大模型 – 大模型算法

KIMI 长序列

看大模型发展趋势 II

ZOMI

大模型业务全流程

数据 & 模型算法

2. 数据处理

- 开源数据
- 数据预处理
- 向量数据库

3. 模型算法

- LLM模型架构
- 多模态一切皆Tokens

模型训练 & 微调

4. 模型训练

- 混合精度
- 梯度检查
- 梯度累积
- …
- 5. 分布式并行
- 训练集群稳定性

6. 模型微调

- 全参微调
- 低参微调
- 指令微调

模型验证 & 推理部署

7. 模型验证

- 下游任务
- 测评标准
- bench mark

8. 推理与智能体

- 量化压缩
- 推理加速
- 9. Agent 智能体

集群算力准备

1. AI 集群建设

- 计算、存储、网络
- AI 集群机房建设
- AI 集群上线与运维

关于本内容

1. LLM 大模型混战：最新 LLM 大模型介绍
2. LLM 长序列讨论：长序列带来问题
3. 看 LLM 大模型趋势：百模厂商的冲击 && 产业思考

2023：大模型元年，ChatGPT 引发

业界地震，LLAMA引导大模型开源

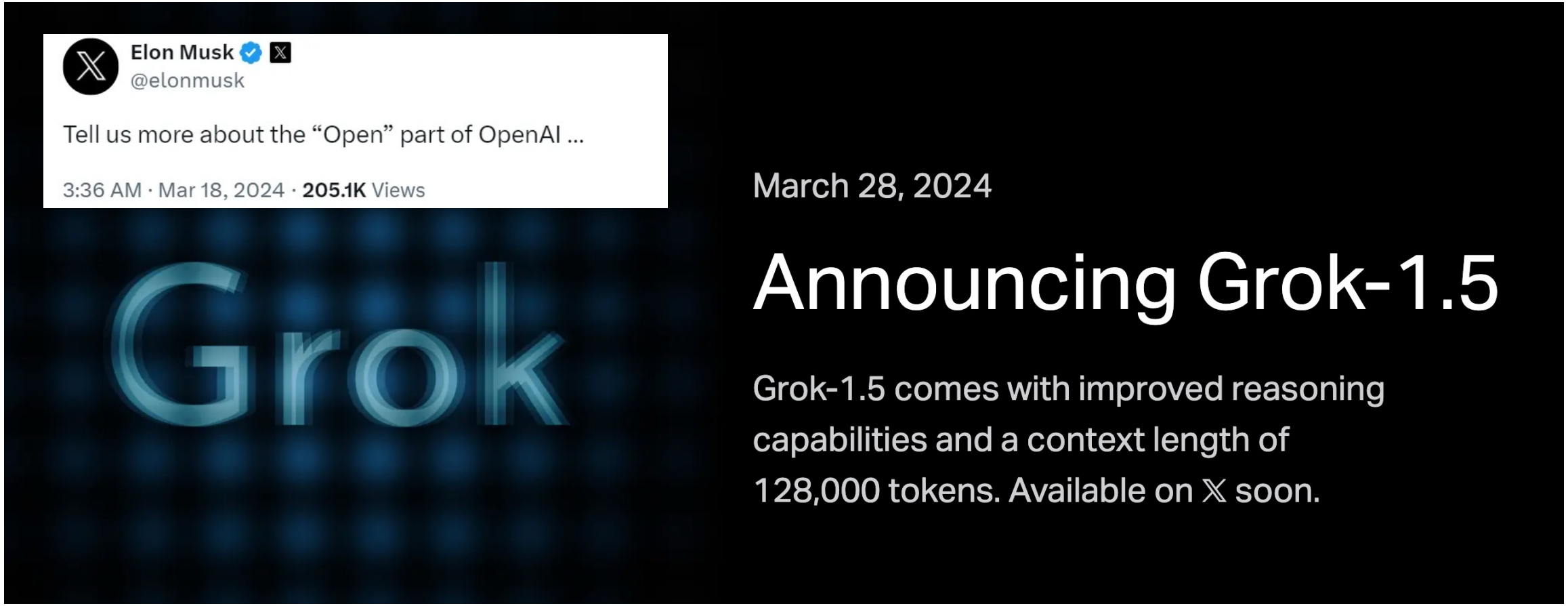
2024：大模型第二阶段，各大厂商发布

LLM 大模型，引发长序列、开闭源之争

1. LLM 混战

马斯克高调宣传 XAI 开源 Grok-1

- Grok-1当前参数量最大的开源大语言模型 https://github.com/xai-org/grok-1

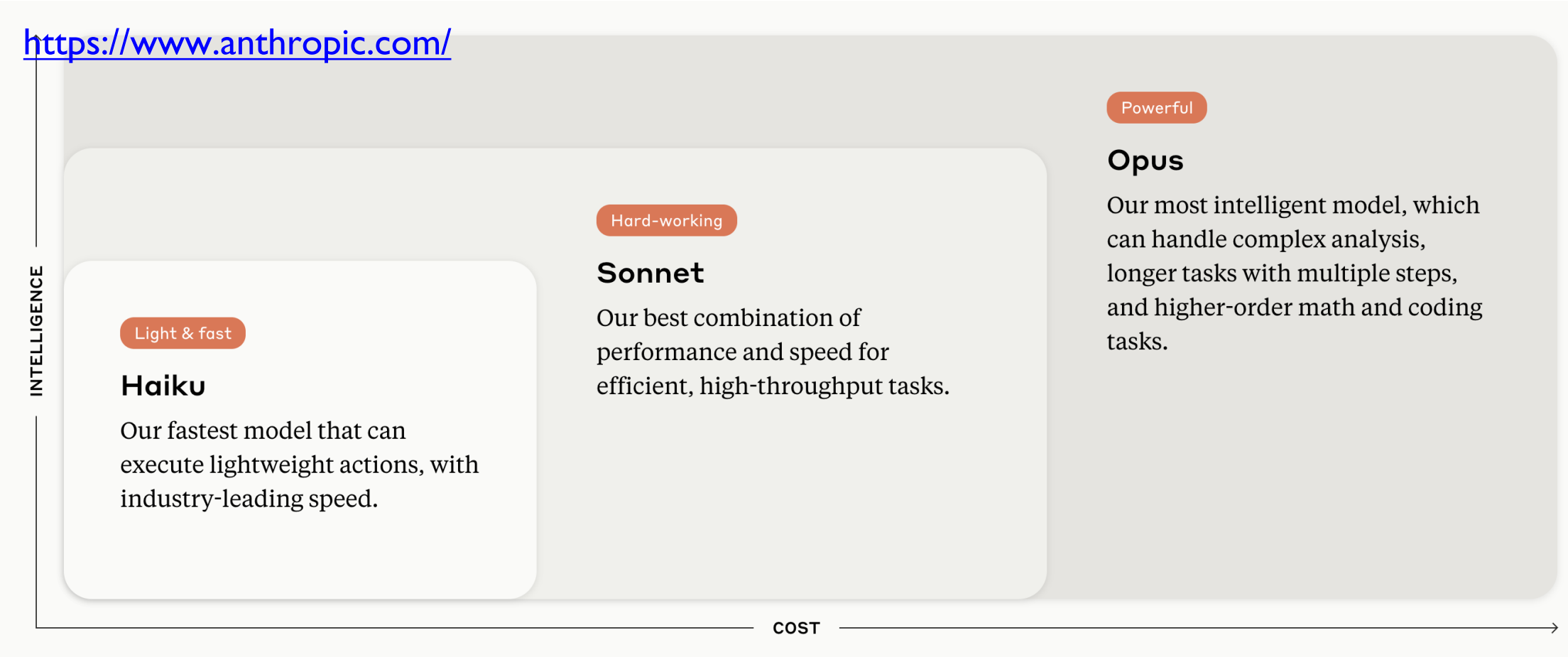


DataBricks 数据湖专业公司开源 DBRX

- 开源大模型王座再易主，1320亿参数DBRX上线
- https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm

Anthropic 国外大模型创业公司发布 Claude

- GPT-4 被全面碾压，Claude 3 一夜颠覆AI界
- https://www.anthropic.com/



一年融资 10 亿美元的月之暗面

- 突破200万字文本交互！一年融资10亿美元，月之暗面模型再升级发布 KIMI Chat
- https://kimi.moonshot.cn/

一年融资 10 亿美元的月之暗面

Kimi概念股一览

代码	简称	总市值(亿元)	3月至今涨跌幅(%)	2023年净利润同比(%)	2023年净利润(亿元)
300133	华策影视	210.26	101.46		
603533	掌阅科技	142.60	83.87		
603721	中广天择	55.37	74.41	-147.66	-0.08
688095	福昕软件	77.61	70.86	-5582.84	-0.99
688787	海天瑞声	55.93	59.98	-205.04	-0.31
002343	慈文传媒	43.79	50.41		
301085	亚康股份	61.52	40.95	-2.94	0.78
301313	凡拓数创	32.48	31.66		
300442	润泽科技	580.35	30.08	50.22	18.00
300182	捷成股份	176.08	29.61		
002432	九安医疗	252.60	28.33	-92.52	12.00
300454	深信服	330.84	24.29	-1.64	1.91
688111	金山办公	1522.80	21.70	17.92	13.18
601801	皖新传媒	168.68	21.14		
603598	引力传媒	57.72	21.14	扭亏	0.53
300634	彩讯股份	95.57	21.03	50.20	3.38
300364	中文在线	249.27	18.29	扭亏	0.88
002858	力盛体育	23.42	14.05	-103.42	-1.55
002929	润建股份	124.62	13.35		
300315	掌趣科技	166.56	12.41	107.94	2.00
600556	天下秀	103.04	12.20		
300541	先进数通	51.77	11.79	50.12	1.60
603322	超讯通信	60.41	-0.83		

注：净利润按照年报、快报、预告区间中值顺序选取

- A股市场上，Kimi概念股约23只。
- 受益于Kimi消息面的利好催化，3月至今，概念股集体大涨。
- 龙头股华策影视3连板，月内累计涨幅高达101.46%。公司与月之暗面进行了深度合作，其内部系统已经接入Kimi。
- 掌阅科技、中广天择、福昕软件等月内涨幅均超过70%。

阶跃星辰大模型新玩家官宣 Step 系列

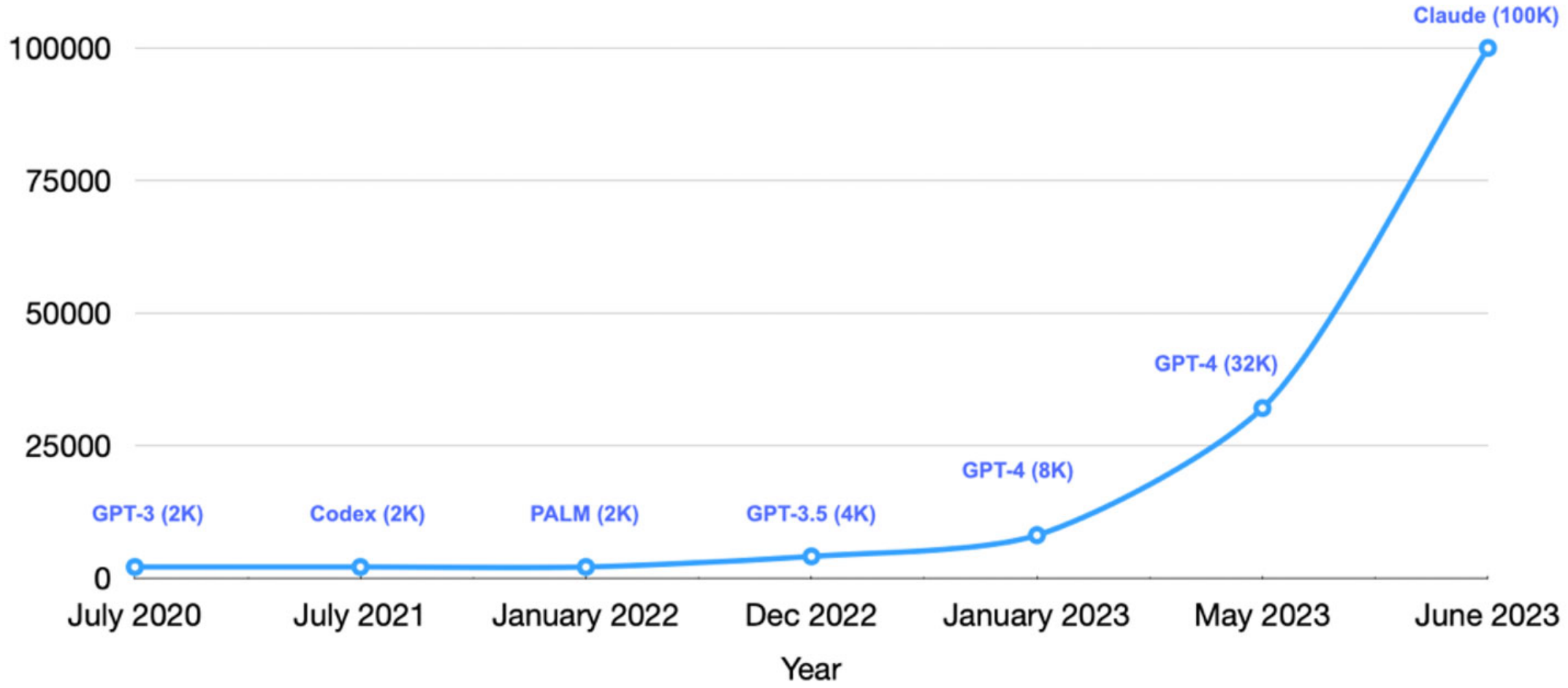
- 官宣Step-1 千亿规模LLM && Step-1V 千亿参数MLM && Step-2 万亿规模 MoE LLM 预览版
- https://www.stepfun.com/#step2

大模型涌现

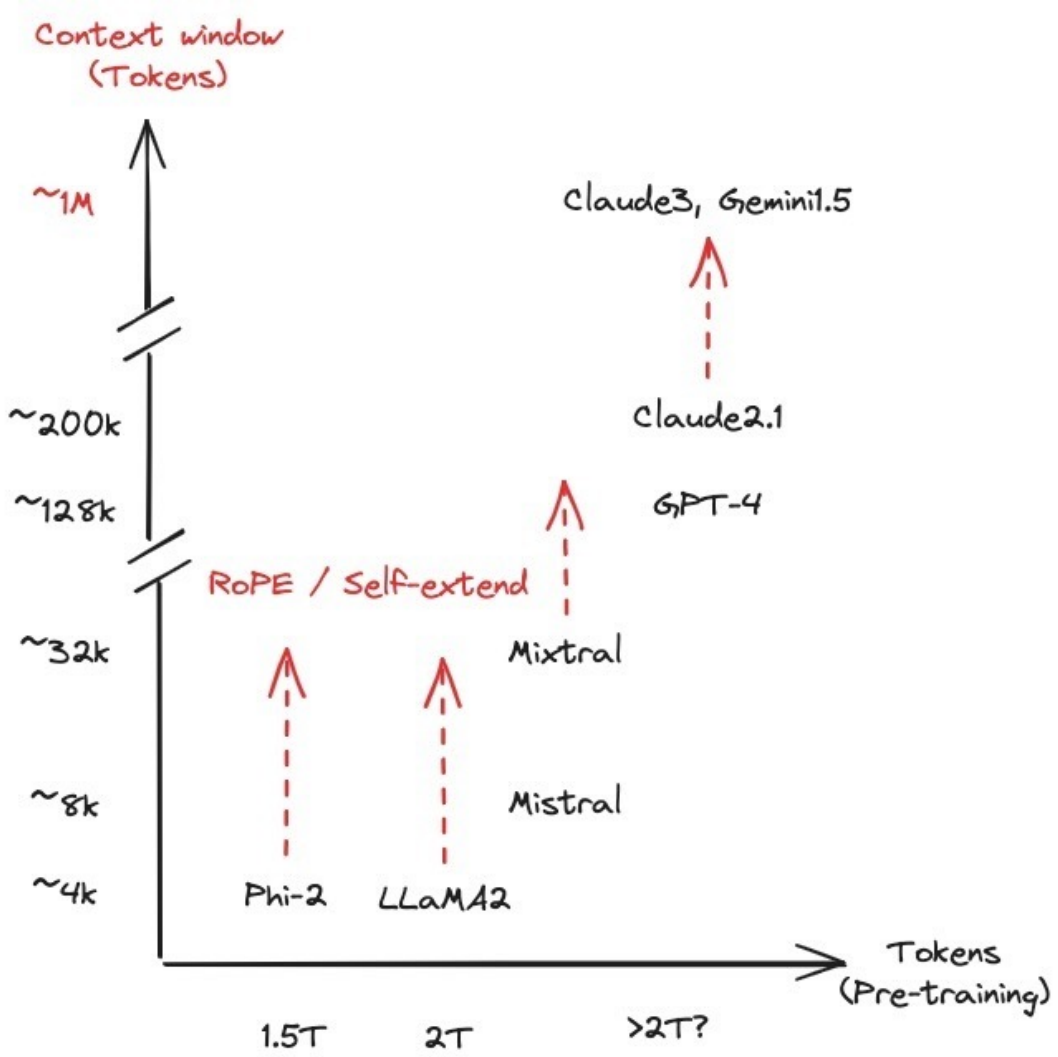
Model	参数量	模型结构	序列长度	数据规模	开源	时间	企业
DBRX	132B	16 * MOE（4 * 36B Act）	训练 32k 推理 32k	12T	开源	2024-03-27	Databricks
Step-1/2	/	MOE	/	/	闭源	2024-03-24	阶跃星辰
Mistral v0.2	7B	8 * MOE（2* 7B Act）	训练 32K 推理 32K	/	Tiny 开源 Large 闭源	2024-03-24	Mistral AI
KIMI Chat	/	/	推理 ~3000K 推理 2000K 中文	/	闭源	2024-03-18	月之暗面
Grok-1	314B	8 * MOE（2* Act）	训练 8K 推理128K	/	开源	2024-03-17	XAI
Qwen1.5	0.5B、1.8B、4B、7B、14B 和 72B	Decoder-only	训练 32k	/	开源	2024-02-07	阿里
Claude 2.1	/	/	推理 200K	/	闭源	2023-11-25	Anthropic

2. LLM 长序列

LLM 大模型序列长度


Claude (100K)
100000
75000
50000
25000
GPT-4 (32K)
GPT-4 (8K)
GPT-3 (2K)
Codex (2K)
PALM (2K)
GPT-3.5 (4K)
0
July 2020
July 2021
January 2022
Dec 2022
January 2023
May 2023
June 2023
Year

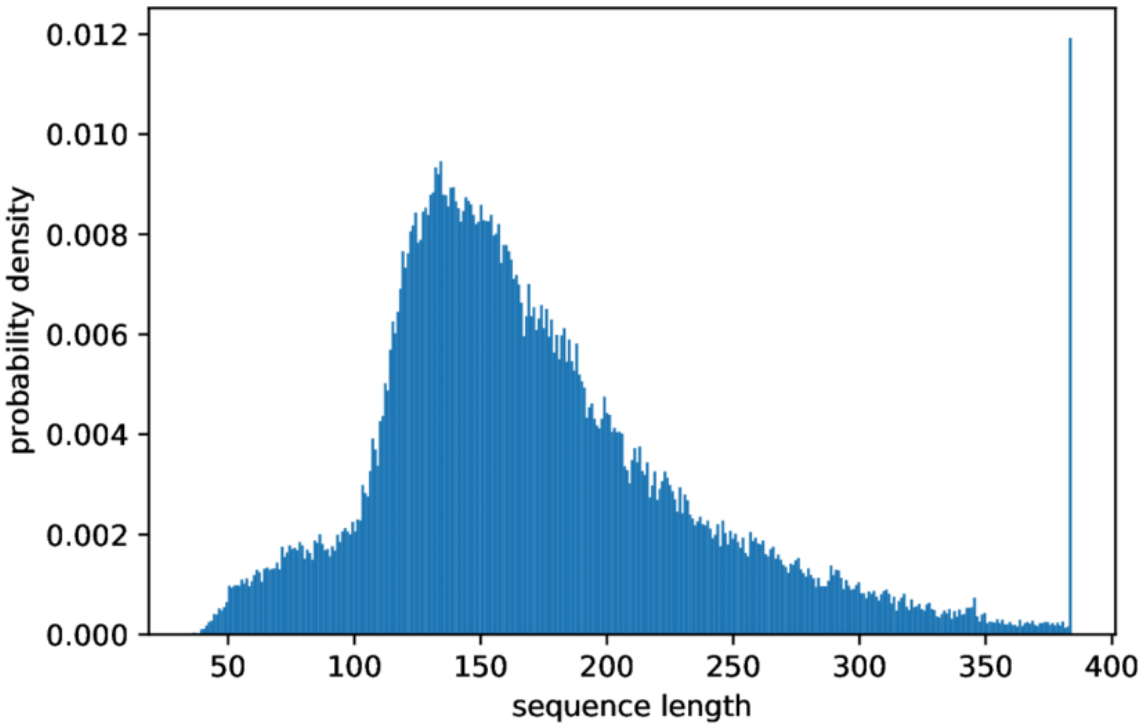
长文本大模型之争

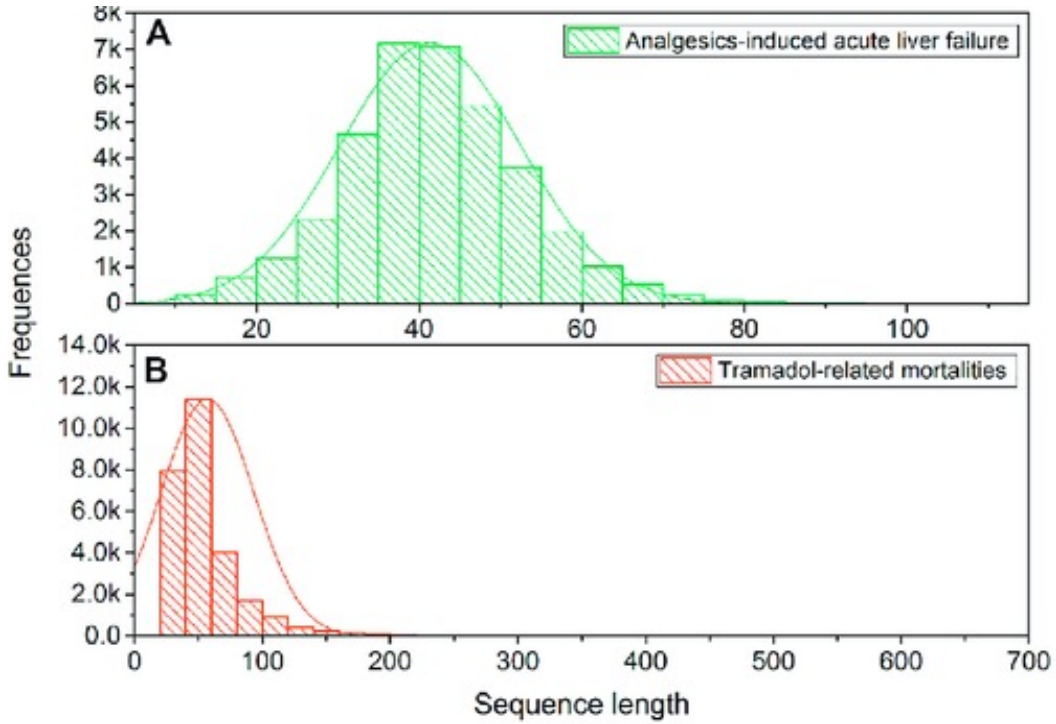


- KIMI 上线时间 2023年10月，可以支持无损上下文长度最多为20万汉字。在5个月的时间内，月之暗面直接将长文本能力提高10倍。
- 长文本能力最强的谷歌gemini 1.5、Claude3支持100万token，kimi 200万汉字上下文长度或已超越海外顶尖大模型水平。

长序列技术挑战 1：训练数据比例低

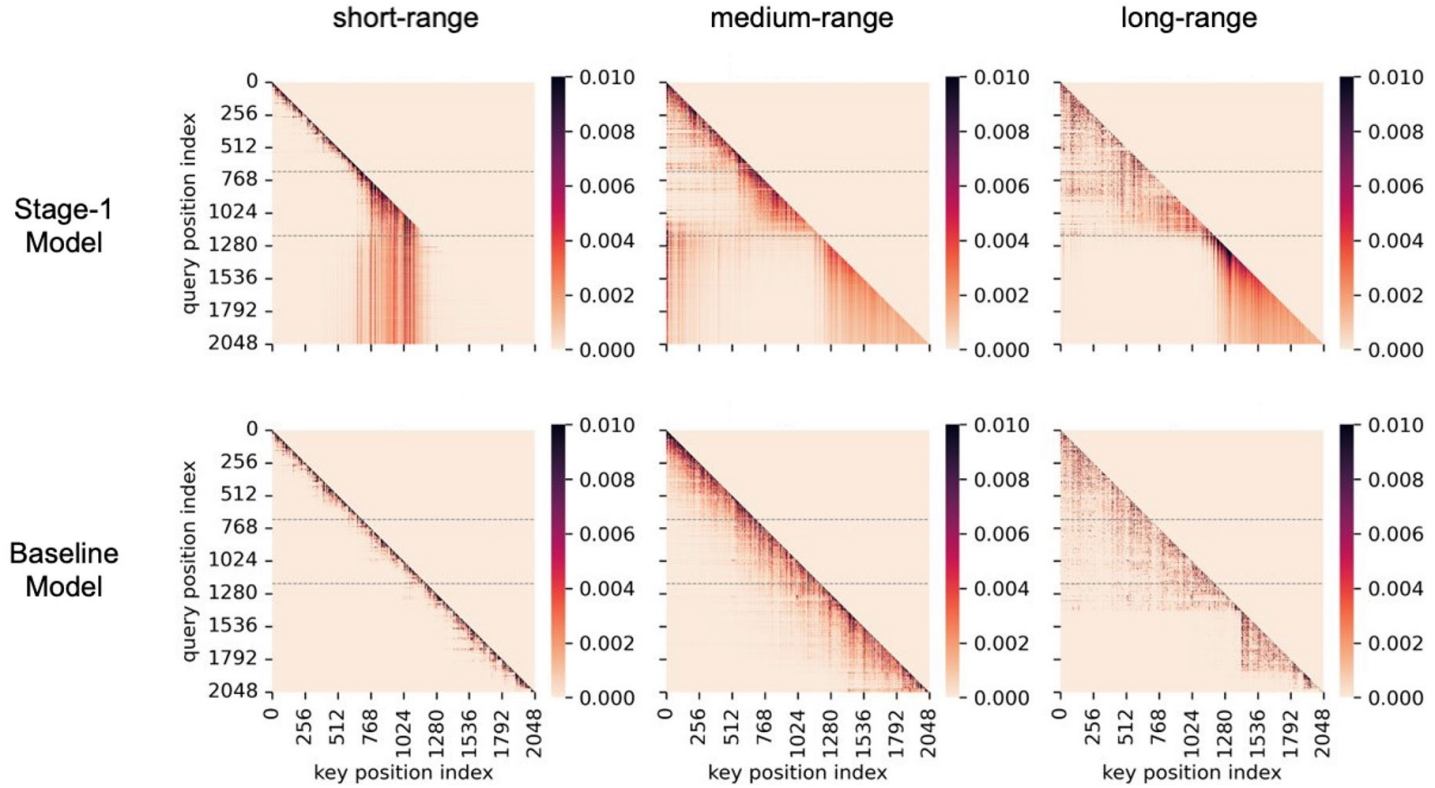
- ~90% 训练数据 Token 小于 1K；
- <1% 训练数据 Token 长度大于 32K；

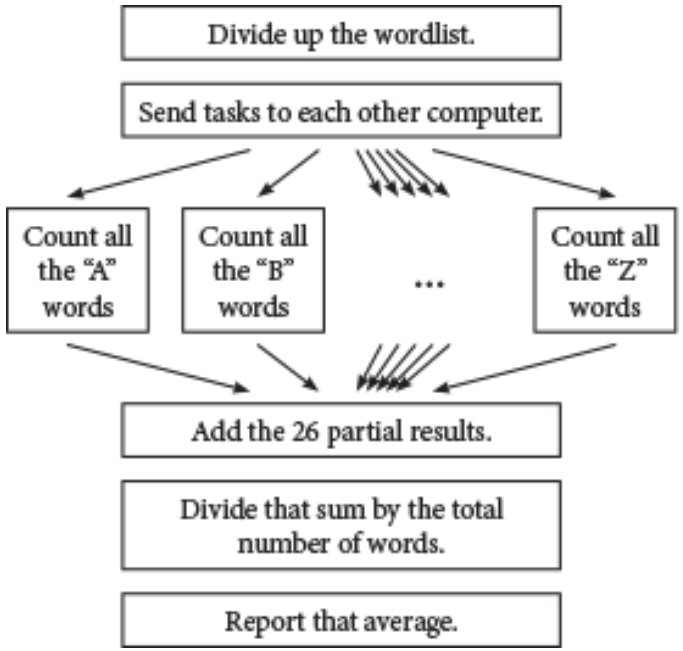




长序列技术挑战 II：训推成本和难度增加

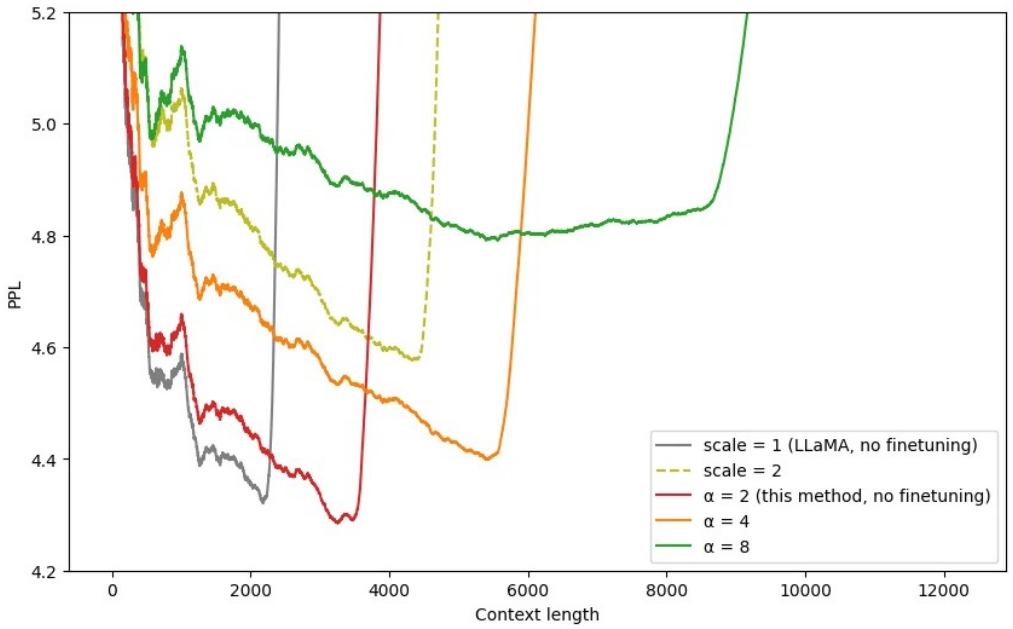
- 计算复杂度随输入序列长度乘平方增加 $O(n^2)$；
- 输入序列长度越长，模型需要序列并行，并行度越高模型利用率 MFU 越低；

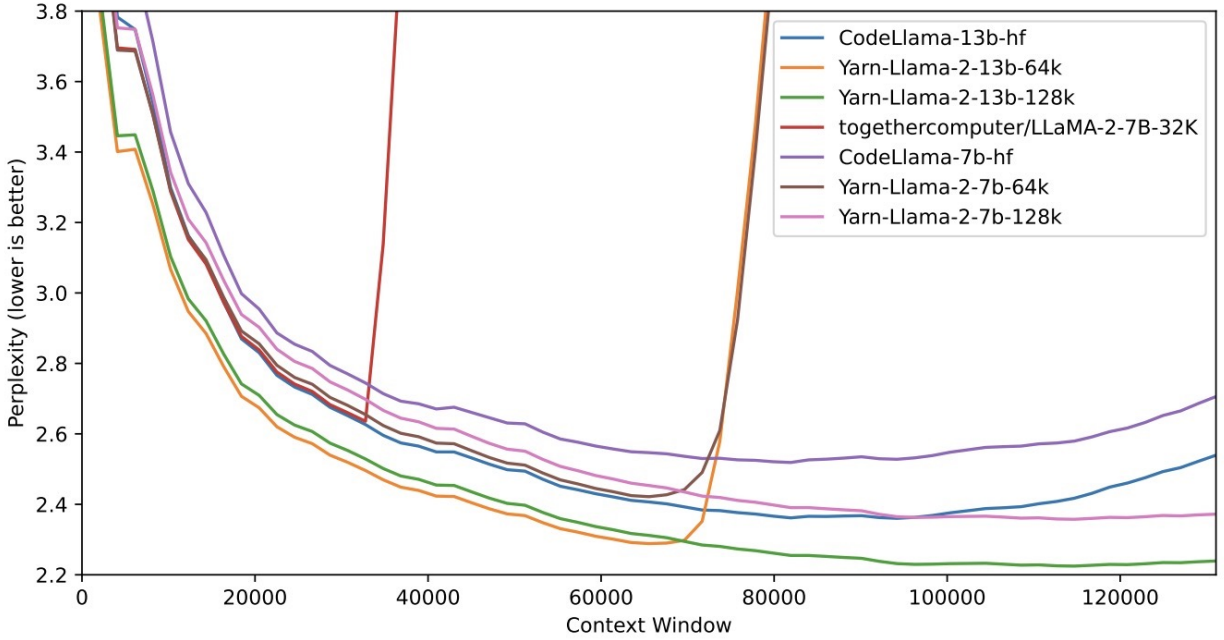




长序列技术挑战 III：模型能力评估难

- 模型性能评估随着序列长度扩展而恶化，执行评估任务越来越慢；
- 如何观察模型在整个序列中的信息；





通过综合算法、数据、工程实现超长序列

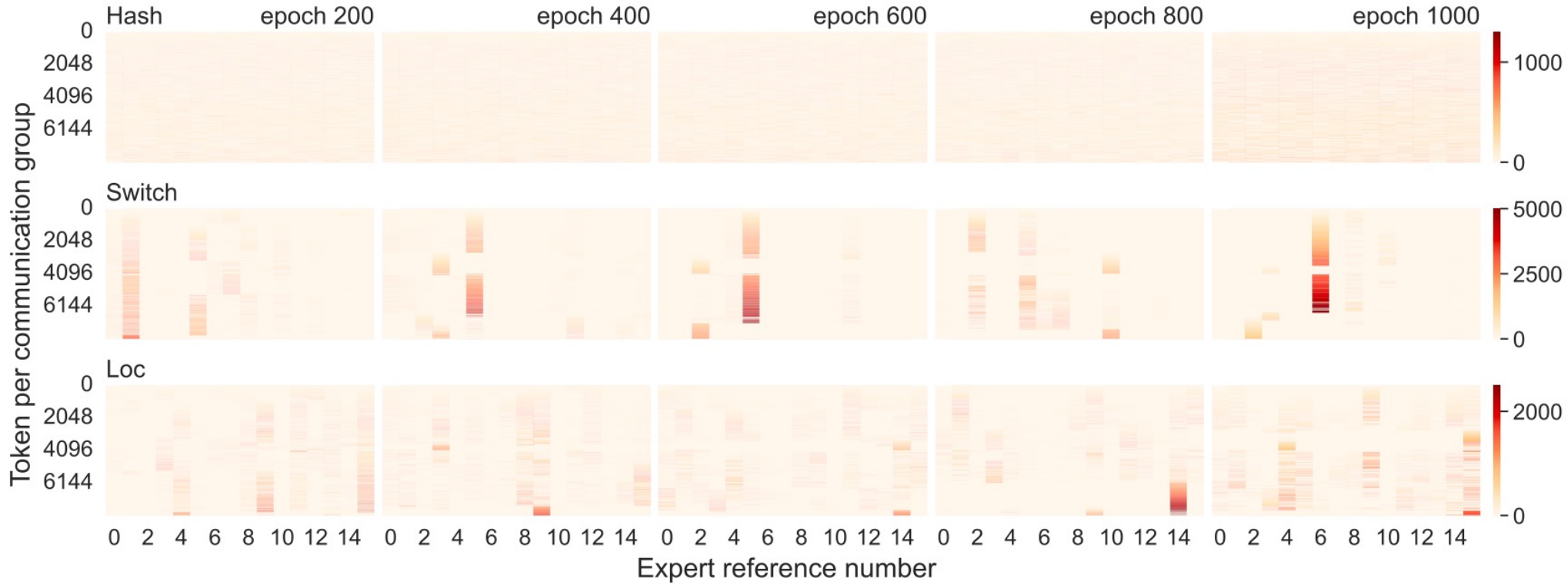
长文本的优点

1. 更好地理解文档：通过扩展上下文窗口，LLM 可以更好地捕捉文档中长距离依赖和全局信息，从而提高摘要、问答等任务的性能。
2. 增强指代消解：更长上下文窗口，帮助 LLM 更好地确定代词所指代的实体，从而提高指代消解的准确性。
3. 改进机器翻译：扩展上下文有助于更好地保留原文的语义，尤其是在专业术语、歧义词等方面，提高翻译质量。
4. 增强few-shot 能力：LLM 可以更好地进行 few-shot 学习，提高新任务泛化能力。LLM 直接通过 Prompt 就可以让模型学到新能力。

3. MOE的挑战

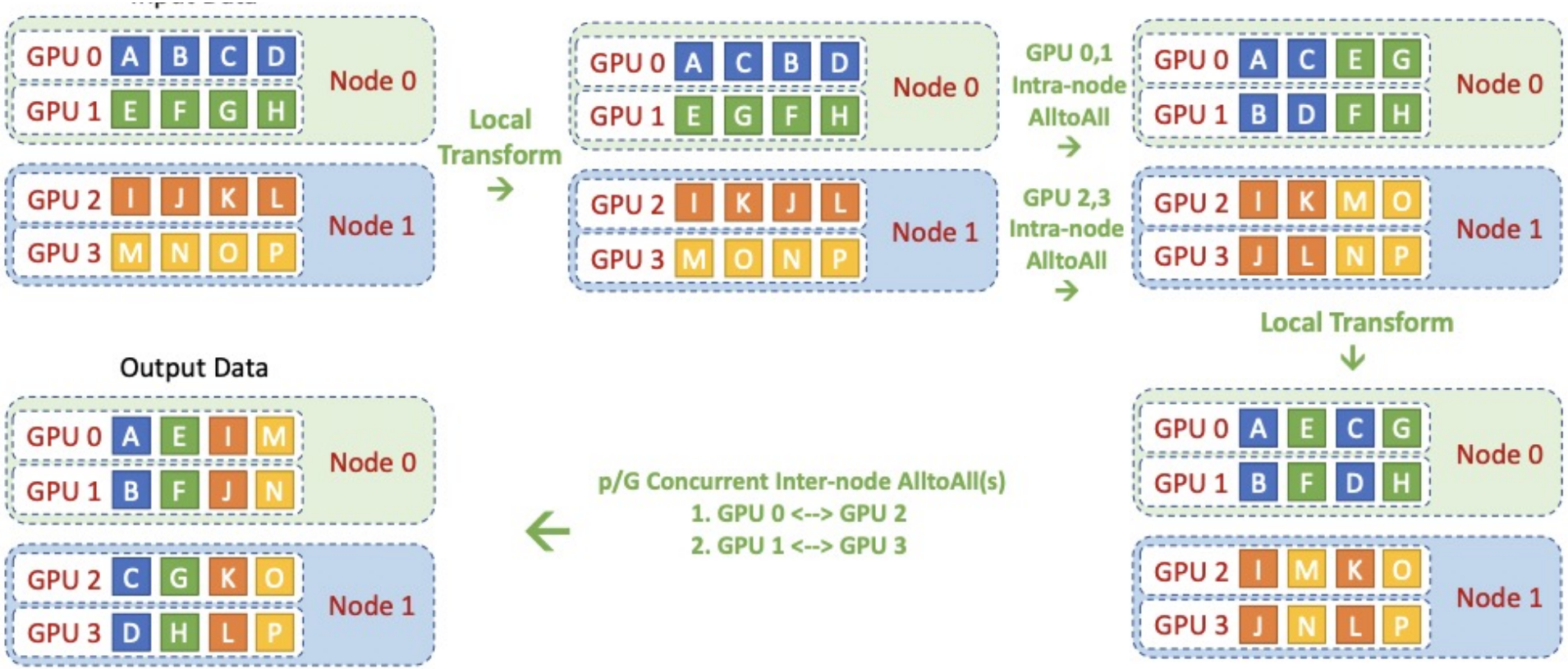
长序列技术挑战 1：训练负载不均衡

- Gate Network需要和Expert Network共同训练，需要新的负载均衡Loss；
- 保证 Expert 训练均衡，减少计算资源浪费（15% Expert Deal 75% Data）；



长序列技术挑战 II：训练有效性

- 不同 Expert 切分到不同 rank，节点间进行 Token 交互；
- MOE 集合通信使用 ALL2ALL，在端到端的通信耗时占比大，影响 MFU；



4. 看 LLM 趋势

LLM 大模型发展趋势与思考

- 对大模型技术趋势思考：

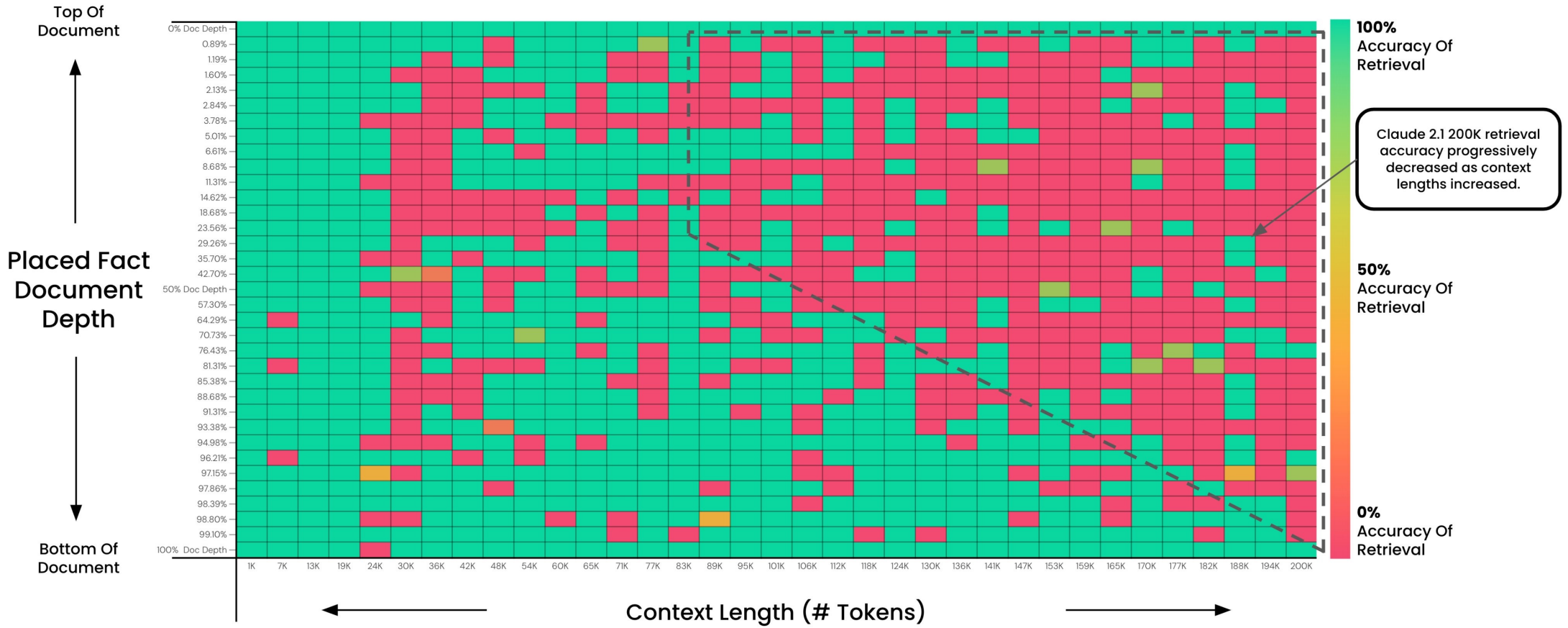
1. 国外逐渐掌握大模型训练能力，大模型技术开始同质化；国内大模型超预期开始出现头部；
2. 长文本、MOE 将会是 24 年大规模 LLM 大模型（7B~300B）云测训练和部署重要趋势；
3. 小规模 LLM 大模型（1B ~ 7B）将助力 AI PC、AI Phone 应用快速落地；

- 对计算厂商、大模型厂商、投融资思考：

1. 23 年主要跟随 LLAMA，24 年国外开源闭源大爆发如何快速响应和支撑业界大模型发展？
2. 百模大战厂商面对快速发展的国内外 LLM 大模型应该如何走出独立道路？
3. KIMI 概念股背后的推手？

长序列 Key Points

- 长序列重要的指标是看召回 Recall (Accuracy of Retrieval)，而不是创作能力 (Generative)



MOE 混合专家 Key Points

- MOE 稀疏结构重点不在模型参数量，而不是模型效果

Benchmark	Grok-1	Grok-1.5	Mistral Large	Claude 2	Claude 3 Sonnet	Gemini Pro 1.5	GPT-4	Claude 3 Opus
MMLU	73% 5-shot	81.3% 5-shot	81.2% 5-shot	75% 5-shot	79% 5-shot	83.7% 5-shot	86.4% 5-shot	86.8 5-shot
MATH	23.9% 4-shot	50.6% 4-shot	—	—	40.5% 4-shot	58.5% 4-shot	52.9% 4-shot	61% 4-shot
GSM8K	62.9 8-shot	90% 8-shot	81% 5-shot	88% 0-shot CoT	92.3% 0-shot CoT	91.7% 11-shot	92% 5-shot	95% 0-shot CoT
HumanEval	63.2% 0-shot	74.1% 0-shot	45.1% 0-shot	70% 0-shot	73% 0-shot	71.9% 0-shot	67% 0-shot	84.9% 0-shot

对百模大战产商思考

- 23 年跟随 LLAMA 甚至魔改 LLAMA（YI 大模型），24 年国外开源闭源大爆发：

1. 继续跟随 LLAMA3 进行魔改？还是基于 Grok、 DBRX进行 L1/L2 进行微调和魔改？
2. LLM 百模大战的厂商核心竞争力在何方？应用服务？政企关系订单？转变文生视频 SORA 赛道？
3. 面对国外王炸级产品，选择什么时候发布自家 LLM 大模型合适？

对百模大战产商思考

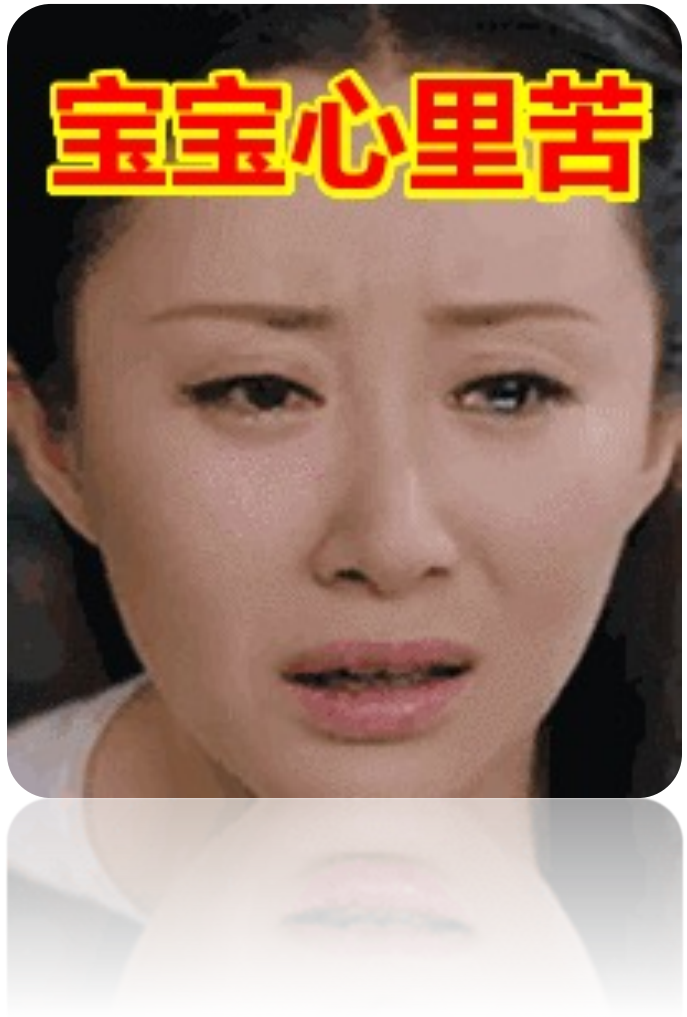
- 23 年跟随 LLAMA 甚至魔改 LLAMA（YI 大模型），24 年国外开源闭源大爆发：

1. 为什么首先做出大模型的不是国内大厂，而是创业公司？
2. 大模型是否只需要真正 N 哥核心人才，而不是乌合之众？
3. 大厂严重 KPI 导向方式不利于大模型方向创新，需要灵活快速适应变化的创业公司？

对计算产商思考

- 23 年适配好 LLAMA 系列及其衍生（ BaiChuan etc..），24 年国外开源大爆发：

1. 如英特尔 Gaudi2 POC 时只能适配 LLAMA 系列模型，计算产商 24 年继续努力跟跑开源大模型?
2. 面向国内计算产商 如寒武纪/燧原 等需要耗费大量人力适配，24 年继续适配还是聚焦底层能力?
3. 如何真正破局？聚焦底层能力，构建模块化性能 && 精度 Benchmark，LLM 大模型自行组装


宝宝心里苦

对投融资思考 I

- A 股出现 KIMI 概念股，火了一波公司，部分券商 · 二级赚一茬：

1. A 股真的是什么都可以炒的，只要有故事… 入局需谨慎，炒股要大胆；可投资的热点仍然在 AI，其他行业（医药、消费等）增长和故事在哪里？
2. 赶紧来了解 AI 行业与 AI 领域，但是不要相信 AI 网课、不要相信李一舟，多给 ZOMI 点赞多关注和推广宣传下 ZOMI；
3. 大模型虽然有技术成果，但仍处于上半场，下半场产品和商业模式还没到，投大模型公司而不是投大模型，包括大模型应用、训推一体化的公司。

对投融资思考 II

- 24 年国外大模型开源和闭源大爆发：

1. 短期内（24 年）仍然是大模型军备期，需要消耗大量算力训练，依然看好 NVIDIA/AMD 等 GPU 公司；
2. 大量算力消耗大头是训练服务器，HBM 内存、RDMA 交换机、光模块仍处于供不应求，看好 XXX；

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.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem