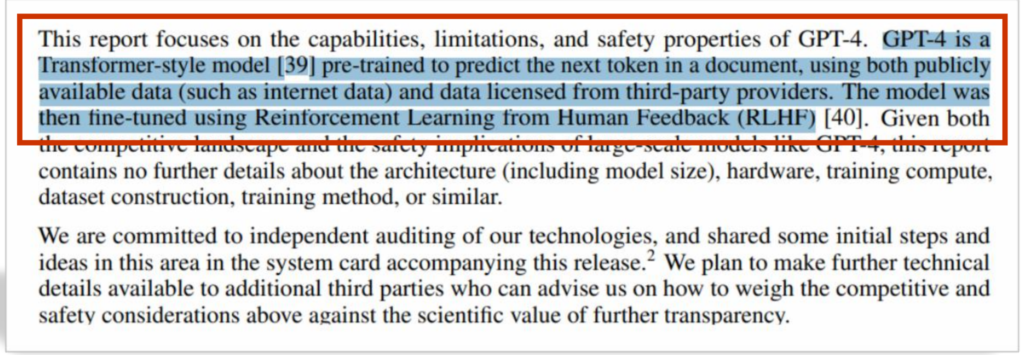
### 2.2 Training Dataset

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



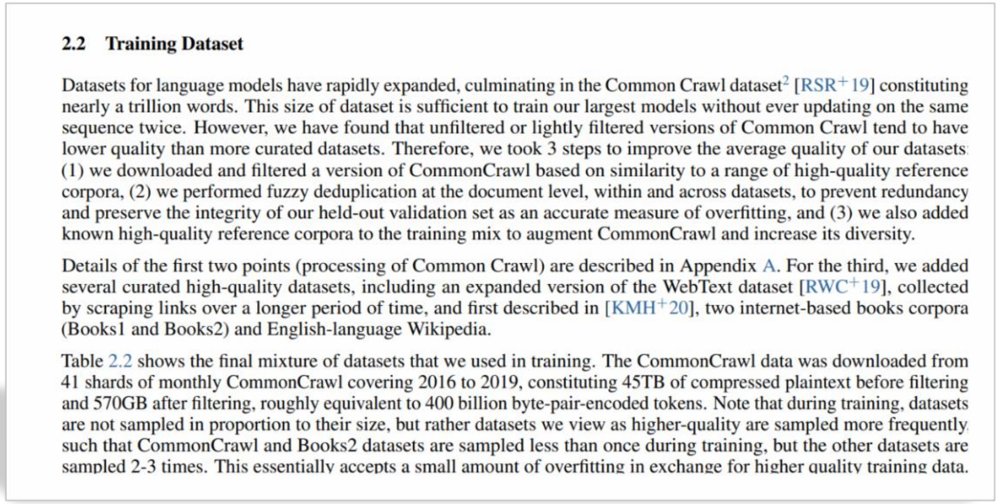
This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then 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”



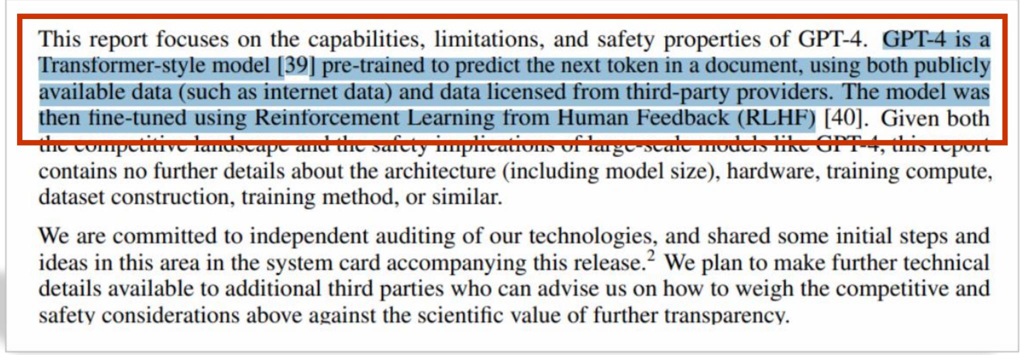
### 2.2 Training Dataset

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.

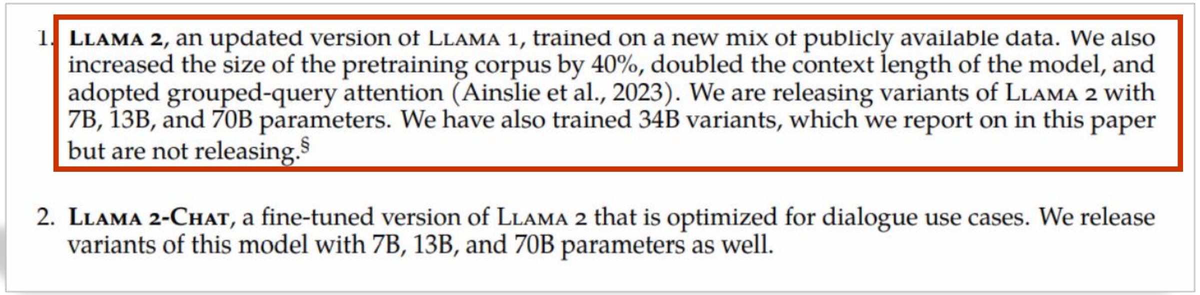
OpenAI GPT3



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



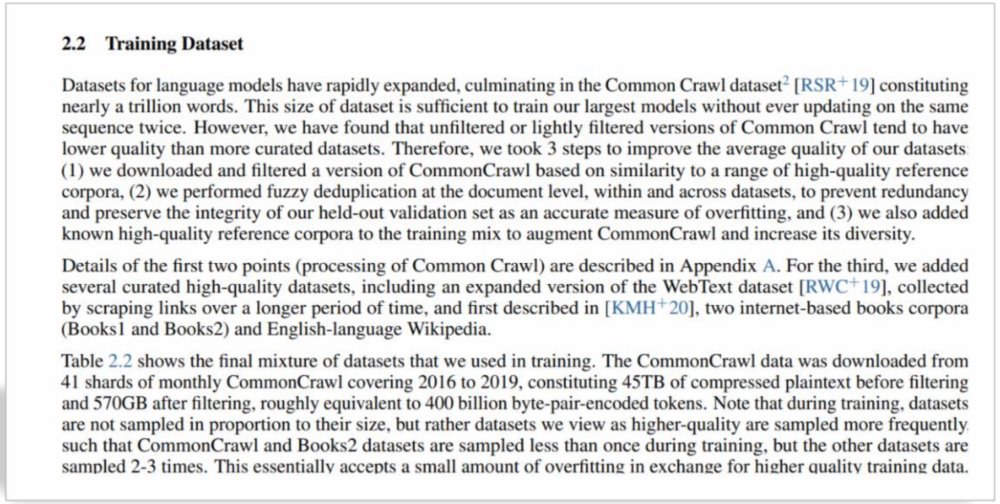
1. Llama 2, an updated version of Llama 1, trained on a new mix of publicly available data. We also increased the size of the pretraining corpus by 40%, doubled the context length of the model, and adopted grouped-query attention (Ainslie et al., 2023). We are releasing variants of Llama 2 with 7B, 13B, 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.

一笔概括：更新了上个版本数据且引入了新的数据

Meta LLaMa 2

# 大语言模型领域的“怪圈文化”



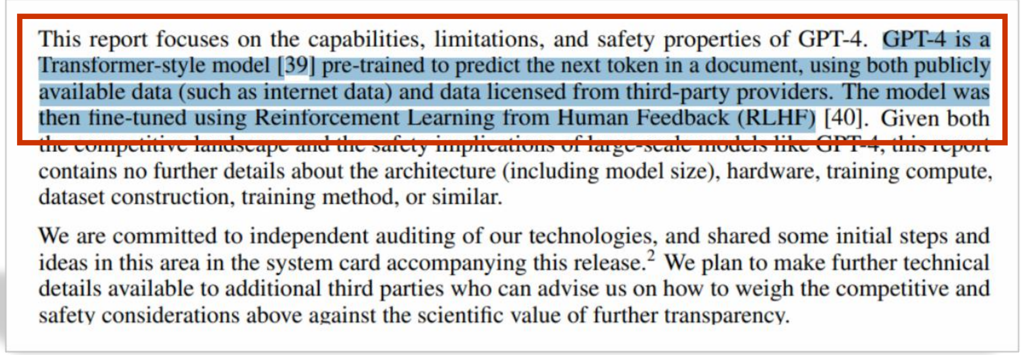
### 2.2 Training Dataset

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

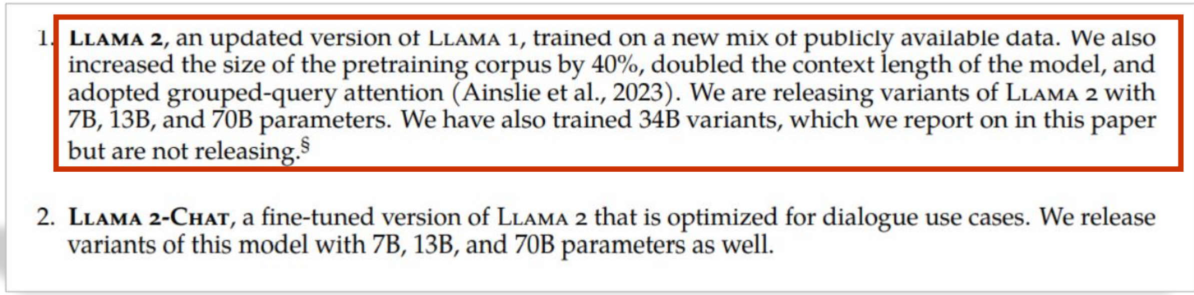


This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then 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.

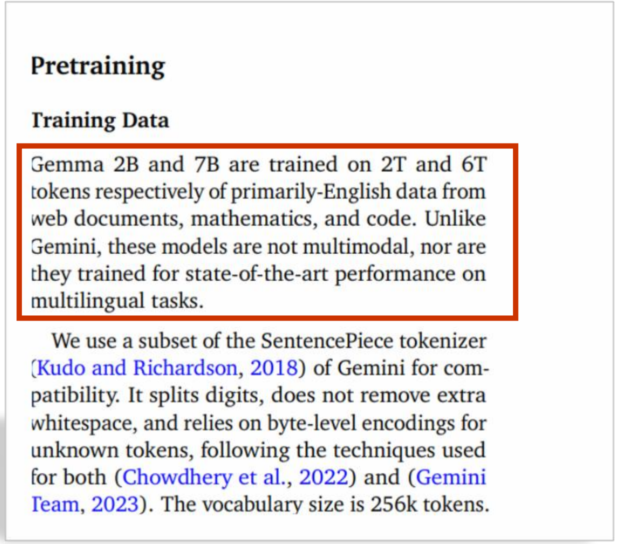
OpenAI GPT4

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**

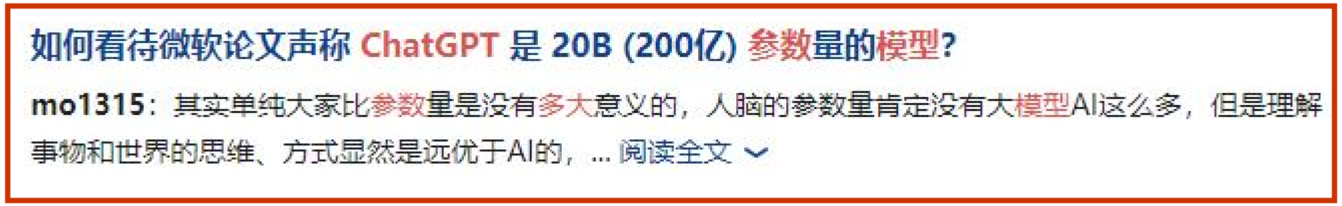
**Training Data**

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.

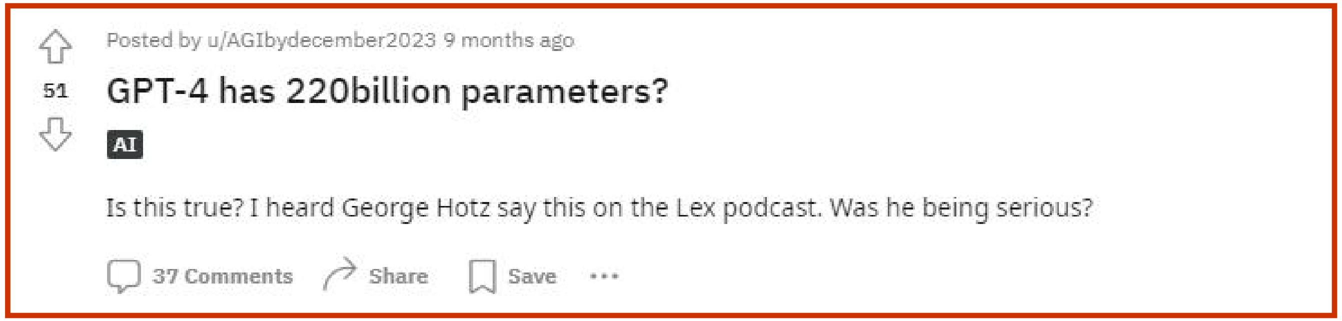
Google Gemma

# 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?

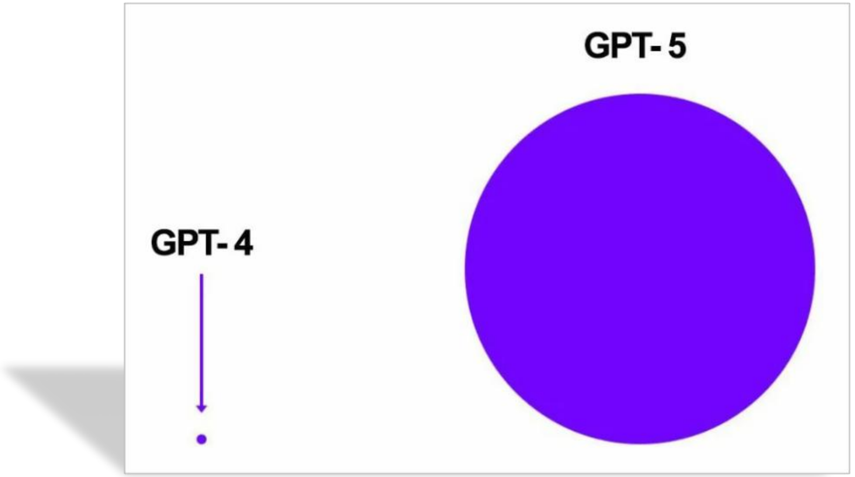
mo1315：其实单纯大家比参数量是没有多大意义的，人脑的参数量肯定没有大模型AI这么多，但是理解事物和世界的思维、方式显然是远优于AI的，... 阅读全文



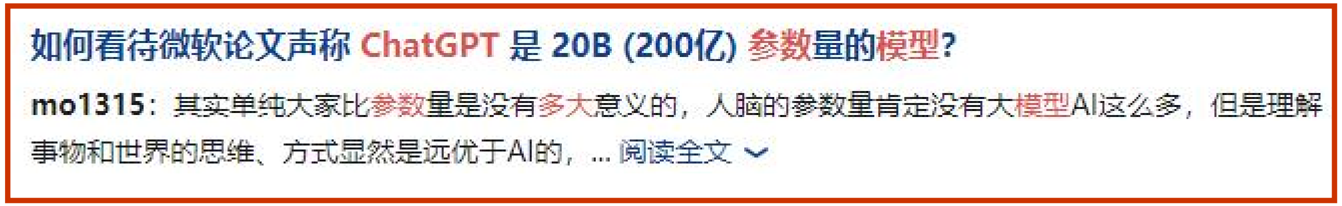
Posted by u/AGIbydecember2023 9 months ago

GPT-4 has 220billion parameters?

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

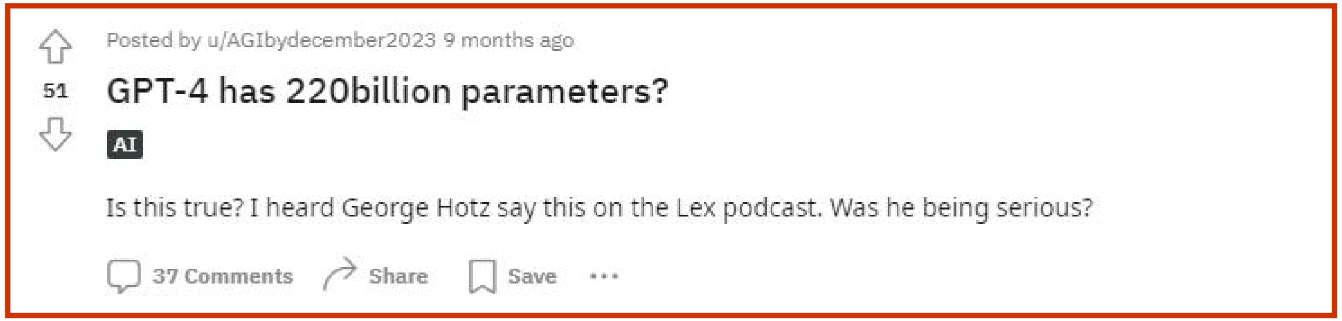


# 大语言模型领域的“怪圈文化”



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 ⌄



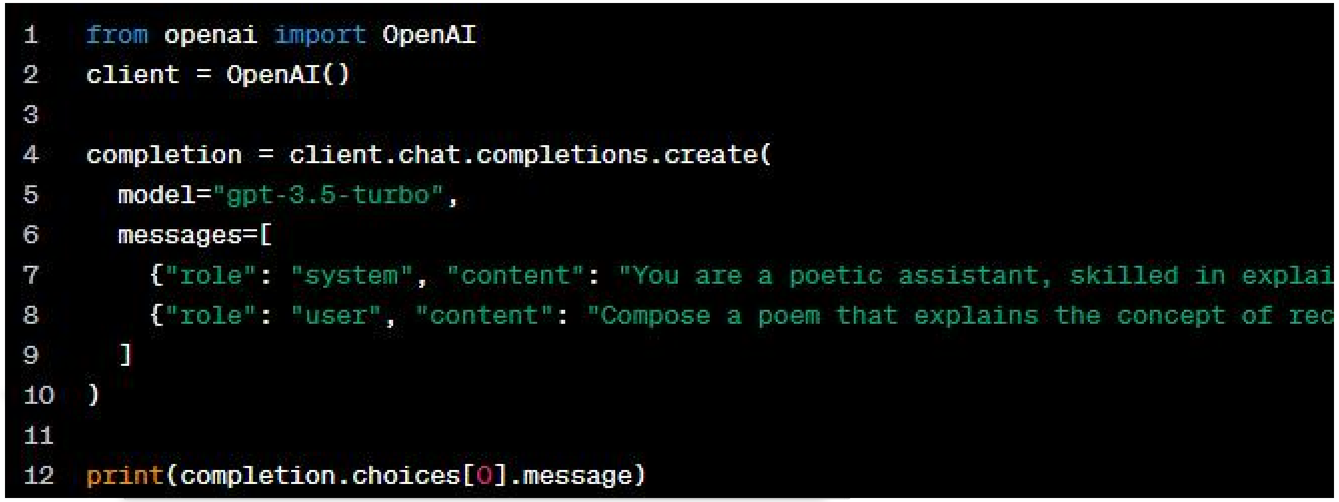
Posted by u/AGIbydecember2023 9 months ago

51

GPT-4 has 220billion parameters?

AI

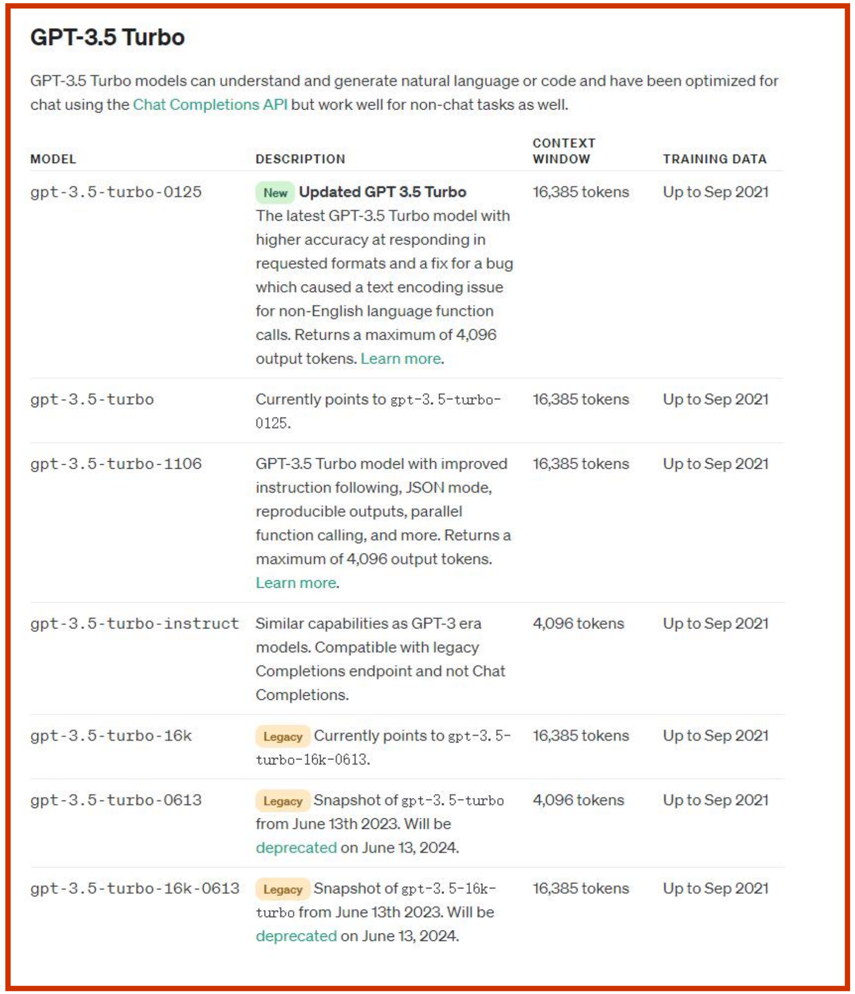
Is this true? I heard George Hotz say this on the Lex podcast. Was he being serious?



```
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

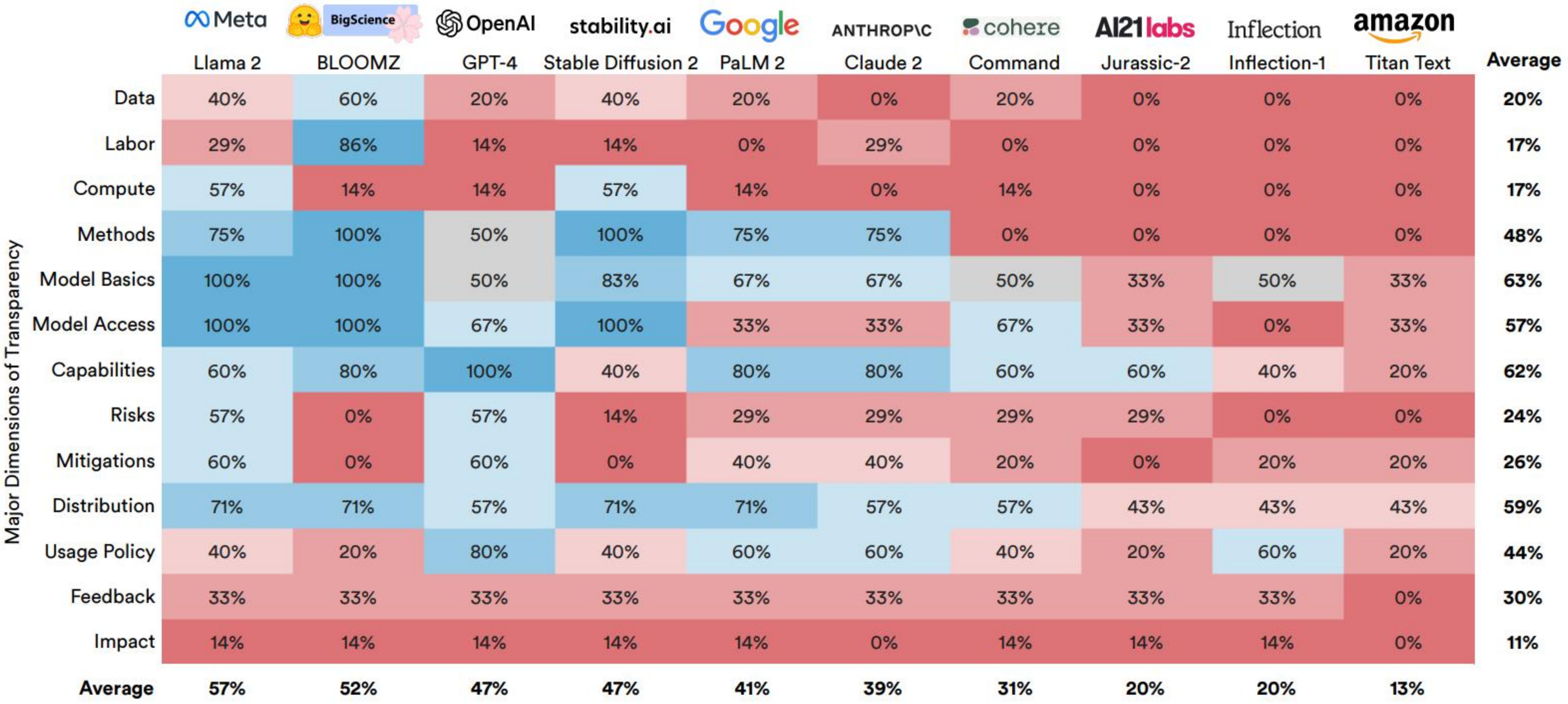
GPT-3.5 Turbo models can understand and generate natural language or code and have been optimized for chat using the Chat Completions API but work well for non-chat tasks as well.

| 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

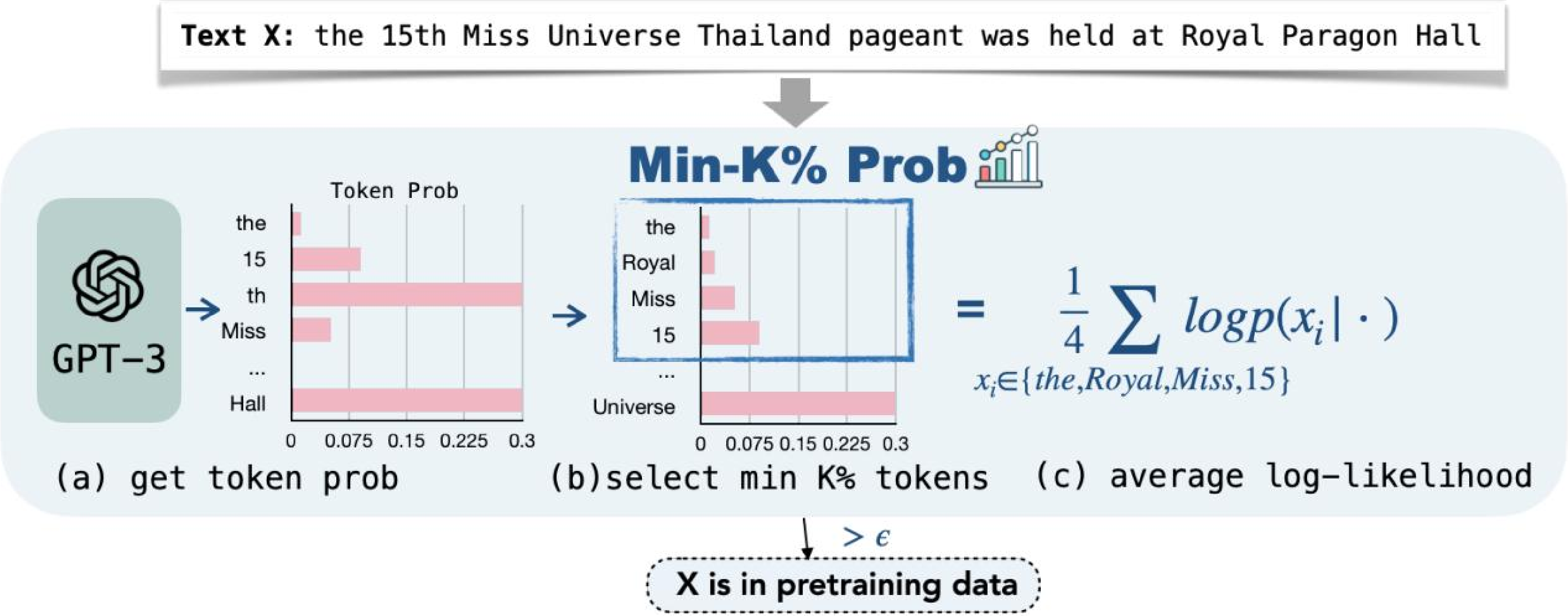


| Major Dimensions of Transparency | Meta | BigScience | OpenAI | stability.ai | Google | ANTHROP\C | cohere | AI21labs | Inflection | amazon | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Llama 2 | BLOOMZ | GPT-4 | Stable Diffusion 2 | PaLM 2 | Claude 2 | Command | Jurassic-2 | Inflection-1 | 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

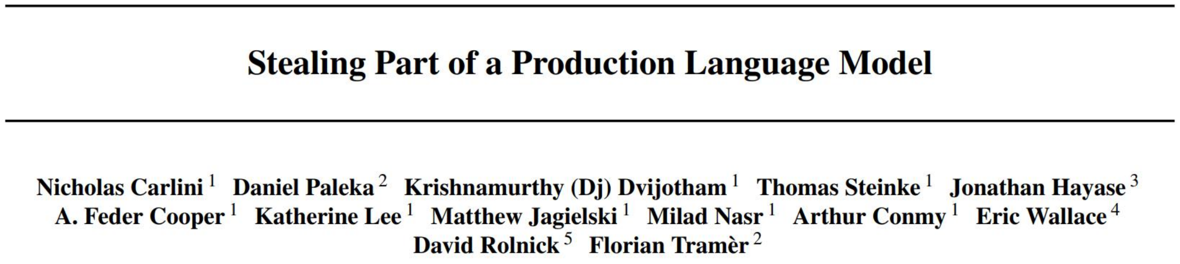


Detecting Pretraining Data from Large Language Models, Shi et al.2024

# “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*"



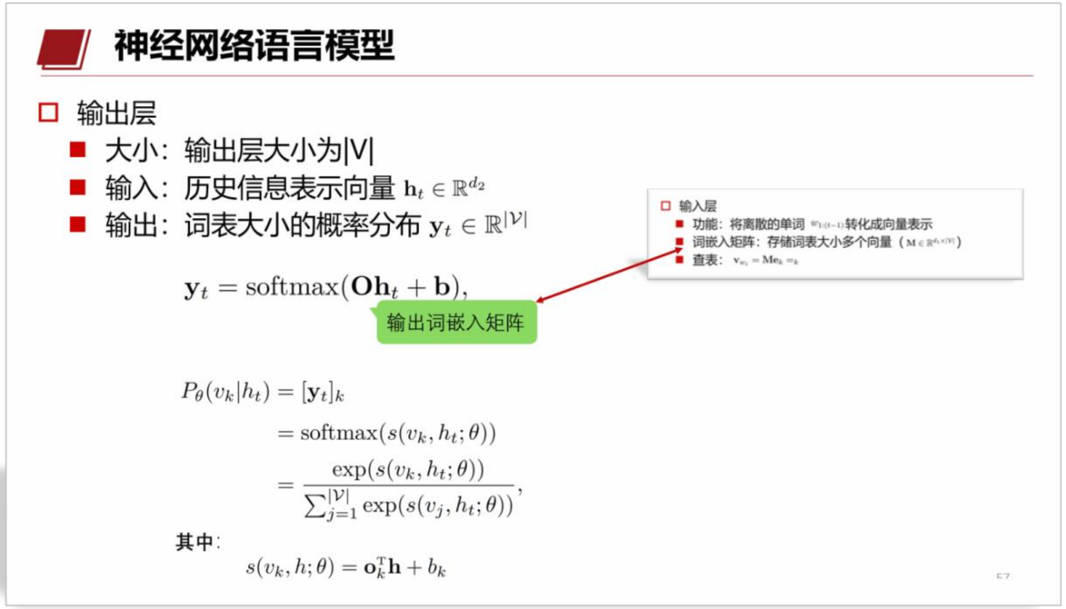
**Stealing Part of a Production Language Model**

**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*”



### 神经网络语言模型

□ 输出层

- 大小：输出层大小为|V|
- 输入：历史信息表示向量 $\mathbf{h}_t \in \mathbb{R}^{d_2}$
- 输出：词表大小的概率分布 $\mathbf{y}_t \in \mathbb{R}^{|\mathcal{V}|}$

$$\mathbf{y}_t = \mathrm{softmax}(\mathbf{O}\mathbf{h}_t + \mathbf{b}),$$

输出词嵌入矩阵

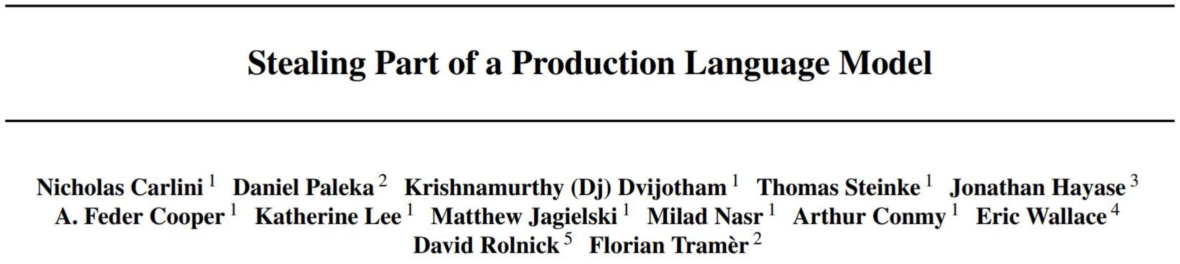
□ 输入层

- 功能：将离散的单词 $w_{1:(t-1)}$转化成向量表示
- 词嵌入矩阵：存储词表大小多个向量（$\mathbf{M} \in \mathbb{R}^{d_1 \times |\mathcal{V}|}$）
- 查表：$\mathbf{v}_{w_i} = \mathbf{M}\mathbf{e}_k =_k$

$$\begin{aligned} P_\theta(v_k|h_t) &= [\mathbf{y}_t]_k \\ &= \mathrm{softmax}(s(v_k, h_t; \theta)) \\ &= \frac{\exp(s(v_k, h_t; \theta))}{\sum_{j=1}^{|\mathcal{V}|} \exp(s(v_j, h_t; \theta))}, \end{aligned}$$

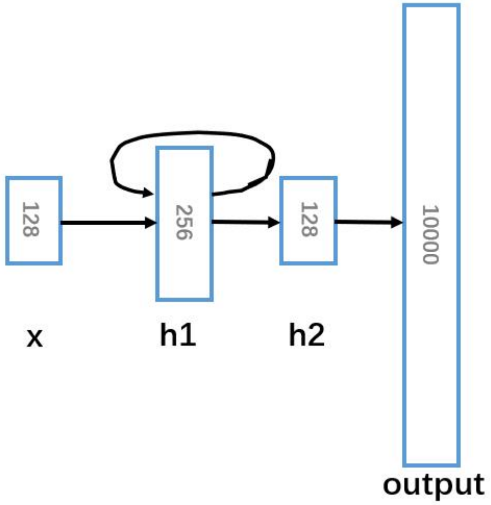
其中：

$$s(v_k, h; \theta) = \mathbf{o}_k^{\mathrm{T}}\mathbf{h} + b_k$$

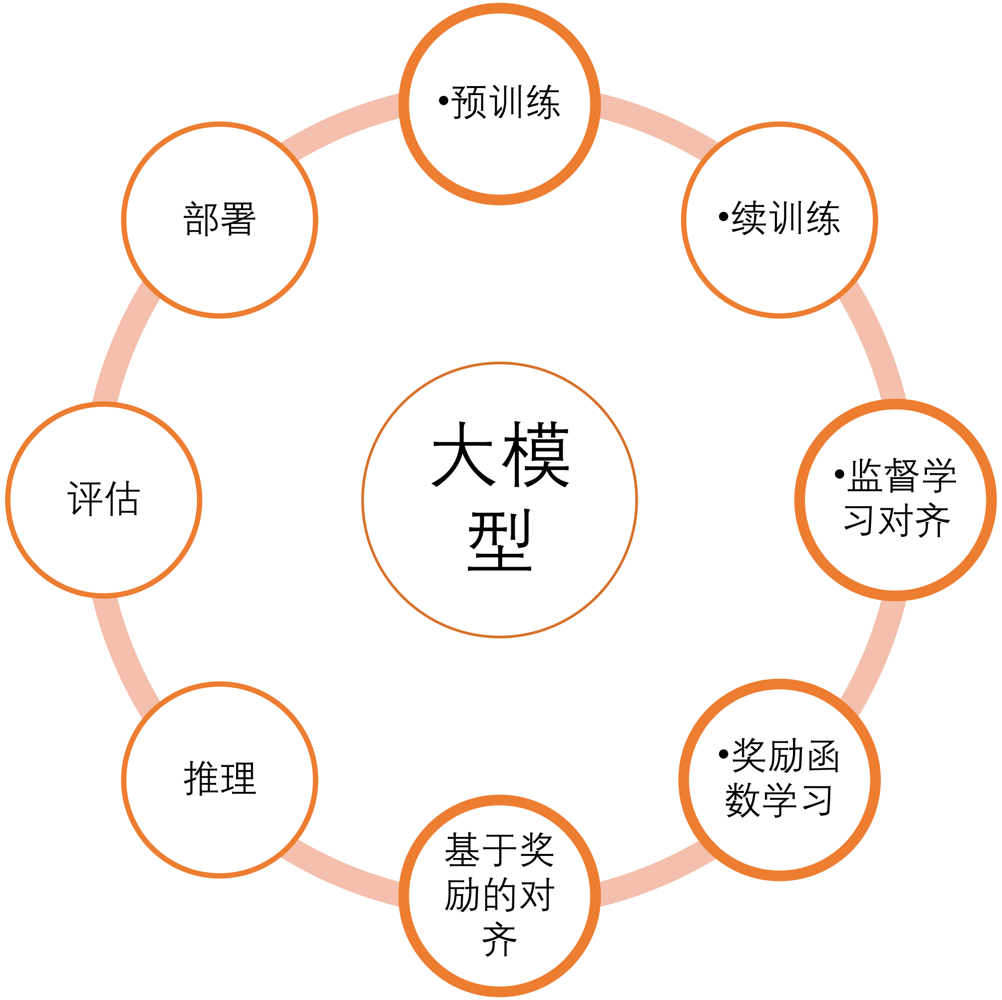


**Stealing Part of a Production Language Model**

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”



大模型

•预训练

•续训练

•监督学习对齐

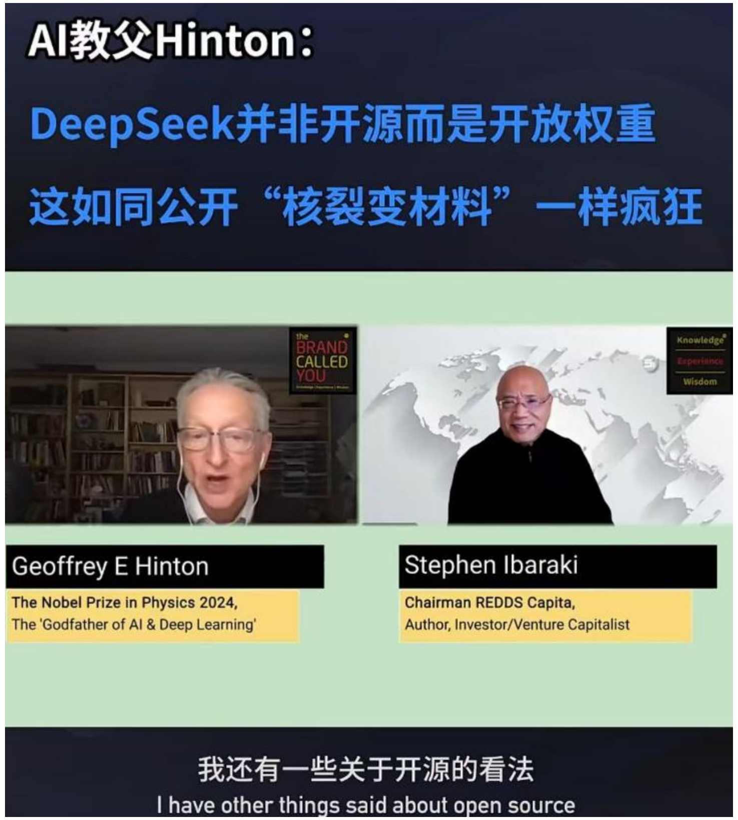
•奖励函数学习

基于奖励的对齐

推理

评估

部署



AI教父Hinton：

DeepSeek并非开源而是开放权重

这如同公开“核裂变材料”一样疯狂

Geoffrey E Hinton

The Nobel Prize in Physics 2024,
The 'Godfather of AI & Deep Learning'

Stephen Ibaraki

Chairman REDDS Capita,
Author, Investor/Venture Capitalist

我还有一些关于开源的看法

I have other things said about open source

# 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

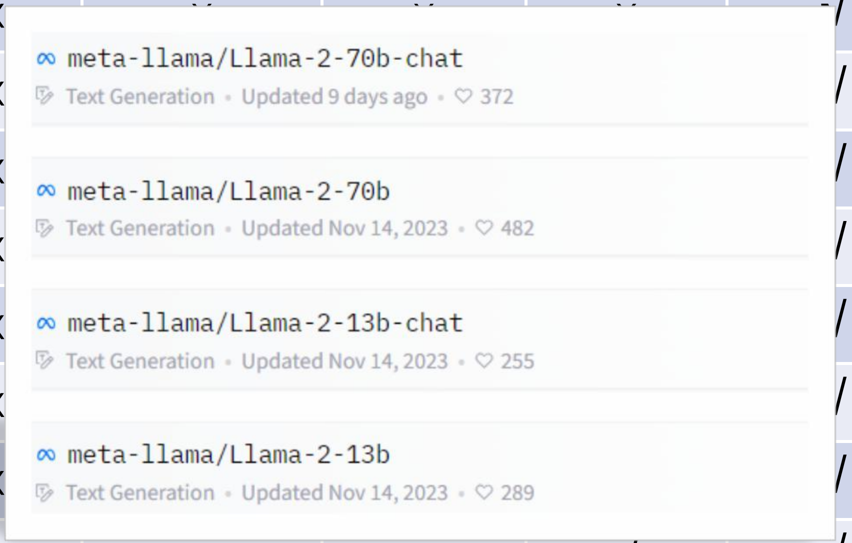
# 大语言模型中的“开源”

| | 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 | x | x | x | √ | √ |
| | 结构信息 | x | | | | √ | √ |
| | 训练策略 | x | | | | √ | √ |
| | 数据信息 | x | | | | √ | √ |
| | 数据处理方式 | x | | | | √ | √ |
| | 数据内容 | 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 | √ | √ | - | √ | √ |



meta-llama/Llama-2-70b-chat
Text Generation • Updated 9 days ago • ♡ 372

meta-llama/Llama-2-70b
Text Generation • Updated Nov 14, 2023 • ♡ 482

meta-llama/Llama-2-13b-chat
Text Generation • Updated Nov 14, 2023 • ♡ 255

meta-llama/Llama-2-13b
Text Generation • Updated Nov 14, 2023 • ♡ 289

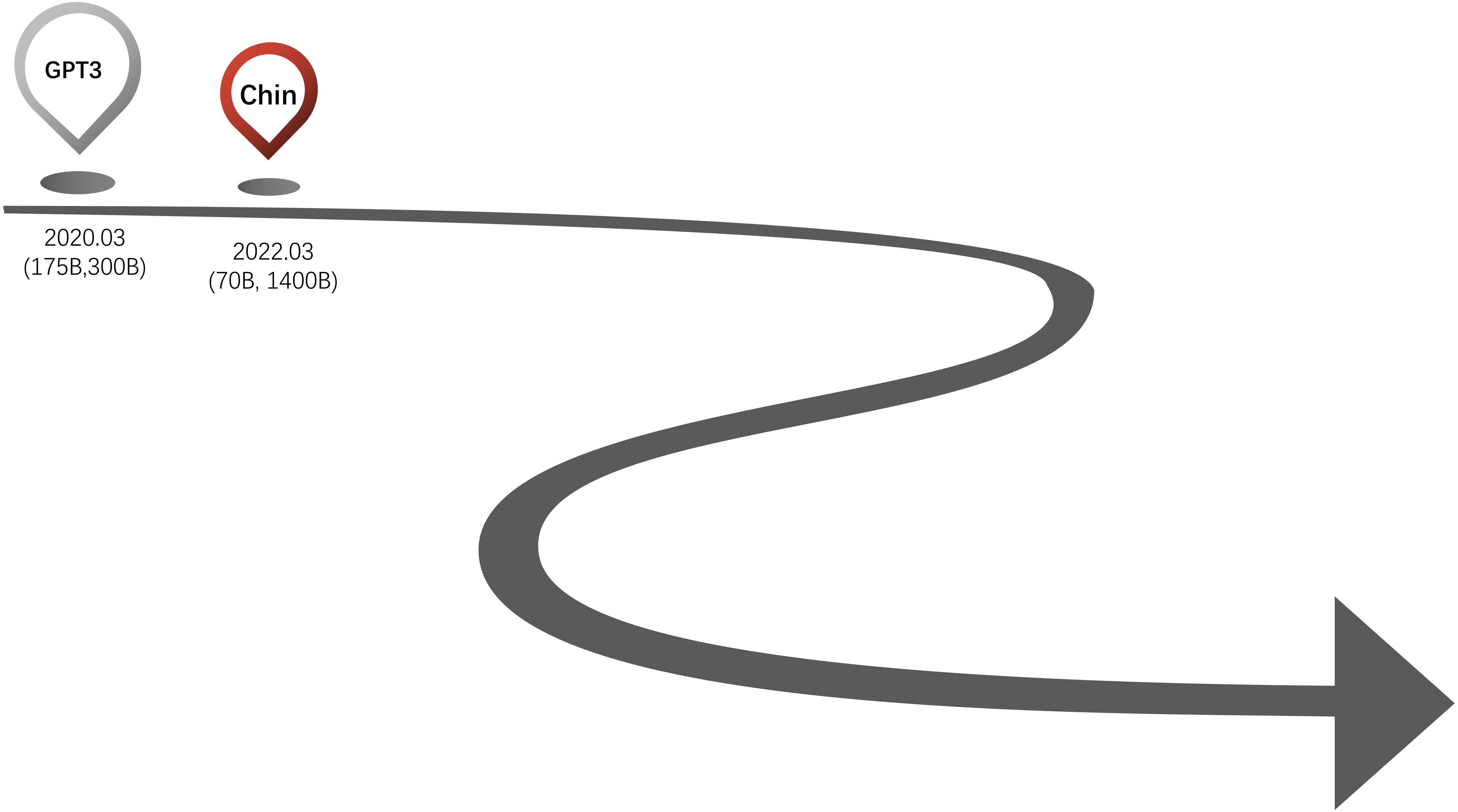
oLMo: Accelerating the Science of Language Models, Groeneveld et al.2024

# Llama家族

GPT3

2020.03
(175B,300B)

# Llama家族



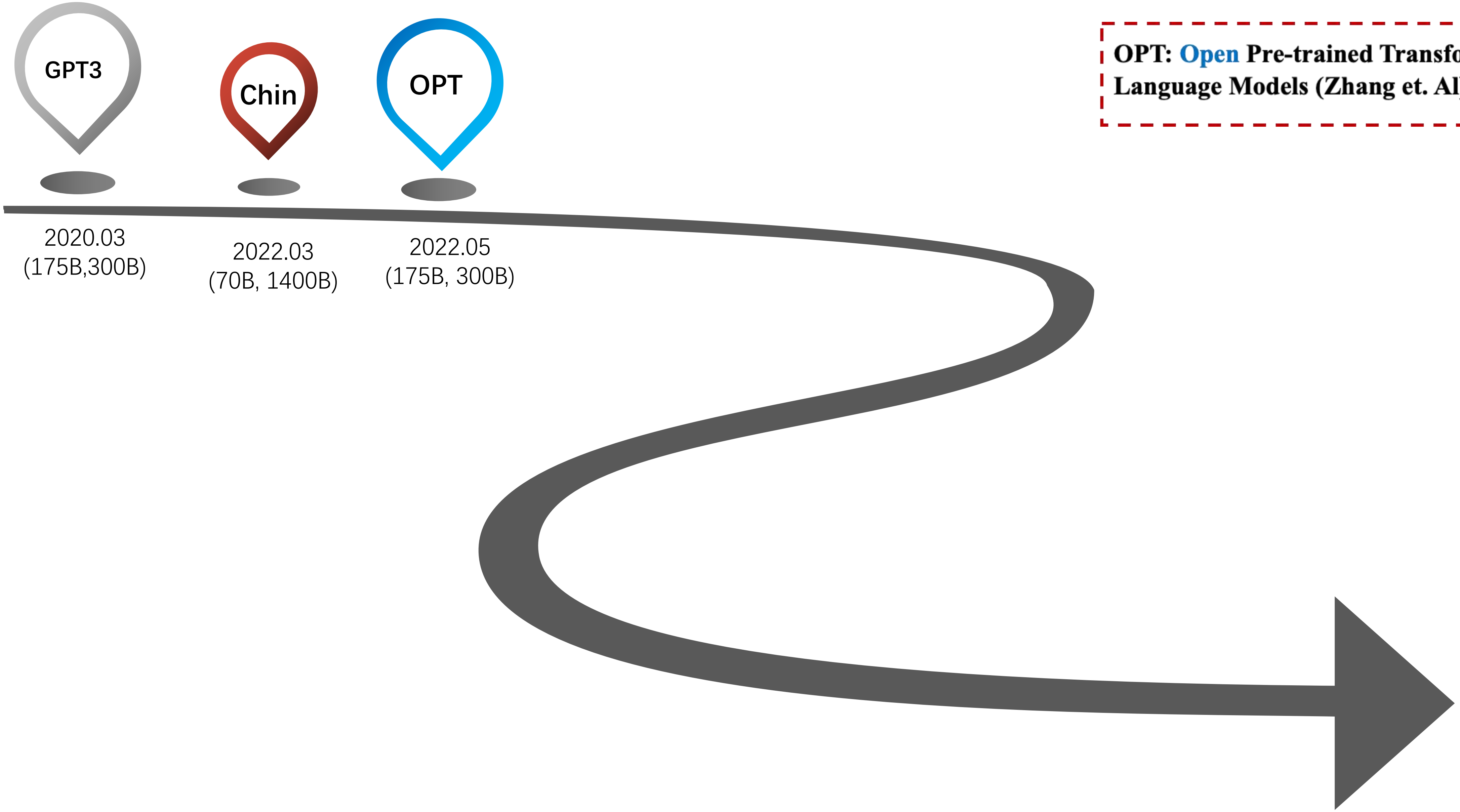
GPT3

2020.03
(175B,300B)

Chin

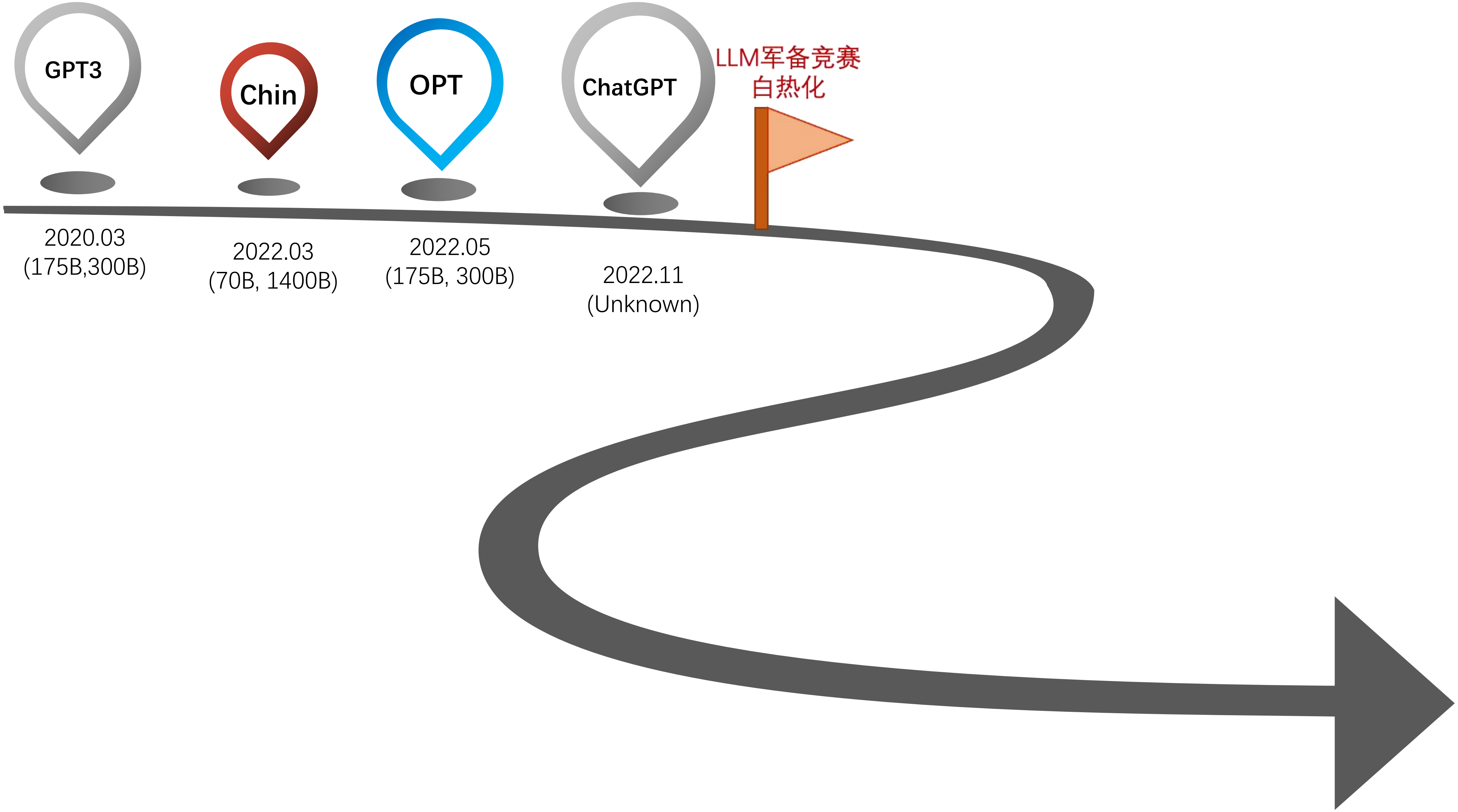
2022.03
(70B, 1400B)

# Llama家族



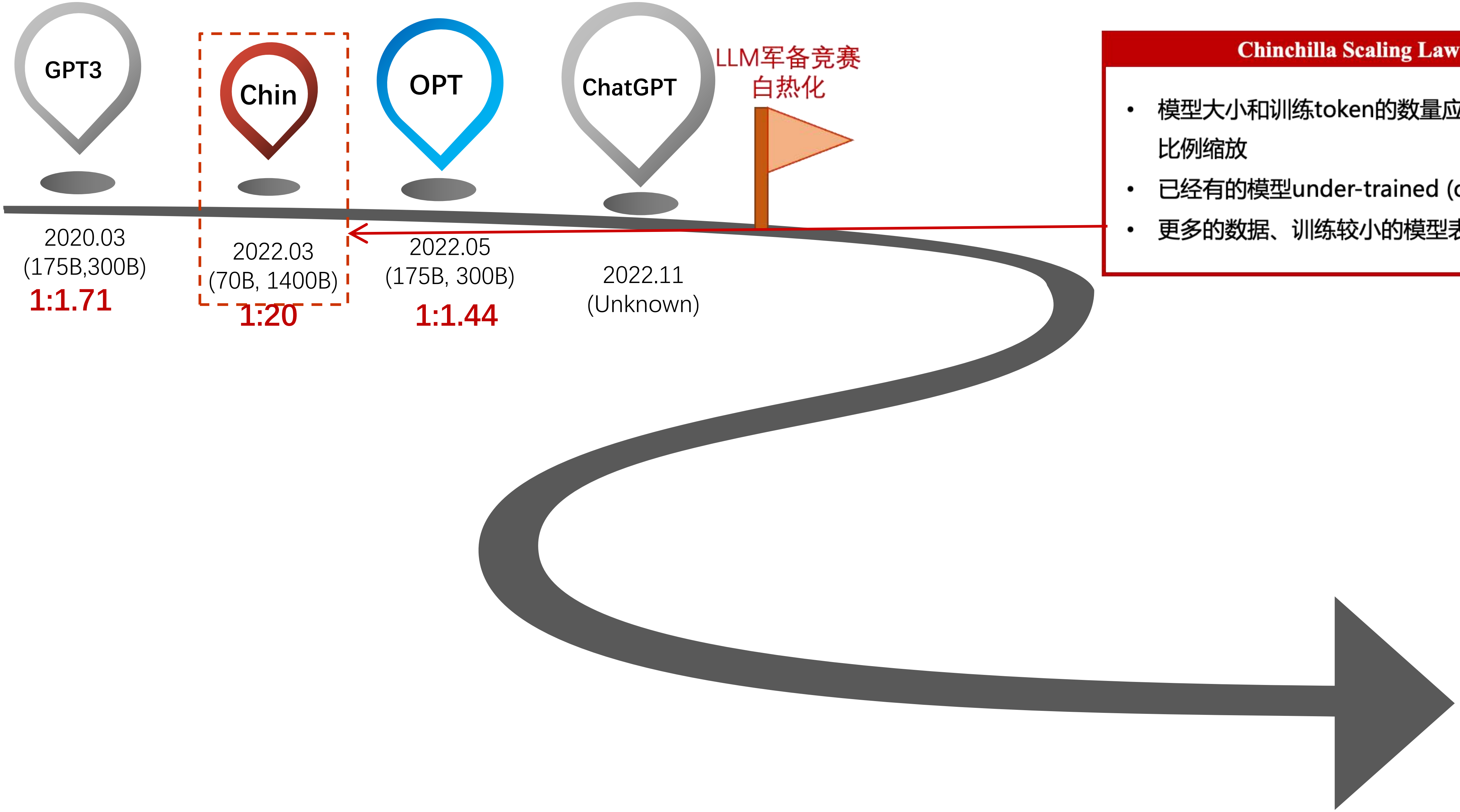
**OPT: Open Pre-trained Transformer Language Models (Zhang et. Al)**

# Llama家族



GPT3
2020.03
(175B,300B)
Chin
2022.03
(70B, 1400B)
OPT
2022.05
(175B, 300B)
ChatGPT
2022.11
(Unknown)
LLM军备竞赛
白热化

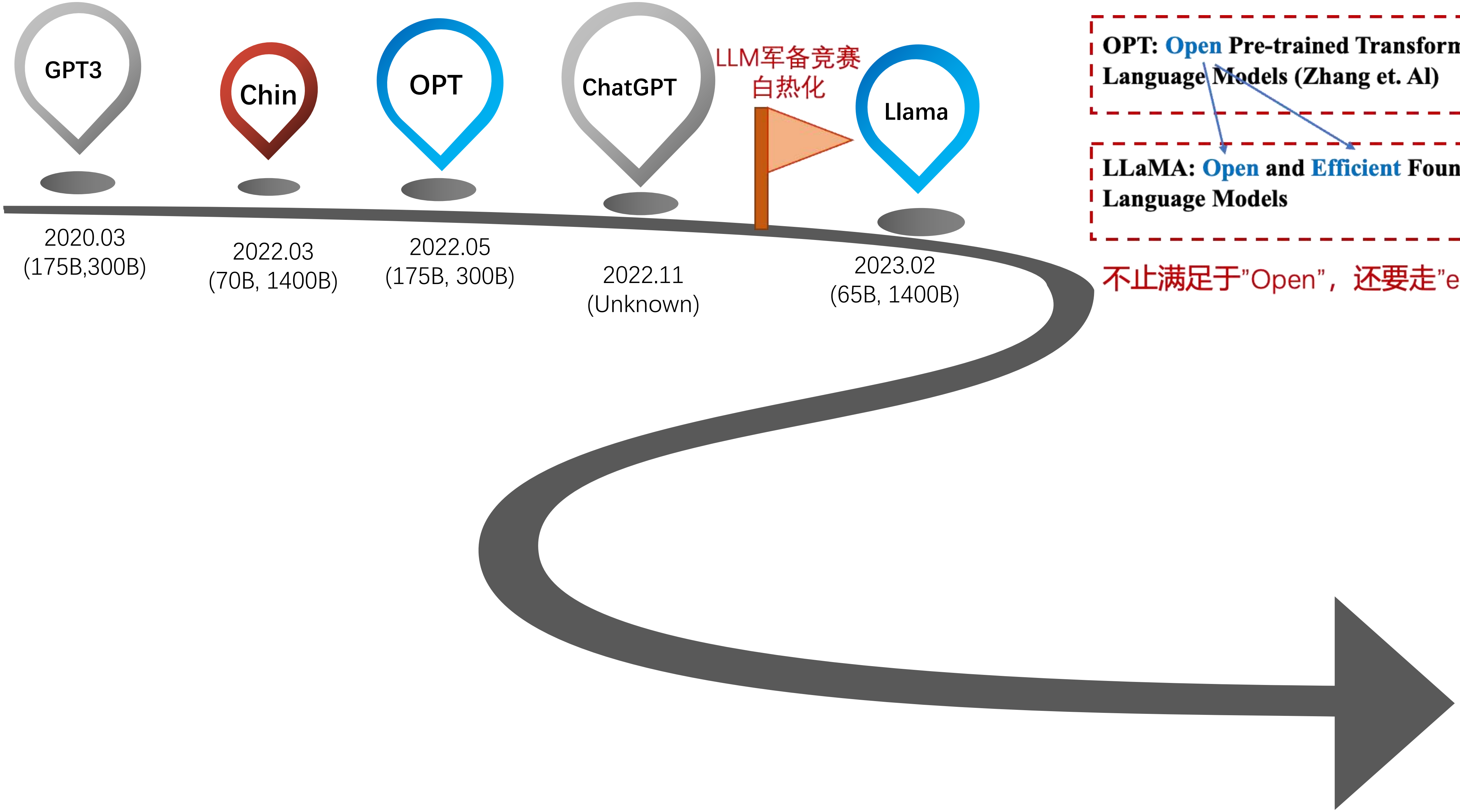
# Llama家族



**Chinchilla Scaling Law**

- 模型大小和训练token的数量应该按相等比例缩放
- 已经有的模型under-trained (over-sized)
- 更多的数据、训练较小的模型表现更好

# Llama 3 诞生背景

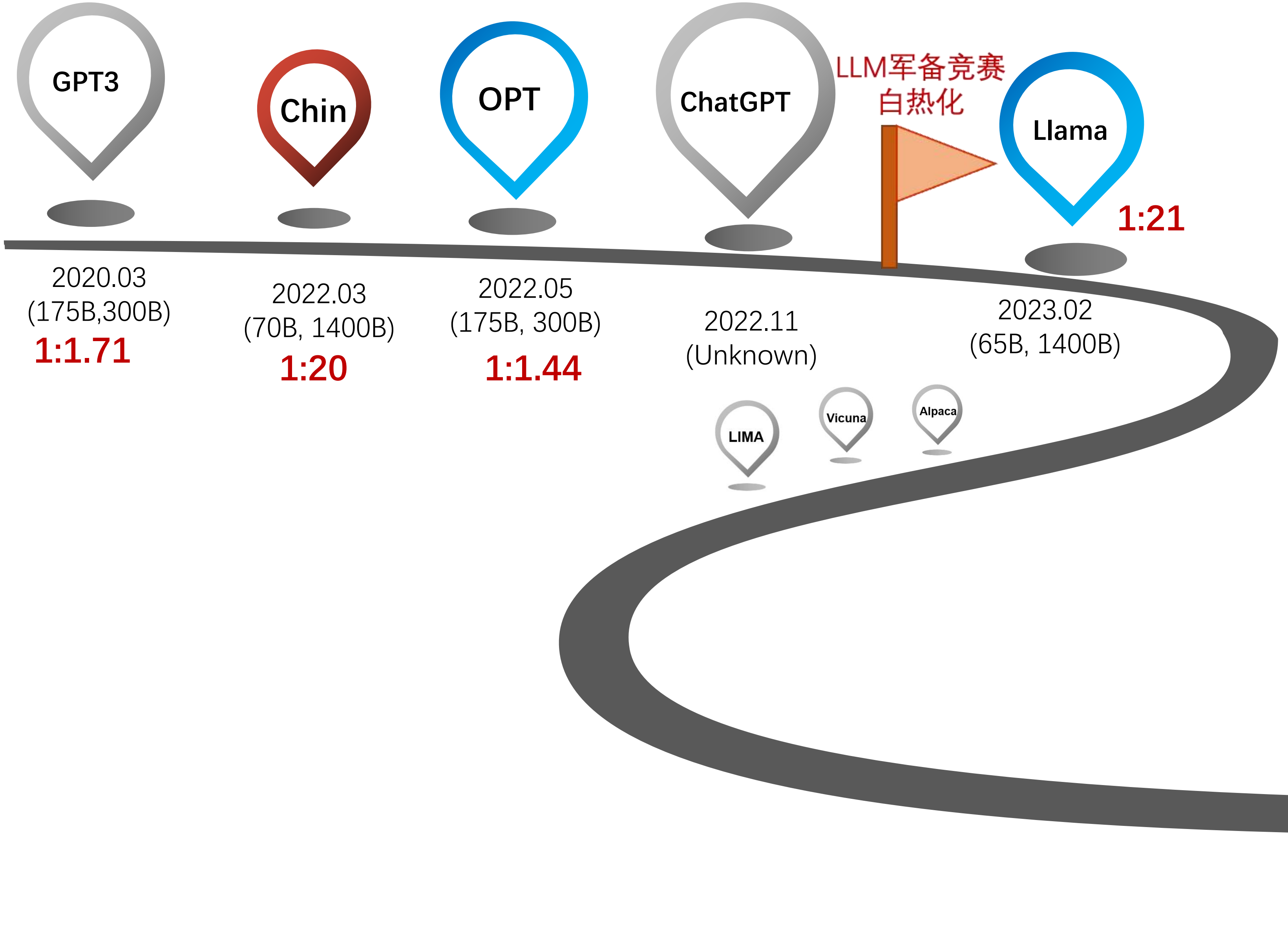


**OPT: Open Pre-trained Transformer Language Models (Zhang et. Al)**

**LLaMA: Open and Efficient Foundation Language Models**

不止满足于"Open"，还要走"efficient"

# Llama家族



GPT3

2020.03
(175B,300B)
1:1.71

Chin

2022.03
(70B, 1400B)
1:20

OPT

2022.05
(175B, 300B)
1:1.44

ChatGPT

2022.11
(Unknown)

LLM军备竞赛
白热化

Llama

1:21

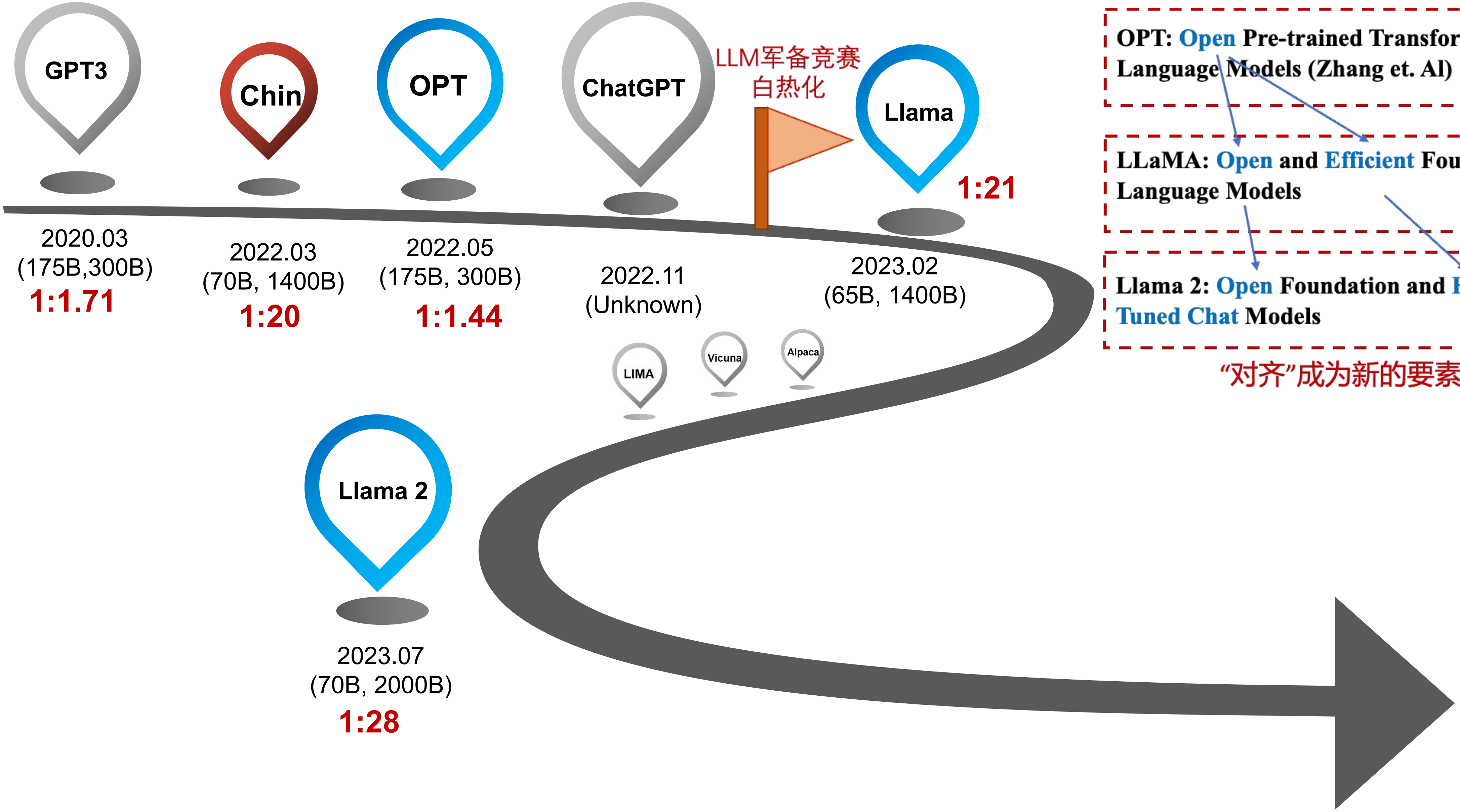
2023.02
(65B, 1400B)

LIMA

Vicuna

Alpaca

# Llama家族



GPT3

2020.03
(175B,300B)
1:1.71

Chin

2022.03
(70B, 1400B)
1:20

OPT

2022.05
(175B, 300B)
1:1.44

ChatGPT

2022.11
(Unknown)

LLM军备竞赛
白热化

Llama

1:21

2023.02
(65B, 1400B)

LIMA

Vicuna

Alpaca

Llama 2

2023.07
(70B, 2000B)
1:28

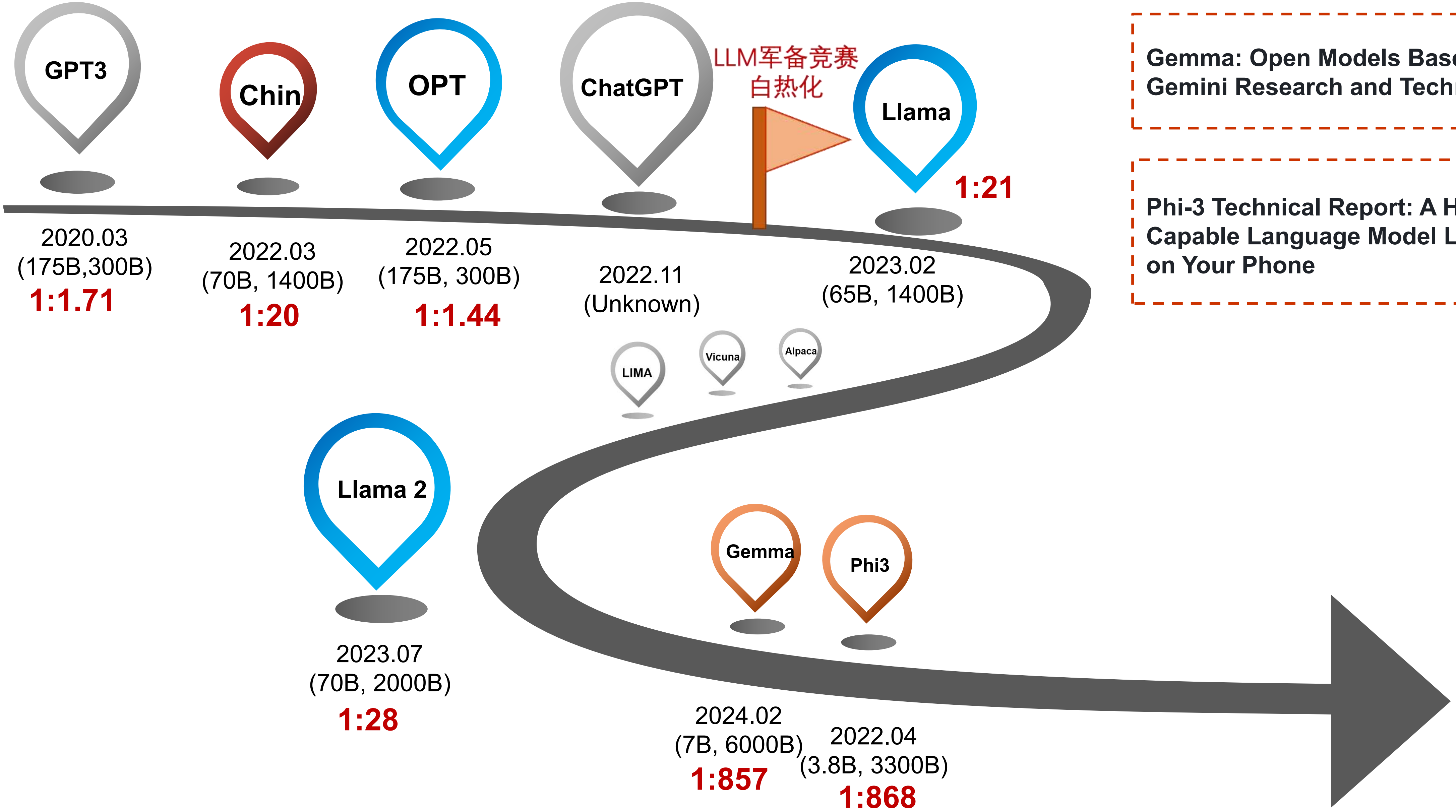
**OPT: Open Pre-trained Transformer Language Models (Zhang et. Al)**

**LLaMA: Open and Efficient Foundation Language Models**

**Llama 2: Open Foundation and Fine-Tuned Chat Models**

“对齐”成为新的要素

# Llama家族



GPT3
2020.03
(175B,300B)
1:1.71

Chin
2022.03
(70B, 1400B)
1:20

OPT
2022.05
(175B, 300B)
1:1.44

ChatGPT
2022.11
(Unknown)

LLM军备竞赛
白热化

Llama
1:21
2023.02
(65B, 1400B)

LIMA

Vicuna

Alpaca

Llama 2
2023.07
(70B, 2000B)
1:28

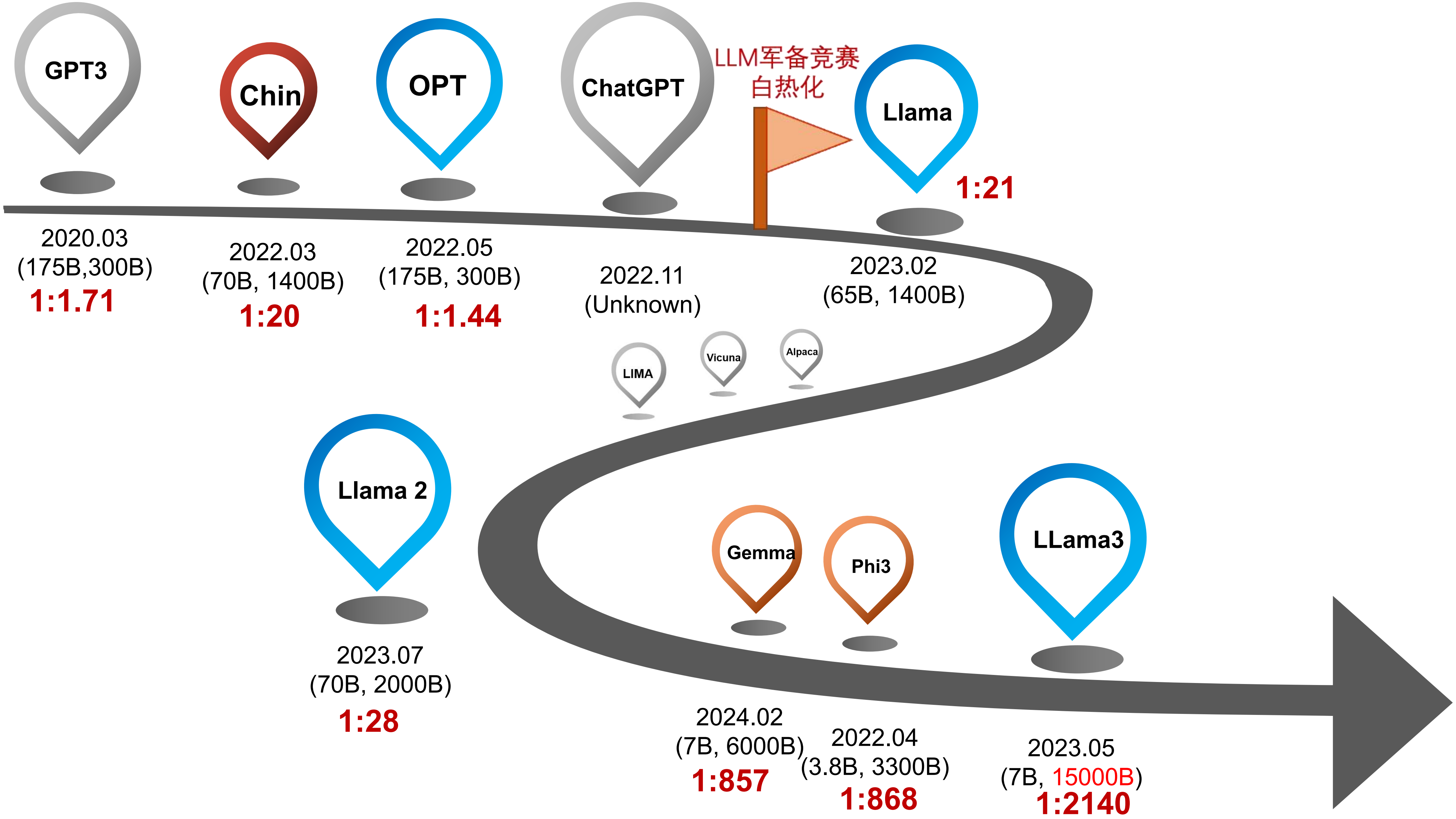
Gemma
2024.02
(7B, 6000B)
1:857

Phi3
2022.04
(3.8B, 3300B)
1:868

Gemma: Open Models Based on Gemini Research and Technology

Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone

# Llama家族



GPT3
2020.03
(175B,300B)
**1:1.71**

Chin
2022.03
(70B, 1400B)
**1:20**

OPT
2022.05
(175B, 300B)
**1:1.44**

ChatGPT
2022.11
(Unknown)

LLM军备竞赛
白热化

Llama
**1:21**
2023.02
(65B, 1400B)

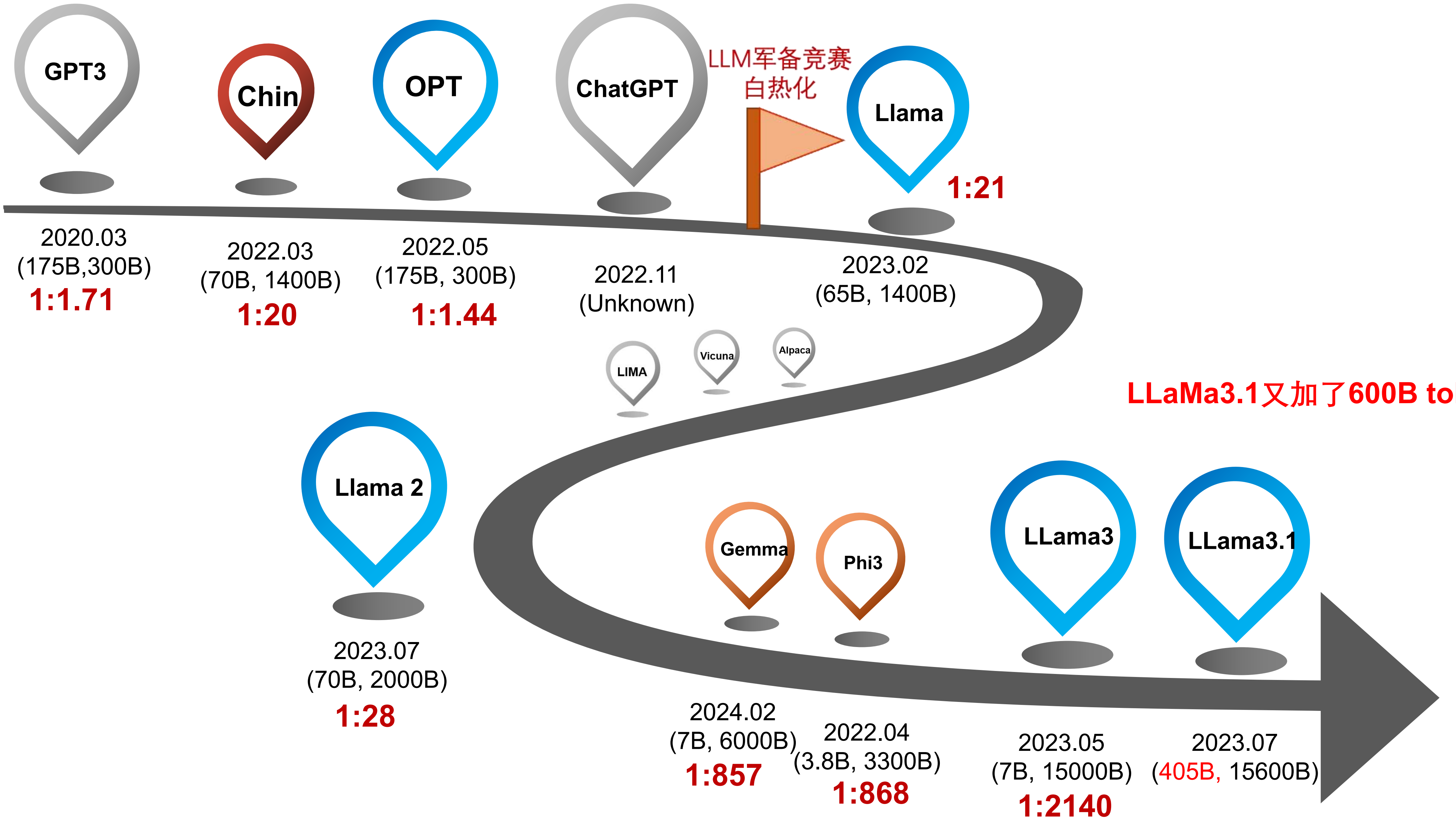
Llama 2
2023.07
(70B, 2000B)
**1:28**

Gemma
2024.02
(7B, 6000B)
**1:857**

Phi3
2022.04
(3.8B, 3300B)
**1:868**

LLama3
2023.05
(7B, 15000B)
**1:2140**

# Llama家族



GPT3

2020.03
(175B,300B)
1:1.71

Chin

2022.03
(70B, 1400B)
1:20

OPT

2022.05
(175B, 300B)
1:1.44

ChatGPT

2022.11
(Unknown)

LLM军备竞赛
白热化

Llama
1:21

2023.02
(65B, 1400B)

Llama 2

2023.07
(70B, 2000B)
1:28

Gemma

2024.02
(7B, 6000B)
1:857

Phi3

2022.04
(3.8B, 3300B)
1:868

LLama3

2023.05
(7B, 15000B)
1:2140

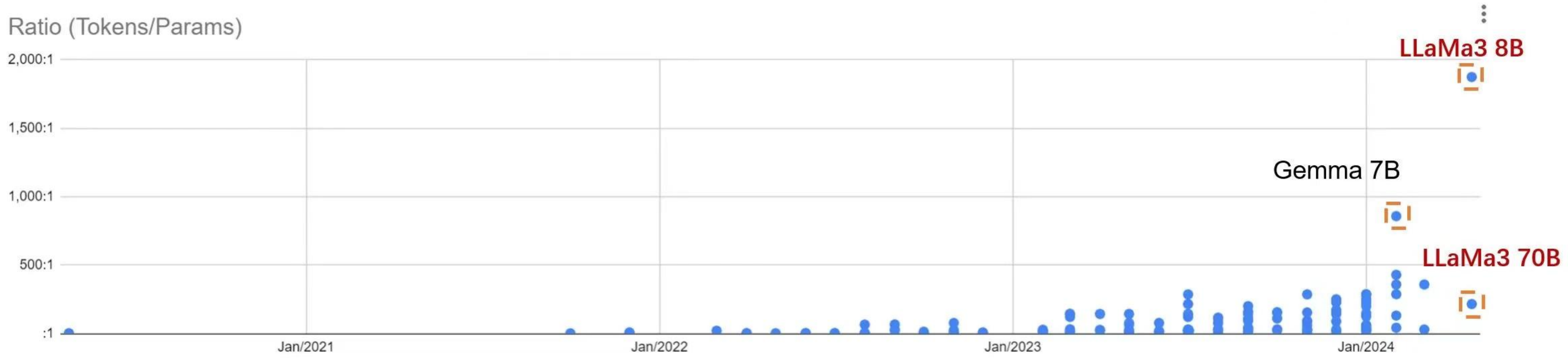
LLama3.1

2023.07
(405B, 15600B)

**LLaMa3.1又加了600B tokens**

# “Chinchilla Scaling Law”视角下的模型发展规律

## Chinchilla Scaling Law 一直被打破



(Chinchilla scaling: Tokens/Params≥20:1)

Gemma 7B: 857:1

Phi3: 868:1

LLaMa3 8B: 2140:1

# Llama 3：预训练

- □ **概述**
  - ■ 初始预训练
    - □ 逐步增加batchsize：4M->8M->16M
    - □ 数据混合：增加非英语数据与数学数据
  - ■ 长上下文预训练
    - □ 逐步增加上下文长度以训练直到模型在成功适应新上下文长度
  - ■ 退火阶段
    - □ 逐步成为常规的预训练阶段
    - □ 总计40M tokens
    - □ 学习率线性递减到0
    - □ 提高高质量数据占比
    - □ 计算模型断点的平均值作为最终模型

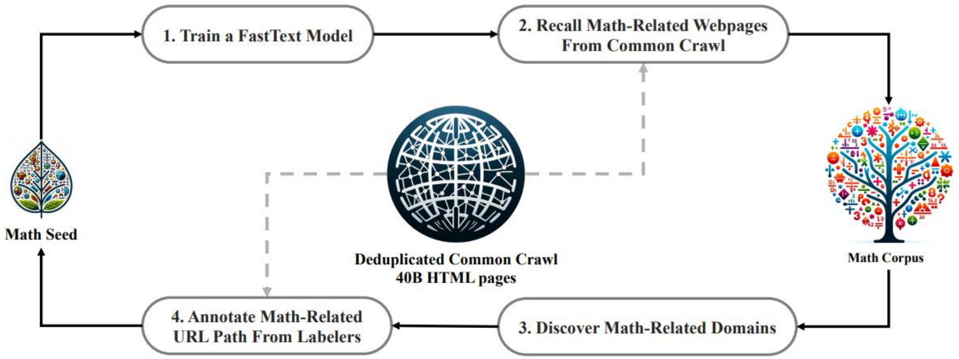
# Llama 3：预训练

## 数据量

| | Training Data | Params | Tokens | ChinLaw |
|---|---|---|---|---|
| LLaMA1 | CommonCrawl<br>C4, Github Wikipedia<br>Books<br>Arxiv<br>StackExchange | 7B | 1000B | 140B |
| | | 13B | 1000B | 260B |
| | | 33B | 1400B | 660B |
| | | 65B | 1400B | 1300B |
| LLaMA2 | A new mix of publicly available online data | 7B | 2000B | 140B |
| | | 13B | 2000B | 260B |
| | | 34B | 2000B | 680B |
| | | 70B | 2000B | 1400B |
| LLaMA3 | A new mix of publicly available online data | 8B | 15600B | 160B |
| | | 70B | 15600B | 1400B |
| | | 405B | 15600B | 8100B |

# Llama 3：预训练

- □ 数据处理
  - 文本提取与清洗
    - □ HTML解析器：从HTML中提取高质量文本
    - □ MarkDown：去除Markdown，对在web数据上训练的模型有害
  - 去重：URL、文档、行 多级别去重
  - 启发式过滤：通过n-gram，脏词过滤低质量文本
  - 基于模型过滤：基于Llama2训练质量分类器
  - 代码与推理：定制的数据提取&清洗pipeline
  - 多语言
    - □ fast text分类器对语言分类
    - □ 基于特定语言的过滤规则和过滤器


1. Train a FastText Model
2. Recall Math-Related Webpages From Common Crawl
Math Seed
Deduplicated Common Crawl 40B HTML pages
Math Corpus
4. Annotate Math-Related URL Path From Labelers
3. Discover Math-Related Domains

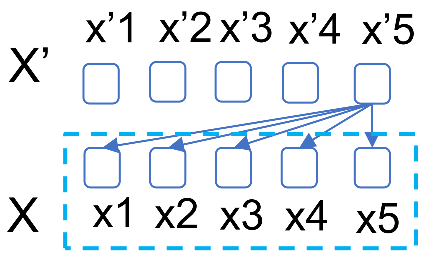
# Llama 3：预训练

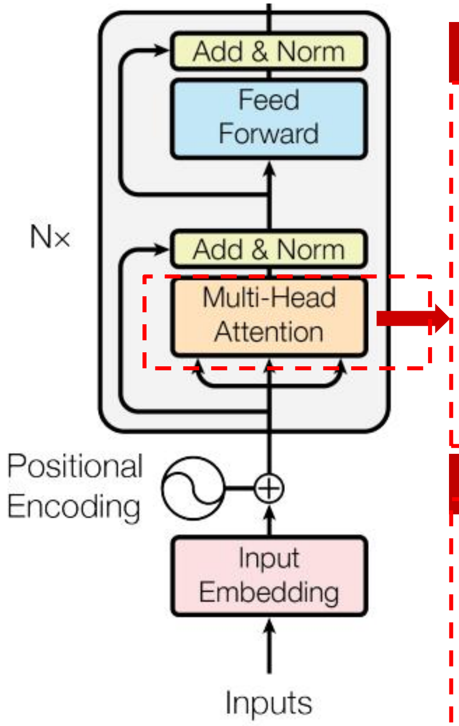
## □ 数据占比

- 知识分类：训练分类器以控制知识比例（如降采样、娱乐等领域）
- Scaling Laws for data mix: 在小模型上实验确定配比，应用到大模型上
- 数据比例：知识50、数学25、代码17、语言8

# Llama 3：预训练

- 训练架构 &细节
  - **Standard transformer architecture**





**Single Head**

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

**Multi Head**

$$Q_i = QW_i^Q, \quad K_i = KW_i^K, \quad V_i = VW_i^V \quad \text{for } i \in \{1, \ldots, h\}$$

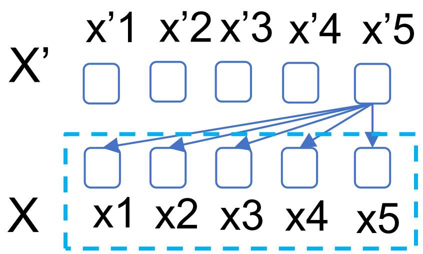
$$\mathrm{Attention}_i(Q_i, K_i, V_i) = \mathrm{softmax}\left(\frac{Q_iK_i^T}{\sqrt{d_k}}\right)V_i$$

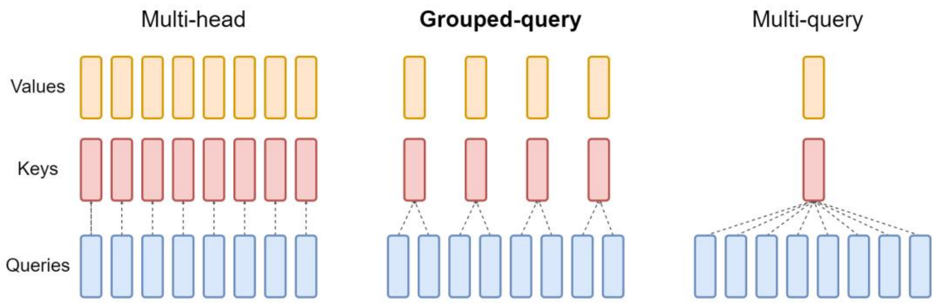
$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_h)W^O$$

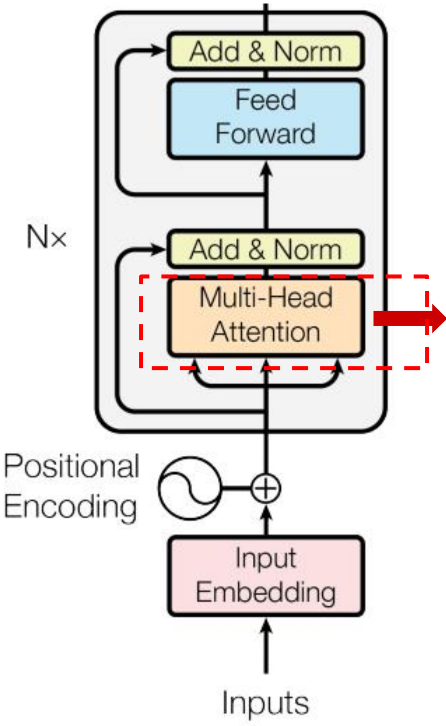
Attention Is All You Need, Ashish et al 2017

# Llama 3：预训练

- 训练架构 &细节
  - Standard transformer architecture
  - **Grouped-query attention**







**Grouped-query Attention**

$$Q = [Q_1, Q_2, \ldots, Q_G]$$

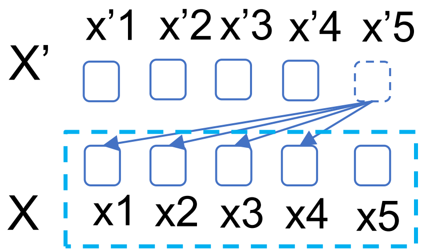
$$\text{Attention}(Q_i, K, V) = \text{softmax}\left(\frac{Q_i K^T}{\sqrt{d_k}}\right) V$$

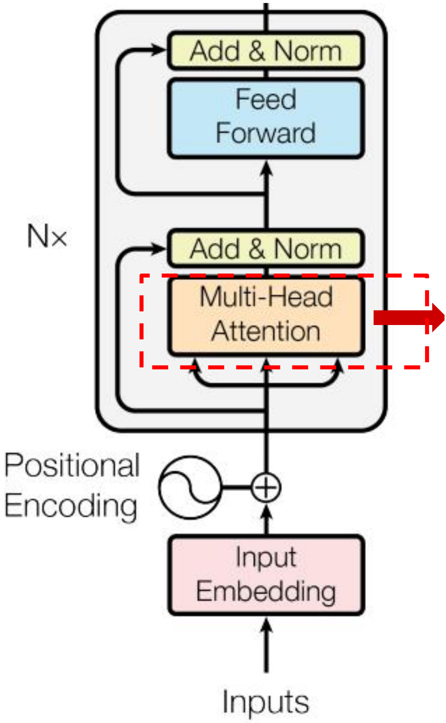
$$\text{GQA}(Q, K, V) = [\text{Attention}(Q_1, K, V), \text{Attention}(Q_2, K, V), \ldots, \text{Attention}(Q_G, K, V)]$$

GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints (Ainslie et al 2023)

# Llama 3：预训练

- 训练架构 &细节
  - Standard transformer architecture
  - Grouped-query attention
  - **KV-cache**





KV-cache

$$K^{(t)}_{\text{cache}} = \left[K^{(t-1)}_{\text{cache}}, K^{(t)}\right]$$

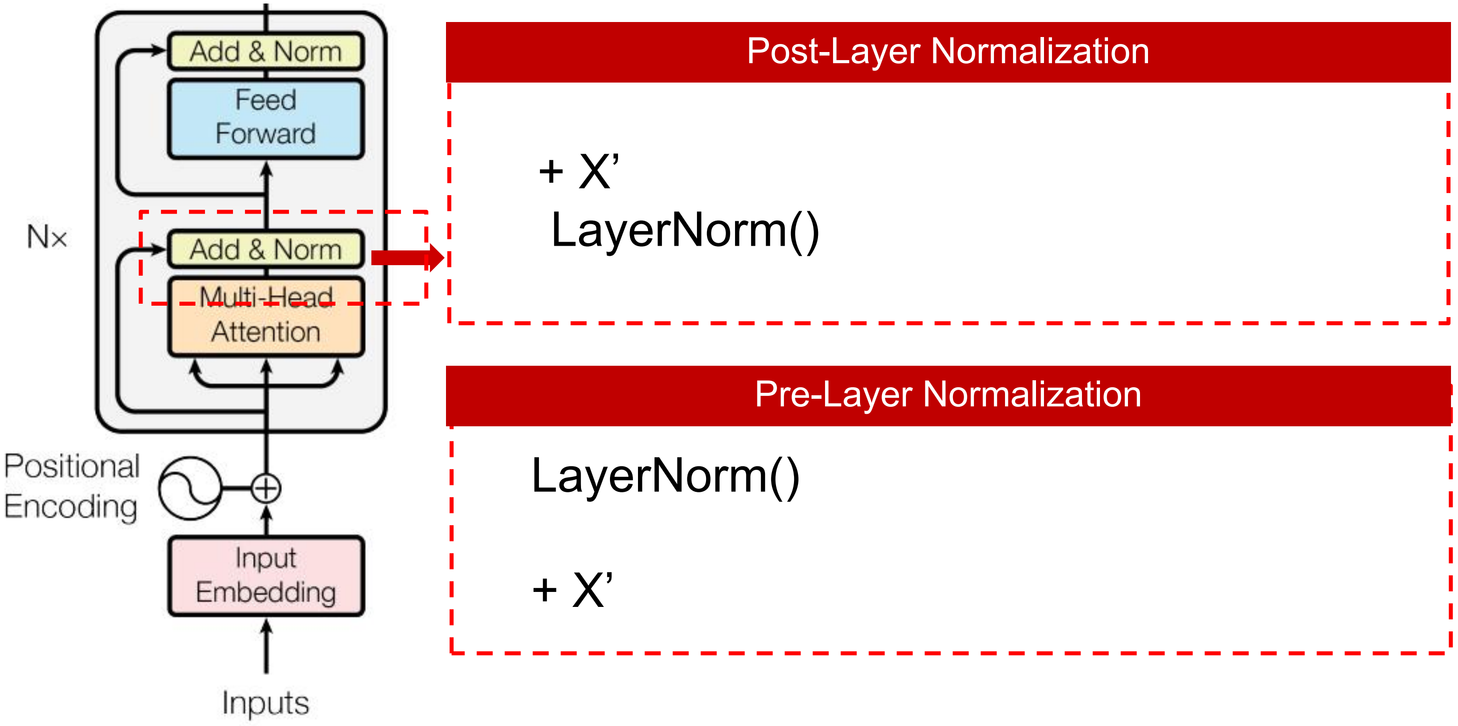
$$V^{(t)}_{\text{cache}} = \left[V^{(t-1)}_{\text{cache}}, V^{(t)}\right]$$

$$\text{Attention}(Q^{(t)}, K^{(t)}_{\text{cache}}, V^{(t)}_{\text{cache}}) = \text{softmax}\left(\frac{Q^{(t)}(K^{(t)}_{\text{cache}})^T}{\sqrt{d_k}}\right) V^{(t)}_{\text{cache}}$$

CacheGen: KV Cache Compression and Streaming for Fast Large Language Model Serving

# Llama 3：预训练

- 训练架构 &细节
  - Standard transformer architecture
  - Grouped-query attention
  - KV-cache
  - **Pre-normalization using RMSNorm**



Post-Layer Normalization

+ X’
LayerNorm()

Pre-Layer Normalization

LayerNorm()

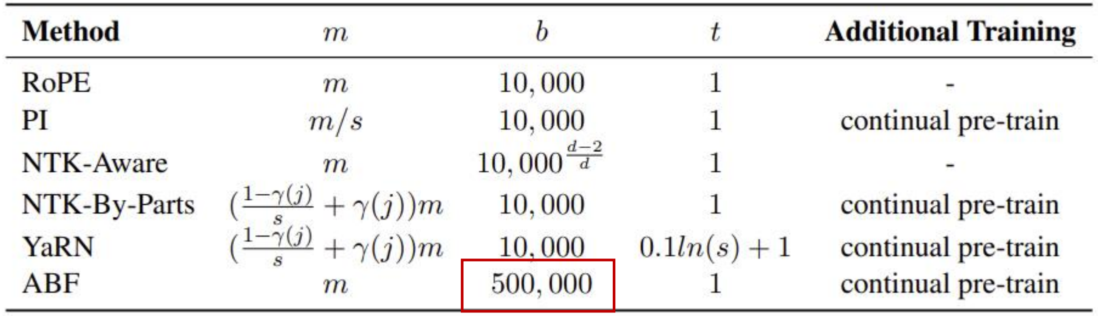
+ X’

# Llama 3：预训练

## 长文本支持

| Llama3 | Llama2 | Llama1 |
|---|---|---|
| 128K | 4K | 2K |

- 长文本预训练：逐渐增加上下文长度
- RoPE base frequency增加到500,000：更好的支持长文本
- 长文本训练阶段使用attention mask



| **Method** | $m$ | $b$ | $t$ | **Additional Training** |
|---|---|---|---|---|
| RoPE | $m$ | $10,000$ | 1 | - |
| PI | $m/s$ | $10,000$ | 1 | continual pre-train |
| NTK-Aware | $m$ | $10,000^{\frac{d-2}{d}}$ | 1 | - |
| NTK-By-Parts | $(\frac{1-\gamma(j)}{s}+\gamma(j))m$ | $10,000$ | 1 | continual pre-train |
| YaRN | $(\frac{1-\gamma(j)}{s}+\gamma(j))m$ | $10,000$ | $0.1ln(s)+1$ | continual pre-train |
| ABF | $m$ | $\boxed{500,000}$ | 1 | continual pre-train |

Extending LLMs' Context Window with 100 Samples, Zhang et al.2023

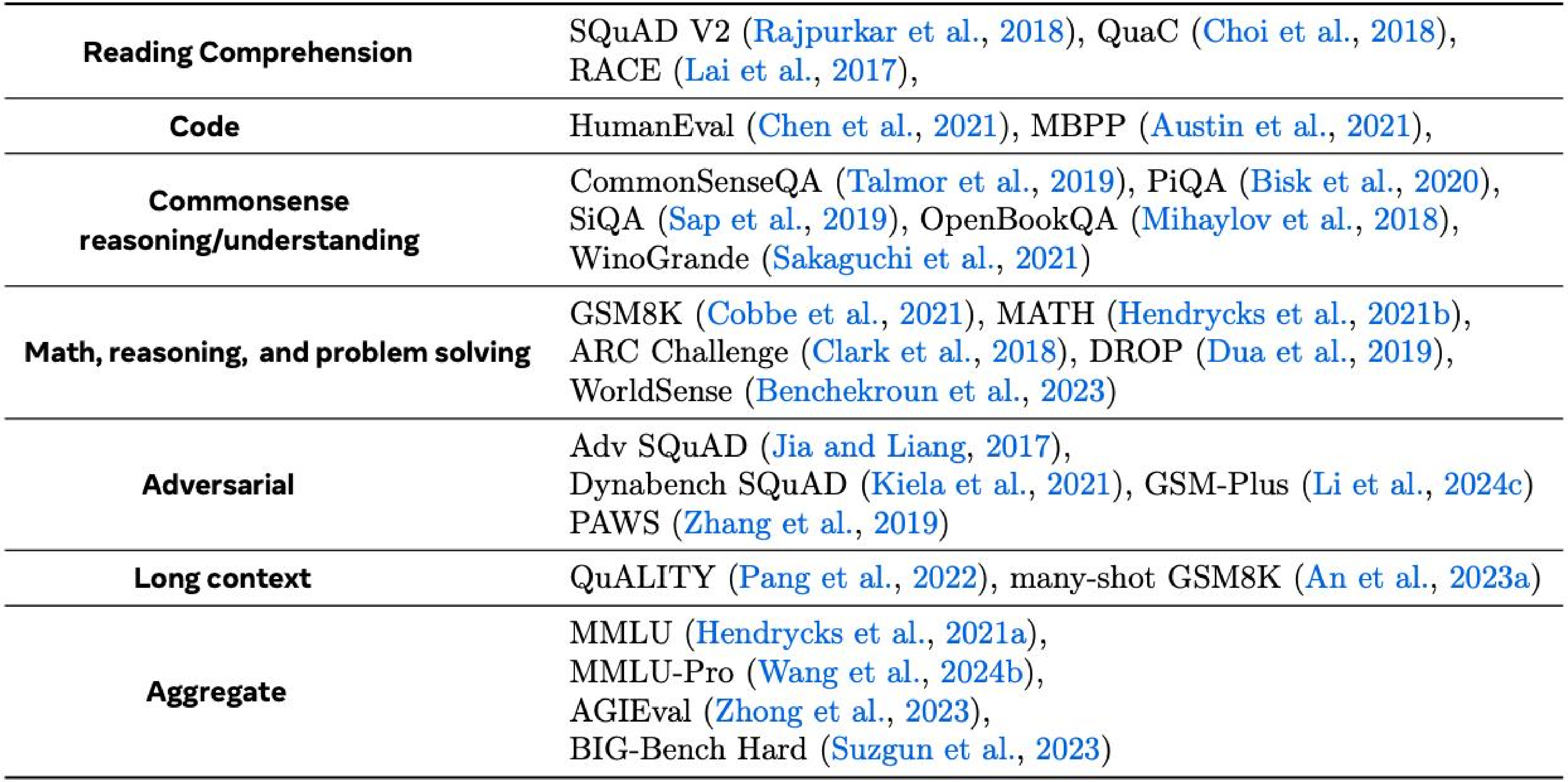
# Llama 3：预训练

## 训练时间

| GPU Hours | Llama3(H100) | Llama2(A100) |
|---|---|---|
| 7B / 8B | 1.46M | 0.18M |
| 70B | 7.0M | 1.72M |
| 405B | 30.84M | / |

# Llama 3：预训练 – 评估

## Bechmark



| | |
|---|---|
| **Reading Comprehension** | SQuAD V2 (Rajpurkar et al., 2018), QuaC (Choi et al., 2018), RACE (Lai et al., 2017), |
| **Code** | HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), |
| **Commonsense reasoning/understanding** | CommonSenseQA (Talmor et al., 2019), PiQA (Bisk et al., 2020), SiQA (Sap et al., 2019), OpenBookQA (Mihaylov et al., 2018), WinoGrande (Sakaguchi et al., 2021) |
| **Math, reasoning, and problem solving** | GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b), ARC Challenge (Clark et al., 2018), DROP (Dua et al., 2019), WorldSense (Benchekroun et al., 2023) |
| **Adversarial** | Adv SQuAD (Jia and Liang, 2017), Dynabench SQuAD (Kiela et al., 2021), GSM-Plus (Li et al., 2024c) PAWS (Zhang et al., 2019) |
| **Long context** | QuALITY (Pang et al., 2022), many-shot GSM8K (An et al., 2023a) |
| **Aggregate** | MMLU (Hendrycks et al., 2021a), MMLU-Pro (Wang et al., 2024b), AGIEval (Zhong et al., 2023), BIG-Bench Hard (Suzgun et al., 2023) |

# Llama 3：数据实例

## □ Commonsense Reasoning - HellaSwag

| | |
|---|---|
| **context** | A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She… |
| **endings** | A. rinses the bucket off with soap and blow dry the dog’s head.<br>B. uses a hose to keep it from getting soapy.<br>C. gets the dog wet, then it runs away again.<br>D. gets into a bath tub with the dog. |
| **label** | C |

# Llama 3：数据实例

## □ Math Problems - GSM8K

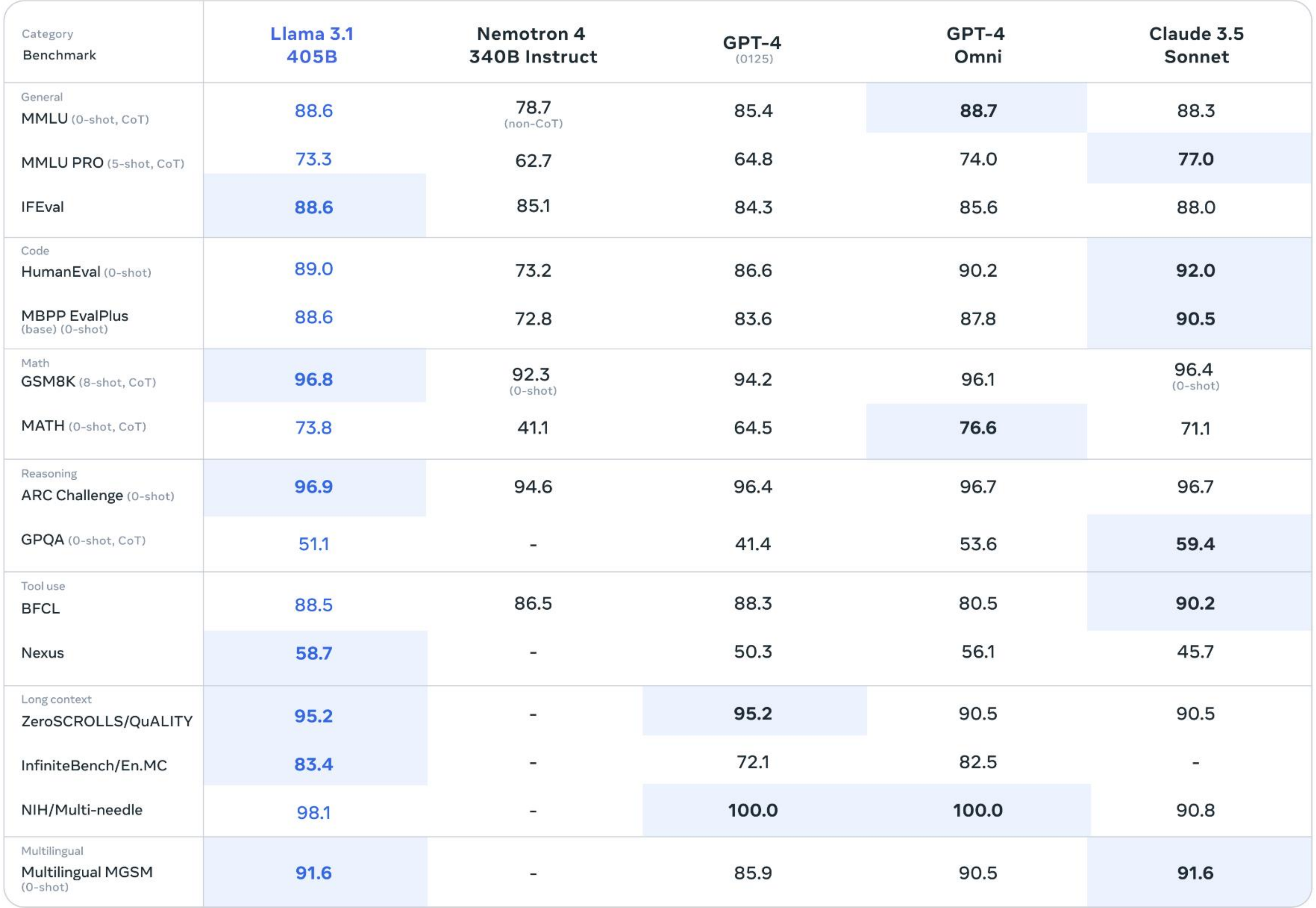
**question**

Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

**answer**

To find out how much Janet makes at the farmers' market, we need to first determine how many eggs are left after she eats some for breakfast and bakes some for her friends. Janet eats 3 eggs for breakfast and bakes 4 for her friends, so she uses a total of 3 + 4 = 7 eggs. Since her ducks lay 16 eggs per day, the number of eggs left is 16 - 7 = 9. She sells these 9 eggs for $2 each, so her daily earnings are 9 x $2 = $18.

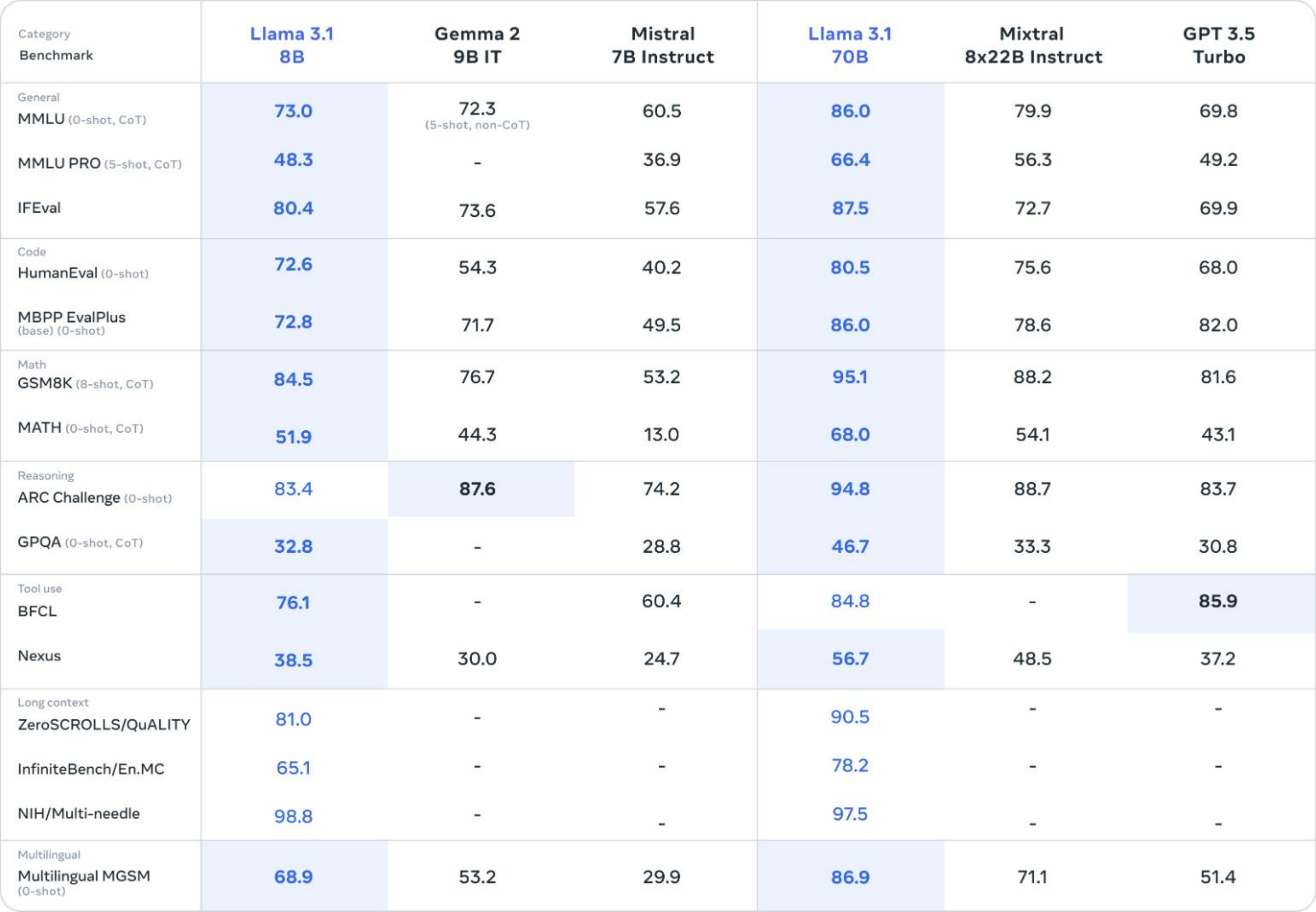
# Llama 3：预训练 – 评估



| Category<br>Benchmark | Llama 3.1<br>405B | Nemotron 4<br>340B Instruct | GPT-4<br>(0125) | GPT-4<br>Omni | Claude 3.5<br>Sonnet |
|---|---|---|---|---|---|
| General<br>MMLU (0-shot, CoT) | 88.6 | 78.7<br>(non-CoT) | 85.4 | **88.7** | 88.3 |
| MMLU PRO (5-shot, CoT) | 73.3 | 62.7 | 64.8 | 74.0 | **77.0** |
| IFEval | **88.6** | 85.1 | 84.3 | 85.6 | 88.0 |
| Code<br>HumanEval (0-shot) | 89.0 | 73.2 | 86.6 | 90.2 | **92.0** |
| MBPP EvalPlus<br>(base) (0-shot) | 88.6 | 72.8 | 83.6 | 87.8 | **90.5** |
| Math<br>GSM8K (8-shot, CoT) | **96.8** | 92.3<br>(0-shot) | 94.2 | 96.1 | 96.4<br>(0-shot) |
| MATH (0-shot, CoT) | 73.8 | 41.1 | 64.5 | **76.6** | 71.1 |
| Reasoning<br>ARC Challenge (0-shot) | **96.9** | 94.6 | 96.4 | 96.7 | 96.7 |
| GPQA (0-shot, CoT) | 51.1 | - | 41.4 | 53.6 | **59.4** |
| Tool use<br>BFCL | 88.5 | 86.5 | 88.3 | 80.5 | **90.2** |
| Nexus | **58.7** | - | 50.3 | 56.1 | 45.7 |
| Long context<br>ZeroSCROLLS/QuALITY | **95.2** | - | **95.2** | 90.5 | 90.5 |
| InfiniteBench/En.MC | **83.4** | - | 72.1 | 82.5 | - |
| NIH/Multi-needle | 98.1 | - | **100.0** | **100.0** | 90.8 |
| Multilingual<br>Multilingual MGSM<br>(0-shot) | **91.6** | - | 85.9 | 90.5 | **91.6** |

Llama 3 405B在benchmark与GPT-4相近/超过

# Llama 3：预训练 – 评估



| Category<br>Benchmark | Llama 3.1<br>8B | Gemma 2<br>9B IT | Mistral<br>7B Instruct | Llama 3.1<br>70B | Mixtral<br>8x22B Instruct | GPT 3.5<br>Turbo |
|---|---|---|---|---|---|---|
| General<br>MMLU (0-shot, CoT) | 73.0 | 72.3<br>(5-shot, non-CoT) | 60.5 | 86.0 | 79.9 | 69.8 |
| MMLU PRO (5-shot, CoT) | 48.3 | - | 36.9 | 66.4 | 56.3 | 49.2 |
| IFEval | 80.4 | 73.6 | 57.6 | 87.5 | 72.7 | 69.9 |
| Code<br>HumanEval (0-shot) | 72.6 | 54.3 | 40.2 | 80.5 | 75.6 | 68.0 |
| MBPP EvalPlus<br>(base) (0-shot) | 72.8 | 71.7 | 49.5 | 86.0 | 78.6 | 82.0 |
| Math<br>GSM8K (8-shot, CoT) | 84.5 | 76.7 | 53.2 | 95.1 | 88.2 | 81.6 |
| MATH (0-shot, CoT) | 51.9 | 44.3 | 13.0 | 68.0 | 54.1 | 43.1 |
| Reasoning<br>ARC Challenge (0-shot) | 83.4 | 87.6 | 74.2 | 94.8 | 88.7 | 83.7 |
| GPQA (0-shot, CoT) | 32.8 | - | 28.8 | 46.7 | 33.3 | 30.8 |
| Tool use<br>BFCL | 76.1 | - | 60.4 | 84.8 | - | 85.9 |
| Nexus | 38.5 | 30.0 | 24.7 | 56.7 | 48.5 | 37.2 |
| Long context<br>ZeroSCROLLS/QuALITY | 81.0 | - | - | 90.5 | - | - |
| InfiniteBench/En.MC | 65.1 | - | - | 78.2 | - | - |
| NIH/Multi-needle | 98.8 | - | - | 97.5 | - | - |
| Multilingual<br>Multilingual MGSM<br>(0-shot) | 68.9 | 53.2 | 29.9 | 86.9 | 71.1 | 51.4 |

# 谢谢各位！