

开源 vs 闭源 大模型哪家强



ZOMI



LLAMA3.1 性能媲美GPT-4和Claude 3.5

- 2024.7.23, Meta发布LLAMA 3.1 开源大模型。根据Meta提供基准测试, 405B (4050亿参数), 从性能上已经可媲美GPT-4和Claude 3.5, 标志着开源大模型发展再进一步。

Category	Benchmark	Llama 3 8B	Gemma 2 9B	Mistral 7B	Llama 3 70B	Mixtral 8x22B	GPT 3.5 Turbo	Llama 3 405B	Nemotron 4 340B	GPT-4 ^(0-shot)	GPT-4o	Claude 3.5 Sonnet
General	MMLU ^(5-shot)	69.4	72.3	61.1	83.6	76.9	70.7	87.3	82.6	85.1	89.1	89.9
	MMLU ^(0-shot, CoT)	73.0	72.3 [△]	60.5	86.0	79.9	69.8	88.6	78.7 [△]	85.4	88.7	88.3
	MMLU-Pro ^(5-shot, CoT)	48.3	—	36.9	66.4	56.3	49.2	73.3	62.7	64.8	74.0	77.0
	IFEval	80.4	73.6	57.6	87.5	72.7	69.9	88.6	85.1	84.3	85.6	88.0
Code	HumanEval ^(0-shot)	72.6	54.3	40.2	80.5	75.6	68.0	89.0	73.2	86.6	90.2	92.0
	MBPP EvalPlus ^(0-shot)	72.8	71.7	49.5	86.0	78.6	82.0	88.6	72.8	83.6	87.8	90.5
Math	GSM8K ^(8-shot, CoT)	84.5	76.7	53.2	95.1	88.2	81.6	96.8	92.3 [◇]	94.2	96.1	96.4 [◇]
	MATH ^(0-shot, CoT)	51.9	44.3	13.0	68.0	54.1	43.1	73.8	41.1	64.5	76.6	71.1
Reasoning	ARC Challenge ^(0-shot)	83.4	87.6	74.2	94.8	88.7	83.7	96.9	94.6	96.4	96.7	96.7
	GPQA ^(0-shot, CoT)	32.8	—	28.8	46.7	33.3	30.8	51.1	—	41.4	53.6	59.4
Tool use	BFCL	76.1	—	60.4	84.8	—	85.9	88.5	86.5	88.3	80.5	90.2
	Nexus	38.5	30.0	24.7	56.7	48.5	37.2	58.7	—	50.3	56.1	45.7
Long context	ZeroSCROLLS/QuALITY	81.0	—	—	90.5	—	—	95.2	—	95.2	90.5	90.5
	InfiniteBench/En.MC	65.1	—	—	78.2	—	—	83.4	—	72.1	82.5	—
	NIH/Multi-needle	98.8	—	—	97.5	—	—	98.1	—	100.0	100.0	90.8
Multilingual	MGSM ^(0-shot, CoT)	68.9	53.2	29.9	86.9	71.1	51.4	91.6	—	85.9	90.5	91.6

Table 2 Performance of finetuned Llama 3 models on key benchmark evaluations. The table compares the performance of the 8B, 70B, and 405B versions of Llama 3 with that of competing models. We **boldface** the best-performing model in each of three model-size equivalence classes. [△]Results obtained using 5-shot prompting (no CoT). [△]Results obtained without CoT. [◇]Results obtained using zero-shot prompting.

关于本内容

1. 一起聊聊李校长~
2. 模型性能分析：开源大模型 vs 闭源大模型
3. 大模型降价潮：开源模型倒逼闭源模型
4. 国内大模型发展：开源LLAMA引导国内百花齐放
5. 大模型的技术发展：MOE？长序列？稠密万亿？为什么 meta 这么大方？

01. 李校长

我们被收智商税了吗？

WAIC 期间，李彦宏表示：开源其实是一种智商税



https://www.bilibili.com/video/BV1S1421b7Ae/?vd_source=26de035c60e6c7f810371fd13d14b6

2024世界人工智能大会暨人工智能全球治理高级别会议

2024 WORLD AI CONFERENCE & HIGH-LEVEL MEETING ON GLOBAL AI GOVERNANCE



李彦宏认为，开源其实是一种智商税

- 当你理性地去想，大模型能够带来什么价值，以什么样的成本带来价值的时候，就会发现，你永远应该选择闭源模型。
- 今天无论是 ChatGPT、还是文心一言等闭源模型，一定比开源模型更强大，推理成本更低。



李彦宏WAIC圆桌访谈：开源模型是智商税，智能体正在爆发

李彦宏认为，开源其实是一种智商税

- ToB 客户，他要选择的是一个性价比最好的模型。一方面，模型要对他的应用产生价值，另外一方面，成本要足够低。很多时候，你看着有用，一算成本不划算，客户就放弃了。这是为什么我刚才讲，开源模型打不过闭源模型。你只要理性的去看待，你的收益是啥，你的成本是啥，你就会发现，最好还是去选择闭源模型。当然，闭源模型不是一个模型，它是一系列的模型，根据你的使用场景去平衡一下，要多好的效果，要多快的推理速度，要多低的成本。模型有非常多的变种，可以根据用户的需求，让他来做选择。
- 闭源模型还有一个开源模型不具备的优势：这些相对来说规模更小的模型，都是从最大最powerful的模型裁剪出来的，裁剪出来这些更小规模的模型，仍然比那些同样规模的开源模型要效果更好。

李彦宏WAIC圆桌访谈：开源模型是智商税，智能体正在爆发

不是第一次了李校长

- 这并非李彦宏第一次公开场合发表类似观点，2024 年 4 月，也曾提出“开源模型会越来越落后”。
- 随后不久，360 集团董事长周鸿祎在发出反对声音称，如果没有开源文化，就不会有 Linux/PyTorch 等众多重要技术成果，甚至互联网发展也会受到限制。



Mark Zuckerberg

- Meta 创始人兼 CEO 马克·扎克伯格亲笔撰写了一篇题为《Open Source AI Is the Path Forward》长文，阐述开源对开发者、对 Meta 以及对全球的重要意义。预测到今年年底，Meta AI 将超过 ChatGPT，成为使用最广泛大模型 AI 助手。

Meta

Open Source AI Is the Path Forward

July 23, 2024

By Mark Zuckerberg, Founder and CEO

<https://about.fb.com/news/2024/07/open-source-ai-is-the-path-forward/>



After LLAMA3.1 Andrew Ng & Yann LeCun



Yann LeCun  
@ylecun

..

Some details on Llama-3.1 at The Verge.

"Meta releases the biggest and best open-source AI model yet: Llama 3.1 outperforms OpenAI and other rivals on certain benchmarks. Now, Mark Zuckerberg expects Meta's AI assistant to surpass ChatGPT's usage in the coming months."

theverge.com/2024/7/23/2420...



Andrew Ng 
@AndrewYNg

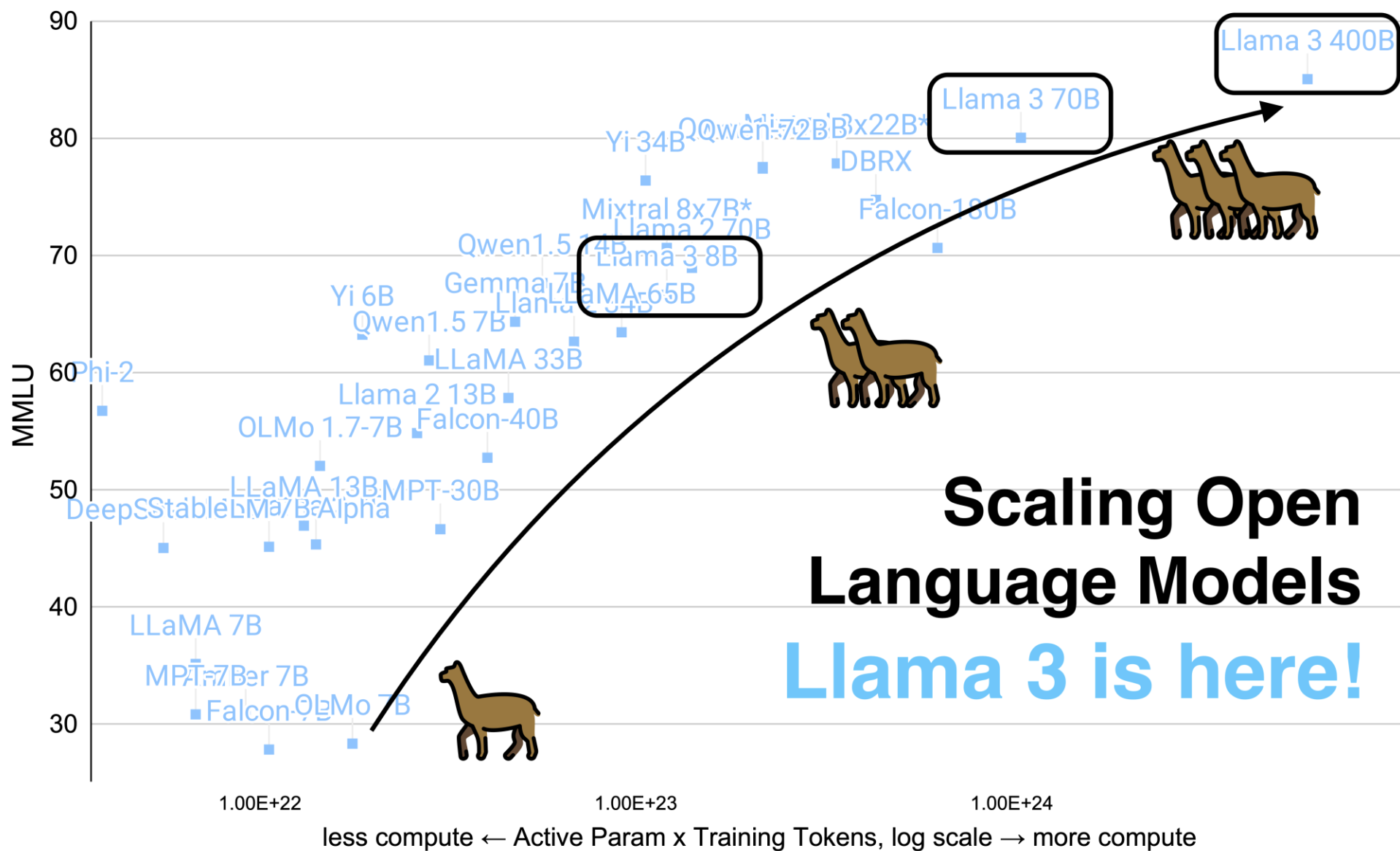
...

Thank you Meta and the Llama team for your huge contributions to open-source! Llama 3.1 with increased context length and improved capabilities is a wonderful gift to everyone.

I hope foolish regulations don't like California's proposed SB1047 don't stop such innovations.

02.

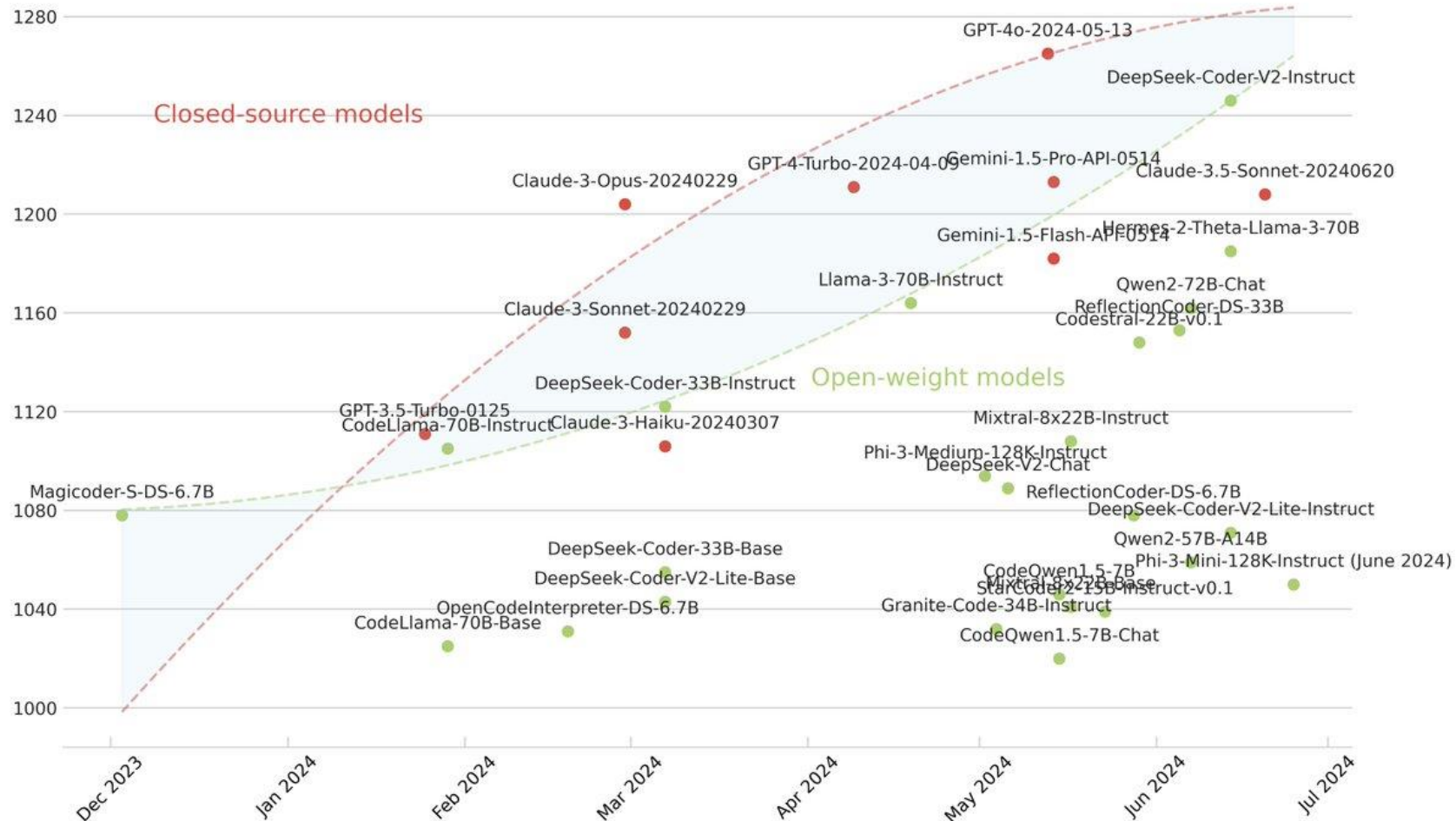
模型性能分析



Code LLMs: Closed-source vs. open-weight models

DeepSeek-Coder-V2-Instruct is a breakthrough in terms of code generation for open-weight models. @maximelabonne

BigCode ELO



Maxime Labonne @ ICML
@maximelabonne

New chart with Code LLMs

Based on the BigCodeBench Leaderboard, it shows the impressive performance of DeepSeek Coder V2 (236B, 21B active), which is on par with the best closed-source models.

Personally, I've switched to Sonnet and I'm very happy with its code performance.



Sunder Ali Khawaja, Ph.D., Senior Member IEEE • 2 度 关注
Researcher (Deep Learning, Responsible AI, Privacy Preservation M...
1 个月前 • 已编辑 •

Closed-source vs. Open-weight LLMs

The gap between closed-source and open-weight models is closing in terms of MMLU. LLMs are plateauing and the gap between closed vs. open is almost closed!

v/ Maxime Labonne <https://lnkd.in/exq9Hghh>

We now face a 6 to 10-month lag, rather than years as was the case when GPT-4 was released.

The release of GPT-4o is not a significant breakthrough with a score of 87.2% on the 5-shot MMLU (source: <https://lnkd.in/etUvuHD>).

On the other side, Meta reported that Llama 3 405B's checkpoint already achieved 86.1% on the same benchmark last month.

This opens the question of having more discriminative benchmarks:

– MMLU-Pro, introduced a week ago (<https://lnkd.in/e255Tn5a>) is a good idea but seems to have issues with duplicates and missing information, as reported by Dorrin_Verrakai on Reddit (<https://lnkd.in/eSY3f5N5>).

– @Imysorg introduced a "hard prompts" category in the arena to evaluate models on more challenging tasks (https://lnkd.in/ej7_DeMb), but this is limited to models on the Chatbot Arena.

ZOMI 本职工作



ZOMI

- 大模型并行加速算法专家
- 训练解决方案架构师
- 大模型训练架构师

ZOMI 本职工作



ZOMI

- 大模型并行加速算法专家
 - 训练解决方案架构师
 - 大模型训练架构师
 - Whatever
-
- 在华为昇腾给国内各大厂做技术支持, So...

近期大家使用最多的开源大模型

1. DeepSeek-V2

- 基于专家网络 (MoE) 大语言模型, 是DeepSeek LLM (dense) 升级版
- 模型尺寸 236B, 其中激活参数21B; 上下文长度支持 128K;

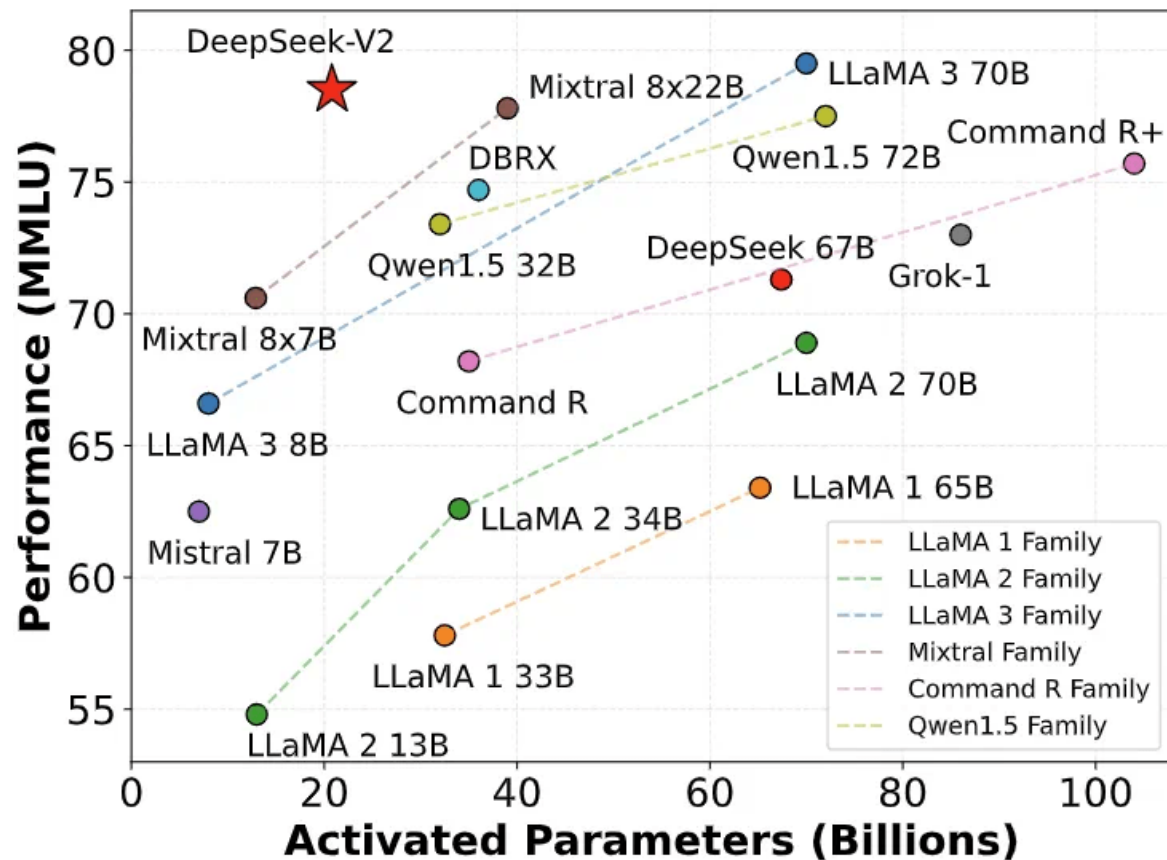
2. QWen2.5

- 尺寸基础和指令调优, 包括 Qwen2-0.5B、1.5B、7B、57B-A14B 和 Qwen2-72B 5 个版本;
- Qwen2 系列所有尺寸的模型都使用了 GQA, 预训练在 32K tokens 数据进行, 在 128K tokens 推理;

3. LLAMA3.1

- 8B、70B、405B 三版本, 405B (~820GB) 是目前最大开源模型;
- 微调后版本使用 SFT 和 RLHF 来对齐可用性与安全偏好;
- 引入长上下文窗口 (~128K Tokens) 和多模态能力支持;

开源大模型 DeepSeek-V2



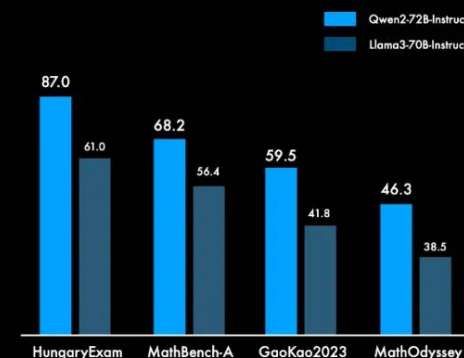
模型	开源/闭源	总分	中文推理	中文语言
gpt-4-1106-preview	闭源	8.01	7.73	8.29
DeepSeek-V2 Chat (RL)	开源	7.91	7.45	8.36
erniebot-4.0-202404 (文心一言)	闭源	7.89	7.61	8.17
DeepSeek-V2 Chat (SFT)	开源	7.74	7.30	8.17
gpt-4-0613	闭源	7.53	7.47	7.59
erniebot-4.0-202312 (文心一言)	闭源	7.36	6.84	7.88
moonshot-v1-32k-202404 (月之暗面)	闭源	7.22	6.42	8.02
Qwen1.5-72B-Chat (通义千问)	开源	7.19	6.45	7.93
DeepSeek-67B-Chat	开源	6.43	5.75	7.11
Yi-34B-Chat (零一万物)	开源	6.12	4.86	7.38
gpt-3.5-turbo-0613	闭源	6.08	5.35	6.71
DeepSeek-V2-Lite 16B Chat	开源	6.01	4.71	7.32

Compare With Others

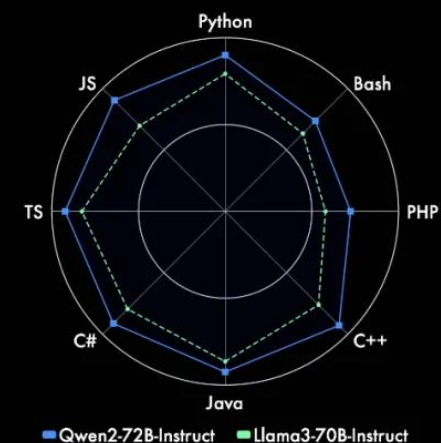
开源大模型 QWen2.X

	Qwen2-72B	Llama3-70B	Mixtral-8x22B	Qwen1.5-110B
MMLU	84.2	79.5	77.8	80.4
MMLU-Pro	55.6	52.8	49.5	49.4
GPQA	37.9	36.3	34.3	35.9
TheoremQA	43.1	32.3	35.9	34.9
BBH	82.4	81.0	78.9	74.8
HumanEval	64.6	48.2	46.3	54.3
MBPP	76.9	70.4	71.7	70.9
MultiPL-E	59.6	46.3	46.7	52.7
GSM8K	89.5	83.0	83.7	85.4
MATH	51.1	42.5	41.7	49.6
C-Eval	91.0	65.2	54.6	89.1
CMMLU	90.1	67.2	53.4	88.3
Multi-Exam	76.6	70.0	63.5	75.6
Multi-Understanding	80.7	79.9	77.7	78.2
Multi-Mathematics	76.0	67.1	62.9	64.4

$\sqrt{\Sigma}$ Math

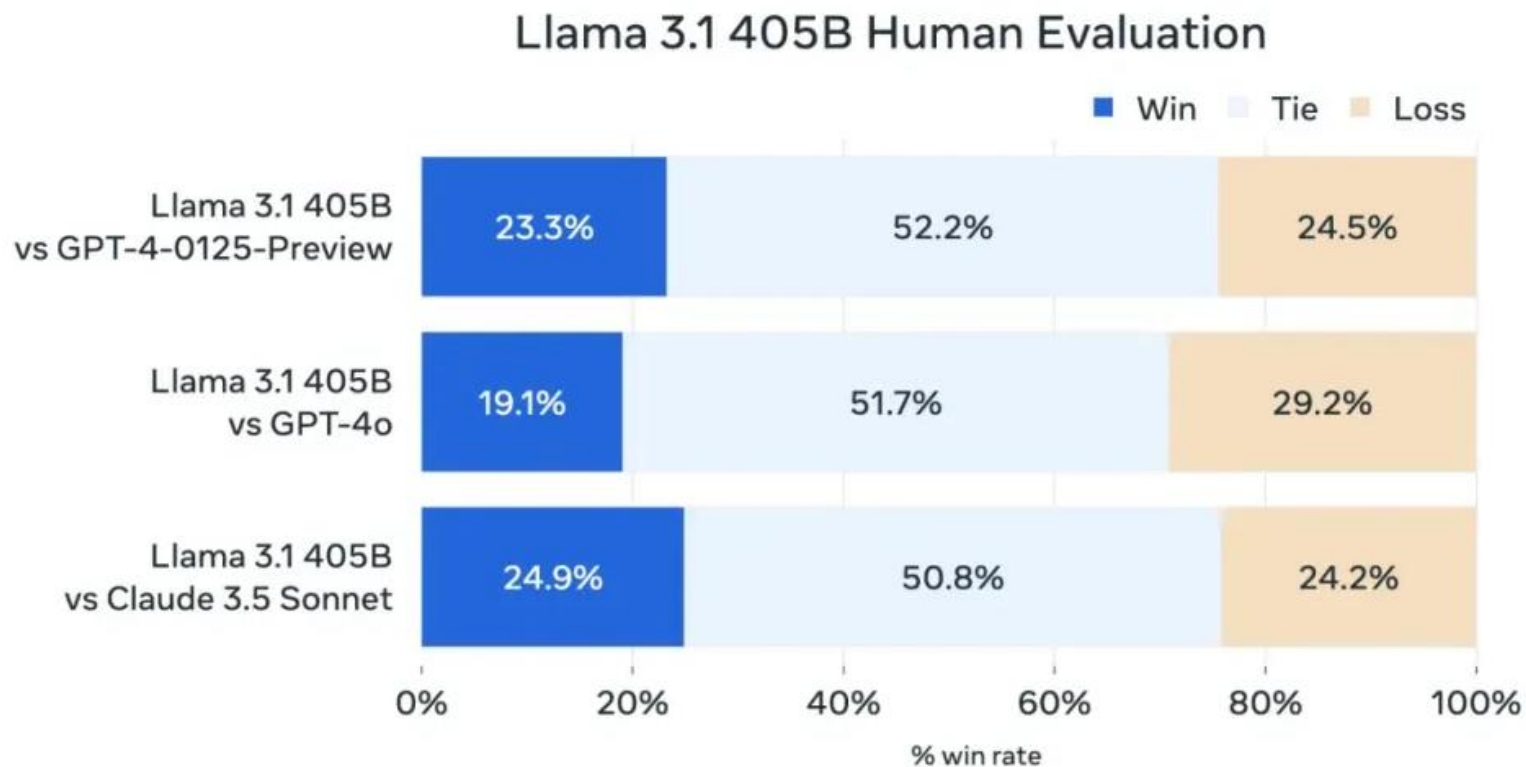


Σ Code



开源大模型 LLAMA3.1

- Llama 3.1发布前，闭源模型略胜一筹，Llama3.1 后开源模型终于可与闭源模型巅峰一战。
- Human Evaluation（评估理解和生成代码、解决抽象逻辑问题能力） 测评Llama 3.1略胜一筹。



开源大模型 LLAMA3.1

Llama 3.1 Performance				Benchmarks			
Category Benchmark	Llama 3.1 8B	Gemma 2 9B IT	Llama 3.1 70B	GPT 3.5 Turbo	Llama 3.1 405B	GPT-4 Omni	Claude 3.5 Sonnet
General							
MMLU Chat (0-shot, CoT)	73.0	72.3 (0-shot, non-CoT)	86.0	69.8	88.6	88.7	88.3
MMLU PRO (5-shot, CoT)	48.3	-	66.4	49.2	73.3	74.0	77.0
IFEval	80.4	73.6	87.5	69.9	88.6	85.6	88.0
Code							
HumanEval (0-shot)	72.6	54.3	80.5	68.0	89.0	90.2	92.0
MBPP EvalPlus (base) (0-shot)	72.8	71.7	86.0	82.0	88.6	87.8	90.5
Math							
GSM8K (8-shot, CoT)	84.5	76.7	95.1	81.6	96.8	96.1	96.4 (0-shot)
MATH (0-sho, CoT)	51.9	44.3	68.0	43.1	73.8	76.6	71.1
Reasoning							
ARC Challenge (0-shot)	83.4	87.6	94.8	83.7	96.9	96.7	96.7
GPQA (0-shot, CoT)	32.8	-	46.7	30.8	51.1	53.6	59.4
Tool use							
BFCL	76.1	-	84.8	85.9	88.5	80.5	90.2
Nexus (0-shot)	38.5	30.0	56.7	37.2	58.7	56.1	45.7
Long context							
ZeroSCROLLS/QuALITY	81.0	-	90.5	-	95.2	90.5	90.5
InfiniteBench/En.MC	65.1	-	78.2	-	83.4	82.5	-
NIH/Multi-needle	98.8	-	97.5	-	98.1	100.0	90.8
Multilingual							
Multilingual MGSM (0-shot)	68.9	53.2	86.9	51.4	91.6	90.5	91.6

1. 通用:

- 在 MMLU、GSM8K 等基准上表现突出。

2. 编程能力:

- 在 HumanEval 和 MBPP 等编程基准上取得高分。

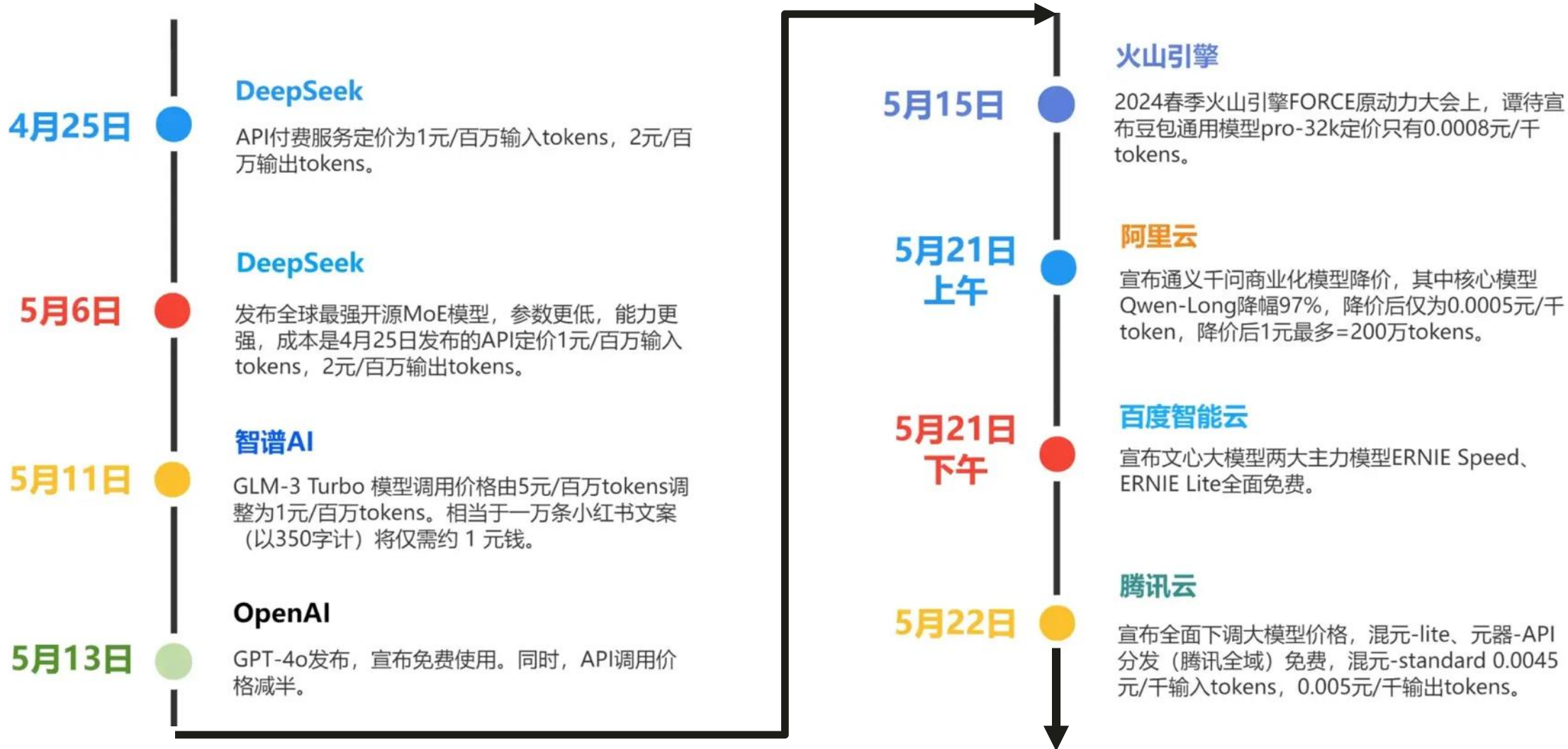
3. 多语言能力:

- 支持多种语言, 尤其在德语、法语等语言上表现良好。

03.

大模型降价潮

大模型 API 降价时间线



5 月份大模型价格战 TimeLine

05.06

幻方量化

05.15

字节

05.21

阿里

05.21

百度

05.22

腾讯

十亿

- 豆包 Lite: 输入 ¥0.0008, 输出 ¥0.001

- Qwen-long: 输入 ¥0.0005, 输出 ¥0.002

- Ernie Lite/Ernie Speed: 免费

- 混元 Lite: 免费

百亿

- DeepSeek V2 236B开源
- 输入 ¥0.001/k token
- 输出 ¥0.002/k token

- 豆包 128K
- 输入 ¥0.005/k token
- 输出 ¥0.009/k token

- Qwen-Trubo: 输入 ¥0.002, 输出 ¥0.006
- Qwen-Plus: 输入 ¥0.004, 输出 ¥0.012

- Ernie 3.5: 输入 ¥0.012, 输出 ¥0.012 or
- 6 个月会员 ¥114(100B Tokens)

- 混元-std: 输入 ¥0.0045, 输出 ¥0.005
- 混元-std 256K: 输入 ¥0.015, 输出 ¥0.06

千亿

- Qwen-Max: 输入 ¥0.04, 输出 ¥0.12

- Ernie 4: 输入 ¥0.12, 输出 ¥0.12 or 1 个月会员 49.9

- 混元 Pro: 输入 ¥0.03, 输出 ¥0.1



04.





































回顾国内大模型

AI 基础大模型情况

AI基础大模型

闭源

开源

 GPT Sore	 PaLM Gemini	 Claude	 文心大模型	 Gemma	 DALL-E	 LLaMA	 Mixtral	 StableDiffusion Stable Audio	 MPT	 DBRX	 OLMo
 盘古大模型	 言犀大模型	 星火大模型	 豆包大模型	 Phi-3	 Grok	 Falcon	 XGen	 openELM	 通义大模型	 混元大模型	 星辰语义
 鸿鹄图文	 K7大模型 可灵、可图等	 云天书	 ABAB大模型	 360智脑	 百川大模型	 日日新 SenseNova	 GLM大模型	 Yi模型	 天工 Skywork	 紫东太初	 姜子牙大模型

备注：以上统计不完全，排名不分先后，部分同时存在开源闭源大模型产品。

开源大模型：优点

1. 透明性和可验证性：

- 开源模型的代码和数据完全公开，任何人都可以检查模型的构建和训练过程，确保其行为的透明性。这有助于发现和修正潜在的偏见和错误，提高模型的可信度。

2. 社区协作和创新：

- 开源社区的协作可以加速技术进步。研究人员和开发者可以在现有模型的基础上进行改进和创新，从而推动整个领域的发展。

3. 教育和学习资源：

- 开源模型为学生和研究人员提供了宝贵的学习资源。他们可以通过实践来理解模型的内部工作原理，进行实验和探索，从而推动教育和研究的发展。

开源大模型：缺点

1. 商业竞争力不足

- 由于代码和数据公开，开源模型可能缺乏独特的商业优势。一些公司可能会担心竞争对手利用这些开源资源获得同等的技术能力。

2. 安全和隐私问题

- 开源模型的公开性也意味着潜在的安全和隐私风险。恶意行为者可能会利用这些资源进行攻击或滥用，导致数据泄露和隐私侵犯。

闭源大模型：优点

1. 商业化和盈利能力

- 闭源模型通常由大型科技公司开发，这些公司可以通过提供API访问和企业解决方案来实现盈利。这种模式可以确保公司获得足够的资金来持续研发和改进模型。

2. 控制和保护

- 闭源模型的开发者可以完全控制模型的使用和分发，保护其知识产权和技术优势。这有助于防止技术被滥用或恶意利用。

3. 资源和支持

- 大型科技公司通常拥有充足的资源和专业团队，可以确保模型的性能和稳定性。闭源模型在技术支持和维护方面通常表现更好，这对商业客户尤为重要。

闭源大模型：缺点

1. 透明性不足

- 闭源模型的代码和数据不公开，外部研究人员和用户无法验证模型的内部机制和行为。这可能导致模型存在潜在的偏见和错误，而用户无法有效发现和纠正。

2. 创新受限

- 由于闭源模型的开发和改进仅限于开发公司内部，外部社区无法直接贡献或改进模型。这可能限制技术的创新和发展速度。

3. 成本和准入门槛

- 闭源模型通常通过付费API提供服务，对于个人开发者和小型企业来说，成本可能较高。同时，使用闭源模型需要遵守严格的使用协议，这也限制了模型的灵活应用。

开源和闭源对比

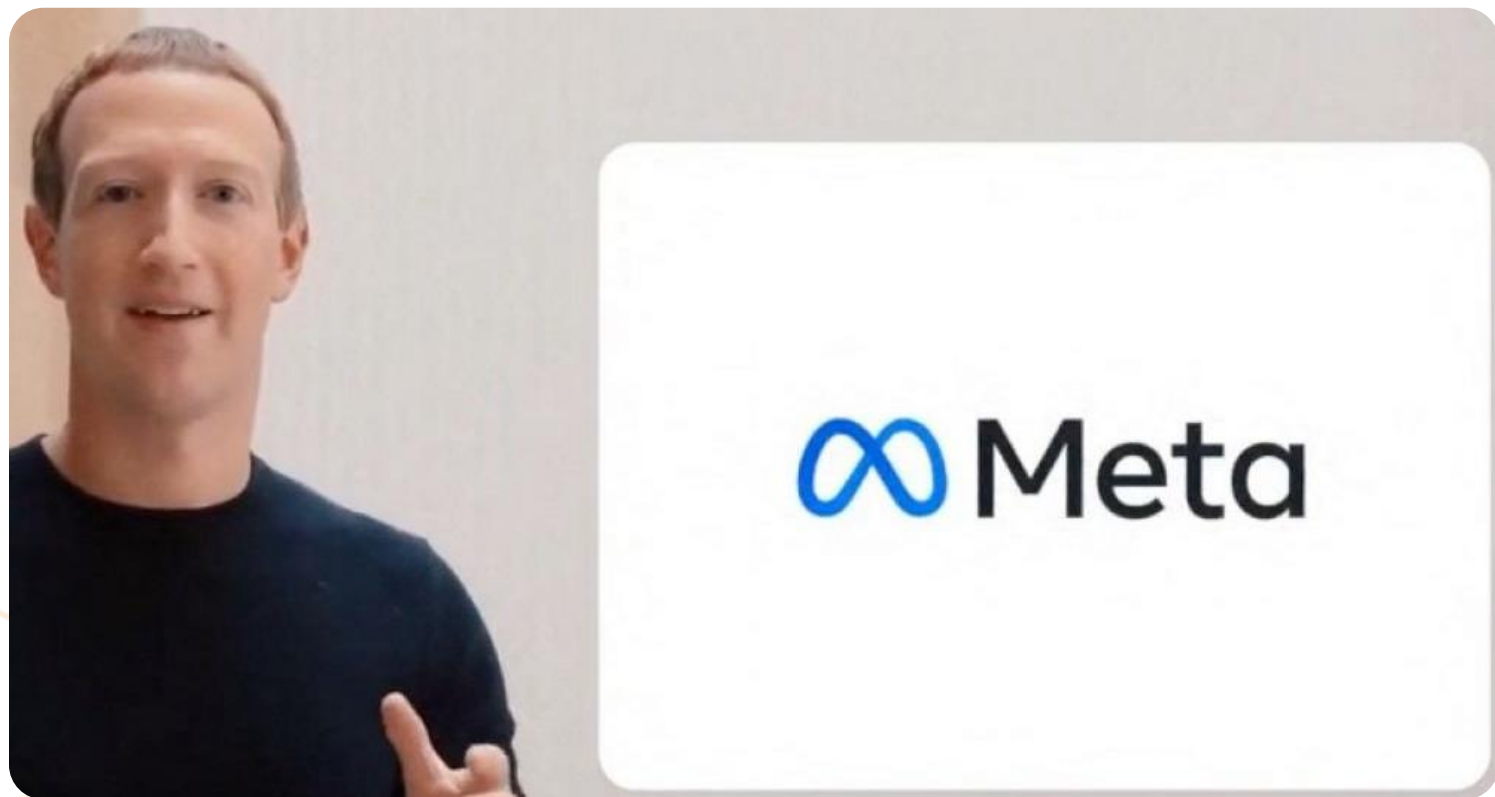
	开源大模型	闭源大模型
透明性与可控性	透明性更高，任何人都可以查看和验证其代码和数据，确保模型的行为符合预期。这对于学术研究和技术验证非常重要。	更注重控制和保护，开发者可以完全掌握模型的使用和分发，防止技术泄露和被滥用。
创新速度与商业应用	通过社区协作，能够快速迭代和创新，推动技术进步。例如，Hugging Face的Transformers库得益于全球开发者的共同贡献，发展迅速。	主要依靠公司内部的研究团队进行创新，虽然可能速度较慢，但通常更具商业应用的针对性和稳定性。例如，OpenAI的GPT-3通过付费API提供强大的商业服务。
资源共享与独占优势	资源共享方面表现突出，任何人都可以利用这些开源资源进行学习和研究，促进技术的普及和应用。	通过保护知识产权，确保技术的独占优势，从而在市场竞争中占据有利地位。这种模式能够为公司带来可观的经济收益，支持进一步的研发投入。
安全与隐私	公开性带来了潜在的安全和隐私风险，恶意行为者可能利用这些资源进行攻击或滥用。	通过控制代码和数据的访问，能够更好地保护用户隐私和数据安全，降低被恶意利用的风险。

05.

大模型发展思考

问题讨论

- 扎心了，为什么小扎这么热心，这么大方开源 LLAMA3.1?



问题讨论

- 扎克伯格接受 AI 博主采访时谈到对 Llama3 「超大杯」 的使用场景期待



Well, the thing that I'm most excited about is seeing people use it to distill and fine tune their own models, right?



为什么 MATE 这么大方？

- **Llama 有机会成为开源 AI 的行业标准，就像 Linux 之于操作系统一样：**
 1. 开源正在迅速缩小差距，Meta 预计未来几代 Llama 将成为业内最先进大模型；
 2. Meta 还将通过建立广泛、开放的生态 AI 系统（已经拥有 AI 框架 PyTorch）；
 3. Scale.AI、Dell等10+公司已准备好帮助企业采用 Llama 训练/蒸馏自定义模型；

Meta

Open Source AI Is the Path Forward

July 23, 2024

By Mark Zuckerberg, Founder and CEO



开源的动力

- 扩大自身企业和品牌的影响力及市场份额需要；
- 吸引开发者参与到开源大模型改进和优化中来；
- 有利于大模型快速迭代与发展，并为企业提供宝贵生态价值。

开源闭源市场份额

1. 开源模式将蚕食闭源模式一定的市场份额，两者在未来较长一段时间内或将形成相互摇摆的博弈格局。
2. 24 年百模大战厂商已经不是进入下半场了，而是进入尾声。目前国内大模型厂商，只有百度、月之暗面等坚持闭源，幻方、阿里、商汤、百川智能、智谱AI更多玩家则开源与闭源兼顾。



最后一公里

1. 入局 L0 大模型基本上机会渺茫，对于垂直行业大模型预训练意义不大，转而采取开源大模型 + 智能体/RAG，解决特定应用场景和问题。
2. 尽管围绕大模型开源与闭源的路线争论从未停歇，但行业共识：没有“最后一公里”的应用与商业化落地，开源与闭源都将失去意义。





Thank you

把AI系统带入每个开发者、每个家庭、
每个组织，构建万物互联的智能世界

Bring AI System to every person, home and
organization for a fully connected,
intelligent world.

Copyright © 2023 XXX Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. XXX may change the information at any time without notice.



Course [chenzomi12.github.io](https://github.com/chenzomi12)

GitHub github.com/chenzomi12/DeepLearningSystem