



德邦证券
Topsperity Securities

证券研究报告 | 行业深度

计算机

行业投资评级 | 优于大市（维持）

2024年7月4日

从技术路径，纵观国产大模型逆袭之路



证券分析师

姓名：陈涵泊

资格编号：S0120524040004

邮箱：chenhb3@tebon.com.cn

研究助理

姓名：王思

邮箱：wangsi@tebon.com.cn

核心逻辑

➤ 海外大模型龙头轮番抢占性能第一宝座，比拼整体性能和多模态交互。

- ✓ 大模型的宝座三次更迭：初代GPT-4o自我革命，持续刷新综合性能；二代谷歌Gemini更极限的上下文理解、更低延时；翘楚Claude3.5聚焦视觉和交互体验。
- ✓ 大模型高地争夺：多模态的理解和响应，原生多模态技术比拼。大模型的效果取决于多模态理解与生成，毫秒级响应，更先进的视觉与音频理解能力，智能感知语气与语态。端到端原生多模态技术、统一神经网络，是竞争的主要角力点。
- ✓ 大模型比较维度升级：从模型到叠加终端，跨设备的使用效果体验。如谷歌推出AI Agent项目Astra模型，可以手机、眼睛镜头对准身边的物品，并向Project Astra提出一些疑问，它几乎能做到零延时地准确回答。

➤ 国内大模型逆袭之路：聚焦长文本，降价迭代提升竞争力。

- ✓ 先文后理：理科目前差距较大，聚焦长文本，国产大模型已有赶超GPT之势，如通义千问、KIMI、山海等。
- ✓ 长文本的三大难度：注意力机制计算复杂度、上下文记忆、最长文本约束难题。
- ✓ 商业上降价，加速迭代卷出未来。头部智谱/字节跳动/阿里/腾讯/百度/讯飞低价迭代，百川智能/月之暗面/零一万物等初创公司并未加入降价行列。从技术来看，降价的背后是训练&推理成本的下降。

➤ **投资建议：**建议关注（1）国产大模型厂商：科大讯飞、商汤、云从科技、格林深瞳、拓尔思、昆仑万维、创业黑马等。（2）接入头部大模型的应用标的：金山办公、万兴科技、福昕软件、虹软科技、彩讯股份、焦点科技、润达医疗、金证股份、泛微网络、金蝶国际等，同时关注Kimi相关标的。

➤ **风险提示：**海外大模型展现闭源趋势，国内大模型技术差距扩大；国内大模型在整体性能上未能达到商业使用的奇点；国内大模型在缺乏算力支持的情况下迭代速度放缓；国内大模型技术路线产生分歧，无法引领未来发展方向。

目录 CONTENTS

- 01 海外龙头轮番抢占第一宝座，
比拼整体性能和多模态交互
- 02 国内大模型逆袭之路：聚焦
长文本，降价迭代提升竞争力
- 03 投资建议
- 04 风险提示



01

海外龙头轮番抢占第一宝座， 比拼整体性能和多模态交互

1.1 高速迭代，海外龙头轮番占领大模型之巅

1.2 大模型高地争夺：多模态的理解和响应，原生多模态技术比拼

1.3 比较维度升级：从模型到叠加终端，跨设备的使用效果体验

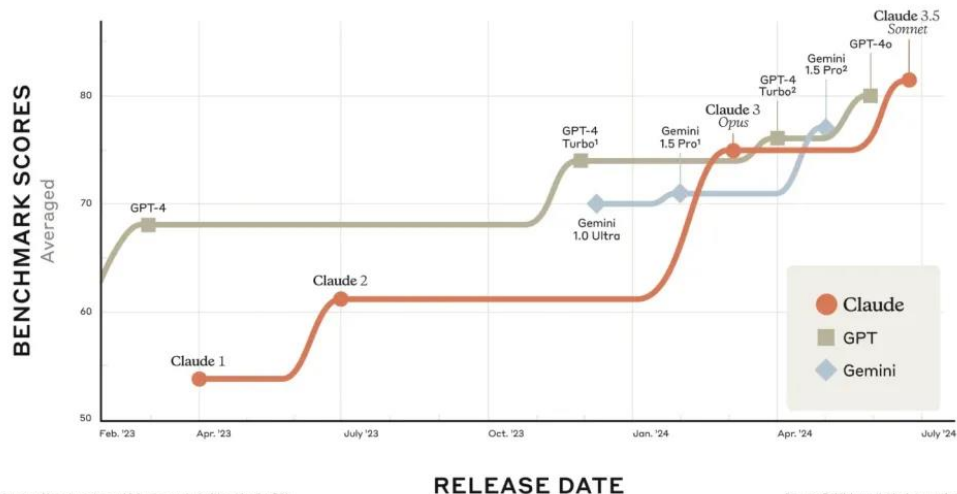
1.1 高速迭代，海外龙头轮换占领大模型之巅

➤ OpenAI、Google、Antropic三大厂商竞相轮换大模型第一宝座：自23年初GPT-4发布以来，在1年左右时间内基本稳定处于大模型最强位置。2024年海外大模型迭代速度有所加快，龙头竞争格局悄然发生变化。

- ✓ 1) 5月，OpenAI发布新的旗舰模型GPT-4o，实现跨模态即时响应，相比GPT-4 Turbo，刷新SOTA实现性能飞跃。
- ✓ 2) 5月，Google发布Gemini 1.5 Pro进阶版，实现200万tokens上下文，具备更强大的推理和理解能力。
- ✓ 3) 6月，Antropic发布Claude 3.5 Sonnet，具备更强的代码和视觉能力，基准测试结果全方位碾压Gemini 1.5 Pro和Llama-400b，大部分优于 GPT-4o，一定程度上暂时代表着当前大模型性能最高水平。

图表：海外龙头竞相抢占大模型第一宝座

AI model release and capabilities timeline



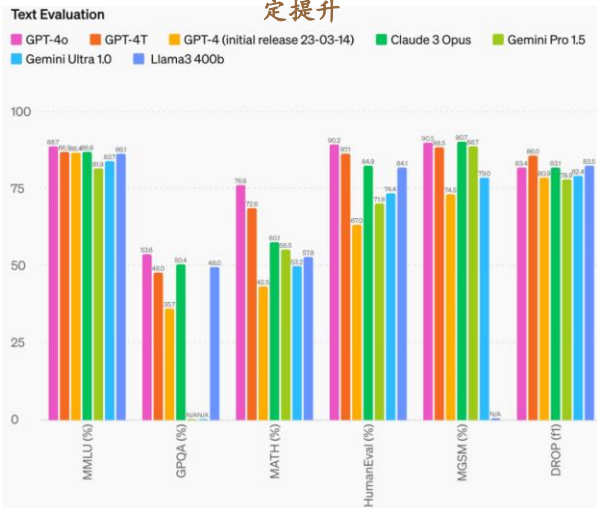
Averaged benchmarks are highest reported without best-of-N: MMLU, GPQA, MATH, MGSM, DROP F1, HumanEval pass@1, MMMU, AI2D, ChartQA, DocQA, Mathvista

Source: Publicly available data; evaluation scores are the average of representative scores found online. 1 = Initial release; 2 = Second release

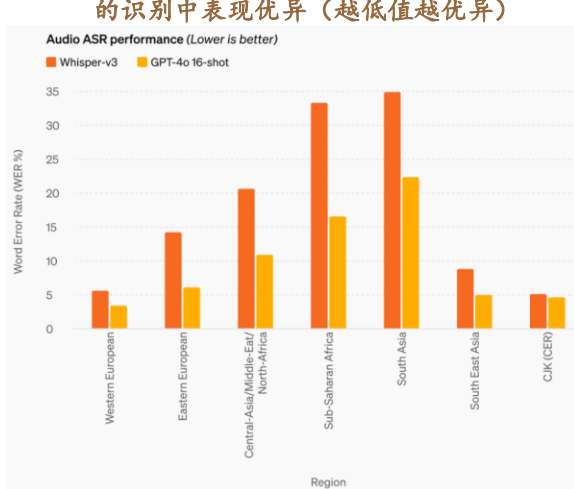
1.1.1 初代：GPT-4o自我革命，持续刷新综合性能

- 在传统基准测试中，GPT-4o在文本、推理和编码智能方面实现了GPT-4 Turbo级别的性能，同时多语言、音频和视觉功能上达到了新的高水位线。
- ✓ **文本推理**：GPT-4o在0-shot CoT MMLU（常识问题）上创下了88.7%的新高分。此外，在传统的5-shot no-CoT MMLU上，GPT-4o创下了87.2%的新高分。相较于GPT-4 Turbo而言，GPT-4o文本推理能力有一定提升。
- ✓ **多语言识别**：与Whisper-v3相比，GPT-4o在多种语言的识别中表现优异，尤其是资源匮乏的语言。
- ✓ **音频翻译**：GPT-4o在音频翻译表现上达到新的高水准，且在MLS基准测试中优于Whisper-v3。

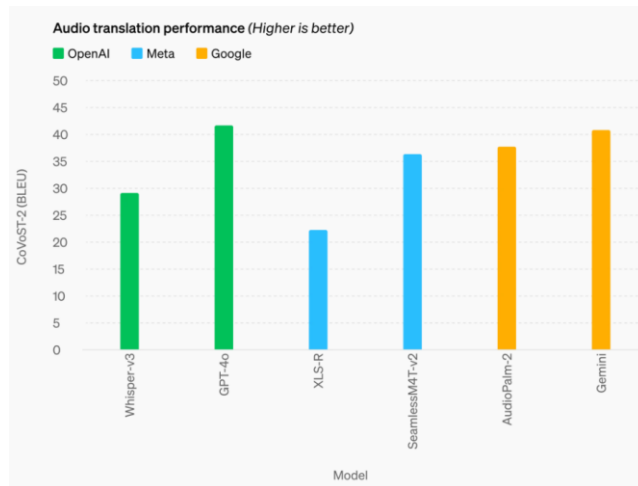
图表：GPT-4o文本推理相较GPT-4 Turbo有一定提升



图表：与Whisper-v3相比，GPT-4o在多种语言的识别中表现优异（越低值越优异）



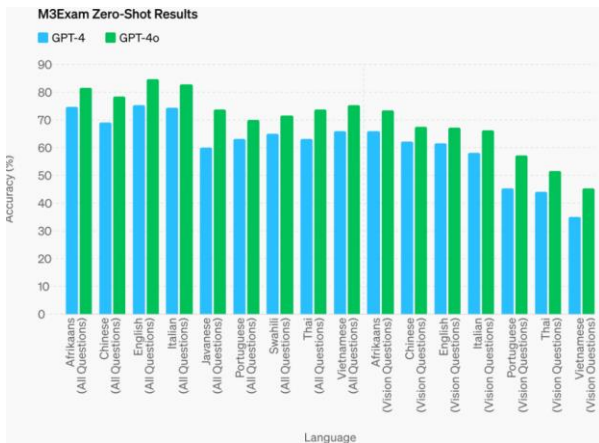
图表：GPT-4o在音频翻译表现上达到新的高水准



1.1.1 初代：GPT-4o自我革命，持续刷新综合性能

- ✓ **M3Exam测试**：M3Exam基准测试既是多语言评估也是视觉评估，由来自其他国家标准化测试的多项选择题组成，有时还包括图形和图表。在所有语言的基准测试中，GPT-4o都比GPT-4更强。
- ✓ **视觉理解**：GPT-4o在视觉感知基准上实现了最先进的性能。具体来看，GPT-4o在MMMU测试中分数达到69.1，而GPT-4 Turbo、Gemini 1.0 Ultra、Gemini 1.5 Pro、Claude Opus分别为63.1、59.4、58.5、59.4。
- ✓ **GPT-4o多模态能力范围显著拓展**：除了文本、图像等常用功能，GPT-4o还支持3D物品合成、文本转字体等多样化功能。
- **GPT-4o API性价比有所提升**：与GPT-4 Turbo相比，GPT-4o速度提升2倍、成本却降低了50%，且速率限制提升5倍。

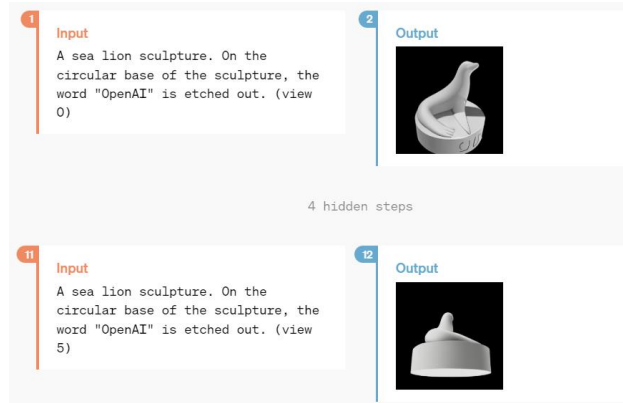
图表：在M3Exam基准测试中，GPT-4o比GPT-4强



图表：GPT-4o在视觉感知基准上实现了最先进的性能

Eval Sets	GPT-4o	GPT-4T 2024-04-09	Gemini 1.0 Ultra	Gemini 1.5 Pro	Claude Opus
MMMU (%) (val)	69.1	63.1	59.4	58.5	59.4
MathVista (%) (testmini)	63.8	58.1	53.0	52.1	50.5
AI2D (%) (test)	94.2	89.4	79.5	80.3	88.1
ChartQA (%) (test)	85.7	78.1	80.8	81.3	80.8
DocVQA (%) (test)	92.8	87.2	90.9	86.5	89.3
ActivityNet (%) (test)	61.9	59.5	52.2	56.7	
EgoSchema (%) (test)	72.2	63.9	61.5	63.2	

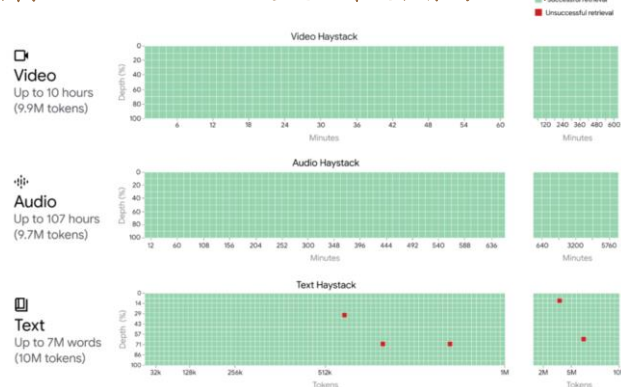
图表：GPT-4o实现3D物品合成



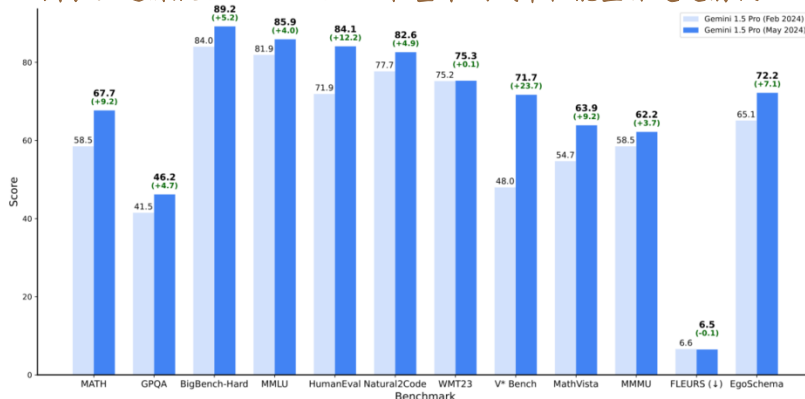
1.1.2 二代：谷歌Gemini更极限的上下文理解、更低延时

- (高性能)进阶版Gemini 1.5 Pro: 谷歌5月发布, 上下文窗口翻倍、具备更强大的推理与理解能力。
- ✓ 上下文: 过往Gemini 1.5 Pro支持100万tokens上下文, 升级后可支持200万tokens (理论极限为1000万tokens), 意味着可输入分析2小时视频、22小时音频、超过6万行代码或者140多万单词。这使得Gemini 1.5 Pro能处理更大量的复杂信息, 生成更准确、更细致的输出。
- ✓ 性能: 通过数据和算法改进, 升级版的Gemini 1.5 Pro增强了模型的代码生成、逻辑推理和规划、多轮对话以及音频和图像理解能力, 在MMMU、A12D、MathVista、ChartQA、DocVQA等多项公共基准测试中取得了显著改进, 在多项图像和视频理解基准测试中也实现了最先进性能。
- ✓ 价格: 输入7美元/百万tokens; 3.5美元/百万tokens (128k上下文)。

图表: Gemini 1.5 Pro理论上下文极限为1000万tokens



图表: 进阶版Gemini 1.5 Pro在基准测试中性能整体超越前代



图表: 进阶版Gemini 1.5 Pro在大多数功能上胜率高于前代模型

Gemini 1.5 Pro	Relative to 1.5 Pro (Feb)	Relative to 1.0 Pro	Relative to 1.0 Ultra
Long-Context Text, Video & Audio	no change	from 32k up to 10M tokens	from 32k up to 10M tokens
Core Capabilities	Win-rate: 78.1% (25/32 benchmarks)	Win-rate: 88.0% (44/50 benchmarks)	Win-rate: 77.8% (35/45 benchmarks)
Text	Win-rate: 78.6% (11/14 benchmarks)	Win-rate: 95.8% (23/24 benchmarks)	Win-rate: 84.2% (16/19 benchmarks)
Vision	Win-rate: 92.3% (12/13 benchmarks)	Win-rate: 95.2% (20/21 benchmarks)	Win-rate: 85.7% (18/21 benchmarks)
Audio*	Win-rate: 80% (4/5 benchmarks)	Win-rate: 60% (3/5 benchmarks)	Win-rate: 40% (2/5 benchmarks)

1.1.2 二代：谷歌Gemini更极限的上下文理解、更低延时

- **(低延时) Gemini 1.5 Flash:** 是为了满足用户对低延迟和低成本的需求而设计的轻量化模型。它针对大规模、大批量、高频的任务进行了优化，服务更具有成本效益。
- ✓ **更高效率和低延迟:** 在输入1万个字符的情况下，1.5 Flash在英法日中语言中实现了最快生成速度。
- ✓ **上下文:** 仍实现了100万tokens的长上下文窗口，开发人员还能注册尝试**200万tokens**。
- ✓ **性能:** 1.5 Flash在跨大量信息的多模态推理方面表现出色，适用于摘要、聊天应用、图像和视频字幕、长文档和表格的数据提取等多种任务。在核心能力测试中，相比1.0 Pro胜率更高，在视觉能力测试中，相比1.0 Ultra胜率更高。
- ✓ **价格:** 128K上下文窗口输入为0.35美元/百万tokens。
- ✓ **支持技术:** 这种强大性能来源于“蒸馏”技术，该技术将1.5 Pro中最重要的知识和技能转移到更小、更高效的模型中。1.5 Flash是一个**Transformer解码器模型**，旨在高效利用TPU，降低模型服务的延迟。例如，它可以并行计算注意力和前馈分量。它使用高阶预处理方法进行训练以提高质量。

图表：Gemini 1.5 Flash与Gemini 1.0对比

Gemini 1.5 Flash	Relative to 1.0 Pro	Relative to 1.0 Ultra
Long-Context Text, Video & Audio	from 32k up to 10M tokens	from 32k up to 10M tokens
Core Capabilities	Win-rate: 82.0% (41/50 benchmarks)	Win-rate: 46.7% (21/44 benchmarks)
Text	Win-rate: 94.7% (18/19 benchmarks)	Win-rate: 42.1% (8/19 benchmarks)
Vision	Win-rate: 90.5% (19/21 benchmarks)	Win-rate: 61.9% (13/21 benchmarks)
Audio	Win-rate: 0% (0/5 benchmarks)	Win-rate: 0% (0/5 benchmarks)

图表：Gemini 1.5 Flash查询的每个输出字符的平均时间最短 (ms)

Language	Gemini 1.5 Flash	Gemini 1.5 Pro	GPT-3.5 Turbo	GPT-4 Turbo	Claude 3 Haiku	Claude 3 Sonnet	Claude 3 Opus
English	1.5	4.3	2.6	6.8	2.2	6.2	10.5
Japanese	4.3	10.9	12.9	35.4	10.7	23.9	46.6
Chinese	5.2	14.1	18.4	53.4	12.8	30	55.3
French	1.9	4.7	4.1	10.7	3.0	6.8	13.4

1.1.3 翘楚：Claude3.5聚焦视觉和交互体验

- 6月，Anthropic发布下一代旗舰大模型**Claude 3.5 Sonnet**，领跑全行业智能水准，也保持中端模型性价比。
- ✓ 上下文&速度&价格：具有20万tokens上下文窗口。Claude 3.5 Sonnet的**运行速度是Claude 3 Opus的两倍**；输入\$3/百万tokens，输出\$15/百万tokens，**成本为Claude 3 Opus的五分之一**。
- ✓ 写作更自然：Claude 3.5 Sonnet在研究生水平推理GPQA、本科生水平推理MMLU和编程能力HumanEval方面树立了新的行业基准。在把握语义的细微差别、幽默和复杂指令上有显著改进，能以更自然、亲和的语气输出高质量的写作内容。
- ✓ 代码方面表现优异。能否根据文字需求改进代码的测试中，3.5 Sonnet成功解决了64%的问题，而3 Opus只解决了38%。只要给予清晰的指令和必要工具，它就能独立编写、编辑和执行代码，并具备复杂推理、故障排除与代码翻译能力。

图表：Claude 3.5 Sonnet在GPQA、MMLU等测试中性能基本领先

	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro	Llama-400b (early snapshot)
Graduate level reasoning GPQA, Diamond	59.4%* 0-shot CoT	50.4% 0-shot CoT	53.6% 0-shot CoT	—	—
Undergraduate level knowledge MMLU	88.7%** 5-shot	86.8% 5-shot	—	85.9% 5-shot	86.1% 5-shot
	88.3% 0-shot CoT	85.7% 0-shot CoT	88.7% 0-shot CoT	—	—
Code HumanEval	92.0% 0-shot	84.9% 0-shot	90.2% 0-shot	84.1% 0-shot	84.1% 0-shot
Multilingual math MGSM	91.6% 0-shot CoT	90.7% 0-shot CoT	90.5% 0-shot CoT	87.5% 8-shot	—
Reasoning over text DROP, F1 score	87.1 3-shot	83.1 3-shot	83.4 3-shot	74.9 Variable shots	83.5 3-shot Pre-trained model
Mixed evaluations BIG-Bench-Hard	93.1% 3-shot CoT	86.8% 3-shot CoT	—	89.2% 3-shot CoT	85.3% 3-shot CoT Pre-trained model
Math problem-solving MATH	71.1% 0-shot CoT	60.1% 0-shot CoT	76.6% 0-shot CoT	67.7% 4-shot	57.8% 4-shot CoT
Grade school math GSM8K	96.4% 0-shot CoT	95.0% 0-shot CoT	—	90.8% 11-shot	94.1% 8-shot CoT

图表：Claude 3.5 Sonnet在内部编码测试中领先于Claude 3模型家族

	Claude 3.5 Sonnet	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku
% of problems which pass all tests	64%	38%	21%	17%

* Claude 3.5 Sonnet scores 67.2% on 5-shot CoT GPQA with maj@32

** Claude 3.5 Sonnet scores 90.4% on MMLU with 5-shot CoT prompting

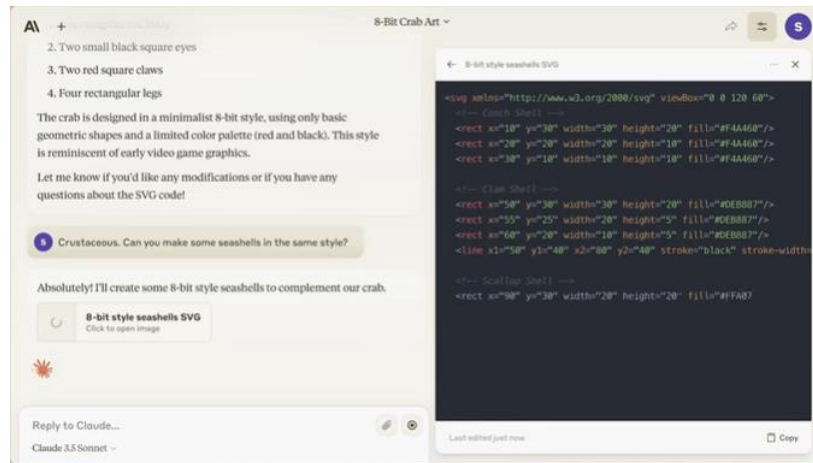
1.1.3 翘楚：Claude3.5聚焦视觉和交互体验

- ✓ **Claude 3.5 Sonnet**：Anthropic迄今为止最强大的视觉模型。在解释图表、图形等视觉推理任务中改进明显。可以准确地从粗略图像中转录文本，并输出更多洞察，这也是零售、物流和金融服务等领域的核心能力。
- ✓ **引入Artifacts变革交互方式**。当要求Claude生成代码、文本或网站设计等内容时，Artifacts会出现在对话旁边的专用窗口中，供用户实时查看、编辑和构建Claude的创作。相当于形成了一个动态工作空间，将AI生成的内容更无缝集成到自己的项目和工作流程中。这项功能标志着Claude从对话式AI向协作工作环境的演变，未来将拓展至团队协作中。

图表：Claude 3.5 Sonnet是Anthropic最强大的视觉模型

	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro
Visual math reasoning <i>MathVista (testmini)</i>	67.7% 0-shot CoT	50.5% 0-shot CoT	63.8% 0-shot CoT	63.9% 0-shot CoT
Science diagrams <i>AI2D, test</i>	94.7% 0-shot	88.1% 0-shot	94.2% 0-shot	94.4% 0-shot
Visual question answering <i>MMMU (val)</i>	68.3% 0-shot CoT	59.4% 0-shot CoT	69.1% 0-shot CoT	62.2% 0-shot CoT
Chart Q&A <i>Relaxed accuracy (test)</i>	90.8% 0-shot CoT	80.8% 0-shot CoT	85.7% 0-shot CoT	87.2% 0-shot CoT
Document visual Q&A <i>ANLS score, test</i>	95.2% 0-shot	89.3% 0-shot	92.8% 0-shot	93.1% 0-shot

图表：Claude 3.5 Sonnet引入Artifacts变革交互方式



01

海外龙头轮番抢占第一宝座， 比拼整体性能和多模态交互

1.1 高速迭代，海外龙头轮番占领大模型之巅

1.2 大模型高地争夺：多模态的理解和响应，原生多模态技术比拼

1.3 比较维度升级：从模型到叠加终端，跨设备的使用效果体验

1.2.1 最终判断依据：多模态的理解、生成和响应

- ▶ **多模态理解与生成，毫秒级响应，实现即时语音对话。** GPT-4o实现毫秒级视觉理解，GPT-4o能够接受文本、音频和图像的任意组合作为输入，并生成文本、音频和图像的任意组合输出。使用语音模式与ChatGPT对话当中，**GPT-3.5与GPT-4平均延迟分别为2.8s、5.4s，而GPT-4o对音频输入的响应时间最短为232毫秒，平均为320毫秒，这与人类在对话中的响应时间相似。**
- ▶ **更先进的视觉与音频理解能力，智能感知语气与语态。**与现有模型相比，GPT-4o展现了出色的视觉和音频理解能力：首先，用户可在对话中随时打断；其次，可根据场景生成多种音调，带有人类般的情绪和情感；直接通过和AI视频通话让它在线解答各种问题。

图表：GPT-4o在与技术人员实时对话



图表：GPT-4o实时感知技术人员情绪



图表：GPT-4o通过视频聊天解决图片中的数学问题

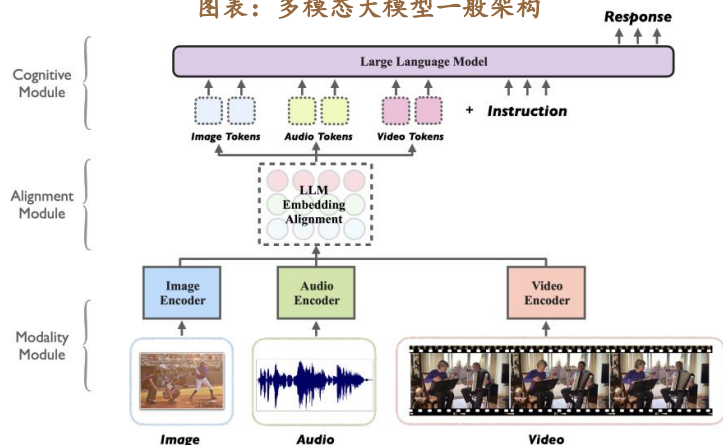


1.2.2 技术争夺：端到端原生多模态技术，统一神经网络

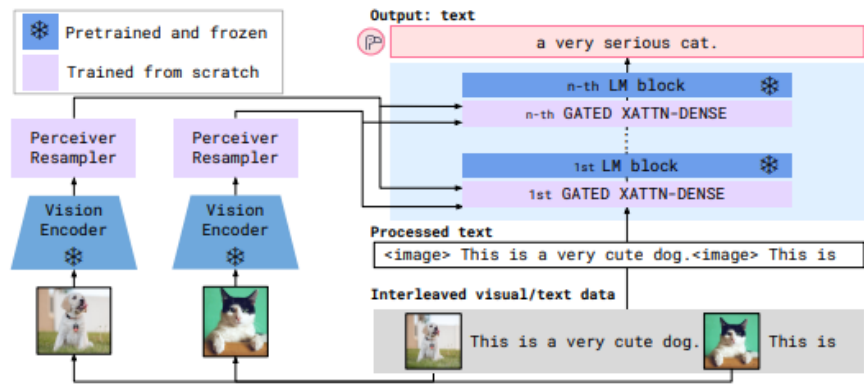


- 传统多模态大模型技术架构一般包括编码、对齐、解码等步骤，逐步整合多模态关联信息，输出目标结果。
- ✓ **编码**：包括视觉、音频、文本等模态编码器，目的是有效处理多个模态信息，转化为可处理状态；
- ✓ **对齐**：不同模态编码器可能不能直接融合，通过建立共同表示空间，将不同模态的表示统一，有效整合多个模态信息；
- ✓ **解码**：编码的反向过程，把模型的内部表示转化为物理世界的自然信号，即输出人类可识别的信息；
- ✓ **特点**：传统的多模态基础模型，通常为每种模态采用特定的编码器或解码器，将不同的模态分离开。
- ✓ **缺点**：限制了模型有效融合跨模态信息的能力。
- 以GPT-4为例，根据Semianalysis猜测，GPT-4多模态模型可能类似于Flamingo架构，它的文本编码器与视觉编码器是分开的，同时引入了交叉注意力机制。

图表：多模态大模型一般架构



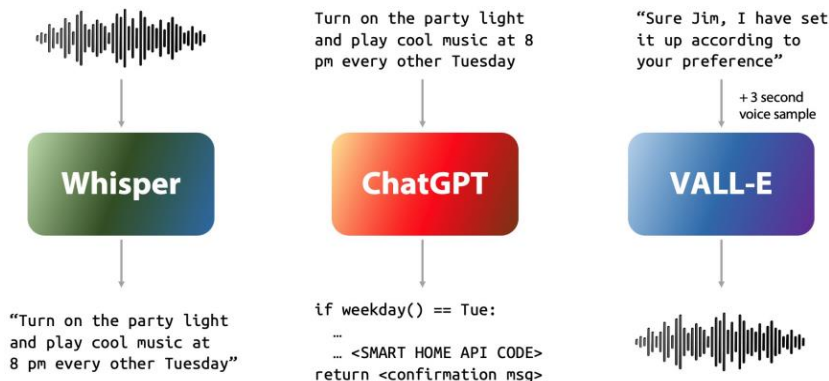
图表：Flamingo模型架构



1.2.2 技术争夺：端到端原生多模态技术，统一神经网络

- ▶ 在语音对话场景，传统语音AI通常经过三步法实现对话功能，在这过程中会丢失很多信息且不能判断情绪变化。三步法具体为：1) 语音识别或ASR：音频到文本，类似 Whisper；2) LLM 计划下一步要说什么：文本1到文本2；3) 语音合成或TTS：文本2到音频，类似ElevenLabs或VALL-E。GPT-4便采用该模式，在这过程中不仅响应速度更慢而且丢失了大量信息，无法直接观察语调、多个说话者或背景噪音，也无法输出笑声、歌唱或表达情感等。
- ▶ GPT-4o为跨模态（文本、视觉和音频）端到端训练新模型，意味着**所有输入和输出都由同一神经网络处理，成为真正的多模态统一模型**，带来性能的飞跃提升。

图表：传统语音AI“语音转文本-问答-文本转语音”三步法模式

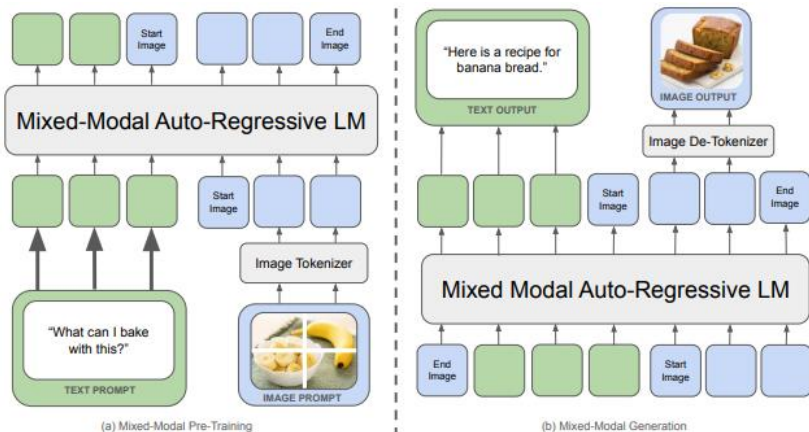


@DrJimFan

1.2.2 技术争夺：端到端原生多模态技术，统一神经网络

- 5月16日，Meta推出混合模态的基座模型Chameleon，与GPT-4o类似，不仅使用了端到端的方式从头开始训练，而且训练时将所有模态（文本、图像、代码）的信息交织混合在一起，并使用统一的Transformer架构处理。
- ✓ 数据处理：除了文本，图像也转化为一系列离散的tokens，最终生成交错的文本和图像tokens序列。
- ✓ 训练数据：训练数据既有纯文本、文本-图像对，也有文本、图像交错出现的多模态文档，共计10万亿tokens。
- ✓ 预训练：一开始就被设计为混合模型，使用统一的架构，以端到端的方式在所有模态（即图像、文本和代码）的交错混合上从头开始训练，而不需要单独的解码器或编码器。

图表：Chameleon模型架构：混合模态与训练和推理



图表：Chameleon训练数据集构成

	纯文本	Llama 2和CodeLlama所使用的 的所有预训练数据	2.9万个 token
第一阶段 (总训练量80%)	文本-图 像对	公开数据，共计14亿对	1.5万个 token
	文本图像 交错	公开数据来源	4000亿 token
第二阶段 (总训练量20%)	先将第一阶段得到的权重降低50%，并混合更高质量的数据让模型继续学习。		

1.2.2 技术争夺：端到端原生多模态技术，统一神经网络



- Chameleon实现了广泛的能力：
 - ✓ 在纯文本任务中保持竞争力：Chameleon-34B的性能在常识推理和阅读理解测试任务上与Mixtral 8x7B和 Gemini Pro等模型相匹配。
 - ✓ 在视觉问答和图像标注基准上：Chameleon-34B超过了Flamingo、IDEFICS和Llava-1.5等模型。

图表：Chameleon-34B文本能力与Gemini Pro等匹配

	Chameleon		Llama-2			Mistral		Gemini Pro	GPT-4
	7B	34B	7B	34B	70B	7B	8x7B	—	—
Commonsense Reasoning and Reading Comprehension									
PIQA	79.6	83.3	78.8	81.9	82.8	83.0	83.6	—	—
SIQA	57.0	63.3	48.3	50.9	50.7	—	—	—	—
HellaSwag	74.2	82.7	77.2	83.3	85.3	81.3	84.4	—	—
	75.6	85.1	—	—	87.1	83.9	86.7	84.7	95.3
	10-shot	10-shot			10-shot	10-shot	10-shot	10-shot	10-shot
WinoGrande	70.4	78.5	69.2	76.7	80.2	75.3	77.2	—	—
Arc-E	76.1	84.1	75.2	79.4	80.2	80.0	83.1	—	—
Arc-C	46.5	59.7	45.9	54.5	57.4	55.5	59.7	—	—
OBQA	51.0	54.0	58.6	58.2	60.2	—	—	—	—
BoolQ	81.4	86.0	77.4	83.7	85.0	84.7*	—	—	—
Math and World Knowledge									
GSM8k	41.6	61.4	14.6	42.2	56.8	52.1 maj08	74.4 maj08	86.5 maj032 CoT	92.0 SFT CoT
	50.9 maj08	77.0 maj032	—	—	—	—	75.1* maj032	—	—
MATH	11.5 maj01	22.5 maj01	2.5	6.24	13.5	13.1 maj04	28.4 maj04	32.6	52.9**
	12.9 maj04	24.7 maj04	—	—	—	—	—	—	—
MMLU	52.1	65.8	45.3	62.6	68.9	60.1	70.6	71.8	86.4

图表：Chameleon-34B视觉问答和图像标注刷新SOTA

	Model	Model Size	COCO	Flickr30k	VQAv2
Pre-trained	Flamingo-80B	80B	113.8 32-shot	75.1 4-shot	67.6 32-shot
	IDEFICS-80B	80B	116.6 32-shot	73.7 4-shot	65.9 32-shot
Chameleon	Chameleon	34B	120.2 2-shot	74.7 2-shot	66.0 2-shot
	Chameleon-SFT	34B	140.8 0-shot	82.3 2-shot	—
	Chameleon-MultiTask	34B	139.1 2-shot	76.2 2-shot	69.6
Fine-tuned	Flamingo-80B-FT	80B	138.1	—	82.0
	IDEFICS-80B-Instruct	80B	123.2 32-shot	78.4 32-shot	68.8 32-shot
Closed Source (finetuning status unknown)	GPT-4V	—	78.5* 8-shot	55.3* 8-shot	77.2
	Gemini Nano 2	—	—	—	67.5
	Gemini Pro	—	99.8* 2-shot	82.2* 4-shot	71.2
	Gemini Ultra	—	—	—	77.8

1.2.2 技术争夺：端到端原生多模态技术，统一神经网络

- 为了进一步评估模型生成多模态内容的质量，论文在基准测试之外引入了人类评估实验：
- ✓ Chameleon在混合模态推理和生成方面提供的全新功能：在开放式问题（即混合图像和文本的问题）的混合模态响应质量方面，人类评估者更喜欢Chameleon模型，而不是Gemini Pro和GPT-4V。它可回答包含文本和生成图像的问题。

图表：Chameleon生成的图像和文本交织的内容

Prompt: I'm bored. Could you show me some cool, quirky-looking birds? Also tell me about them in 2-3 sentences.



Response: Absolutely! Here are three birds with some unique features that you might find interesting:

Keel-Billed Toucan The Keel-Billed Toucan is a brightly colored bird native to South American rainforests and woodlands. Its most distinctive feature is its colorful beak, which can grow up to seven inches long and has a distinctive, curved shape. The beak is used for plucking fruit from trees and for social displays. Despite its large size, the Keel-Billed Toucan is a skilled flyer and can often be seen flitting through the treetops. Here is an image of this vibrant bird:

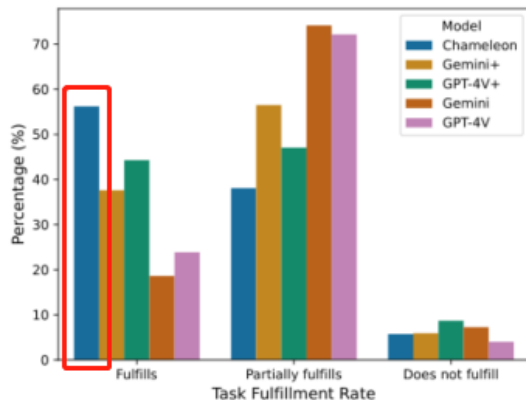
Puffin Puffins are small, aquatic birds that are found in the northern hemisphere. They have distinctive black and white feathers, as well as a brightly colored beak that is used for attracting mates and defending their territory. Puffins are known for their distinctive "puffin" call, which sounds like a loud, guttural chuckle. They feed on fish, which they catch by diving into the water from the air. Puffins are colonial birds, and they breed in large colonies on rocky cliffs and islands. Here is an image of this adorable bird:



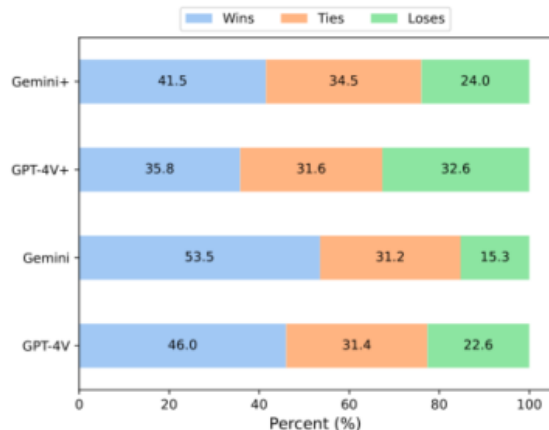
Golden Pheasant The Golden Pheasant is a brightly colored bird native to the forests of China. It is a member of the pheasant family and is known for its striking plumage, which features a golden-yellow body, red face, and green tail feathers. The Golden Pheasant is a ground-dwelling bird, and it feeds on a diet of insects, seeds, and fruits. It is a popular game bird, and is often hunted for its meat and feathers. In captivity, the Golden Pheasant is sometimes kept as an ornamental bird, due to its striking appearance. Here is an image of this beautiful bird:

I hope these images and descriptions have helped to brighten your day! If you have any other questions or requests, please don't hesitate to ask.

图表：在混合模态响应质量方面，人类评估者更喜欢Chameleon



(a) The prompt task fulfillment rates.



(b) Chameleon vs. the baselines: Gemini+, GPT-4V+, Gemini, GPT-4V.

01

海外龙头轮番抢占第一宝座， 比拼整体性能和多模态交互

1.1 高速迭代，海外龙头轮番占领大模型之巅

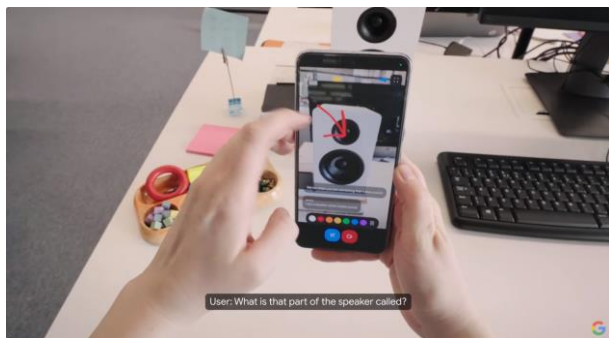
1.2 大模型高地争夺：多模态的理解和响应，原生多模态技术比拼

1.3 比较维度升级：从模型到叠加终端，跨设备的使用效果体验

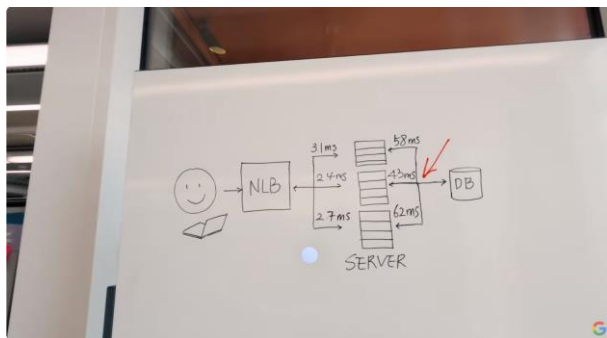
1.3 比较维度升级：叠加终端，跨设备的使用效果体验

- 谷歌推出AI Agent项目Astra模型，具备类似GPT-4o的能力，可以实现跨文本、音频、视频多模态实时推理。Astra和GPT-4o的使用效果类似，用户可以通过它和AI实时对话，以及视频聊天。在发布会上，工作人员在演示视频中将手机镜头对准身边的物品，并向Project Astra提出一些疑问，它几乎能做到零延时地准确回答。
- **Astra特点：**1) 利用用户设备上的摄像头和麦克风为日常活动提供无缝辅助；2) 具备多模态能力，无缝处理音频、图像、视频和文本输入，为用户提供全面的体验；3) 深度理解视觉内容，并具备**内容回溯能力**；4) **语调丰富**。
- **Astra VS. GPT-4o：**两者均具备音频、图像、视频等多模态处理能力，不同的是，**GPT-4o**专注于对多模态输入的实时交互，**Astra**专注于通过外界设备无缝捕获和解释视觉信息，使其能够根据现实世界的观察提供与上下文相关的响应。

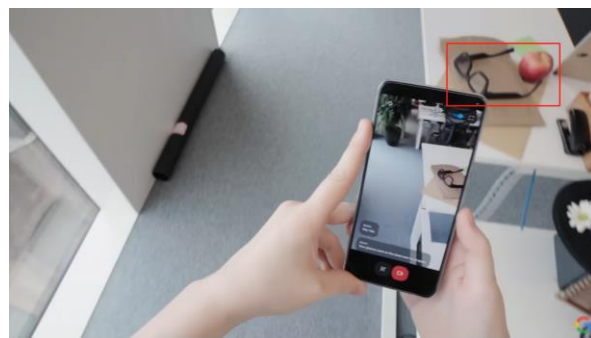
图表：工作人员使用手机与Astra互动



图表：工作人员佩戴眼镜后与Astra互动



图表：Astra回溯视频内容找回眼睛

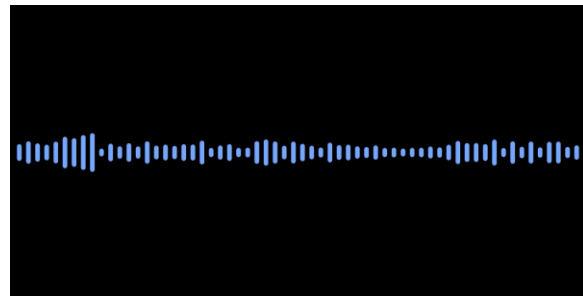


1.3 比较维度升级：叠加终端，跨设备的使用效果体验

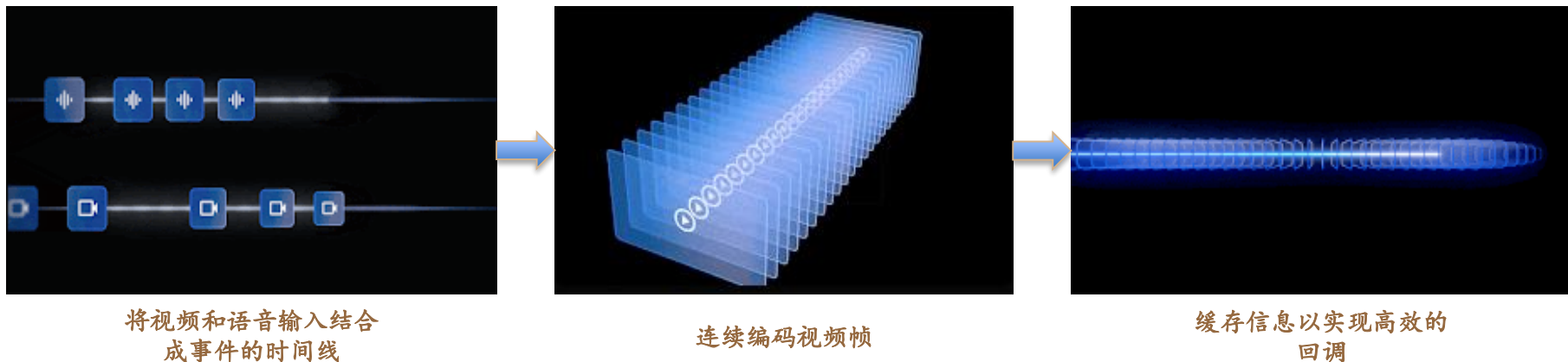
➤ Astra将视频帧和语音编码到时间轴。

- ✓ 技术支持：Astra首先通过智能手机摄像头从网络以及周围的世界中提取信息，之后通过连续处理和编码视频帧和语音输入，创建事件的时间线，并缓存信息以便快速回调。它可以识别物体、回答询问，甚至还能记住摄像头视角内不再存在的过往画面。简而言之，它通过智能手机摄像头观察您周围的世界，感知、响应并记住它。
- ✓ 此外，谷歌还利用领先的语音模型，增强了Astra的发音，使其具有更广泛的语调。Astra可以更好地理解他们所处的语境，并在对话中快速做出反应。

图表：Astra具有广泛的语调



图表：Astra内部原理



02

国内大模型逆袭之路：聚焦长文本，降价迭代提升竞争力

2.1 先文后理：聚焦长文本，国产大模型已有赶超GPT之势

2.2 长文本的三大难题和解决之道

2.3 商业上降价，加速迭代卷出未来

2.1 先文后理：理科能力差距较大，注重文科能力的提升

图表：国内外最新大模型的综合能力比较

模型名称	总分	理科					文科				
		计算	逻辑推理	代码	工具使用	知识与百科	长文本	角色扮演	语义理解	生成与创作	传统安全类
GPT-4-Turbo-0125	79.13	82.8	75.2	87.9	78.6	81.6	69.8	76.2	79.8	73.4	86
GPT-4-Turbo-0409	77.02	81.6	71.2	87.6	76.6	79	67	72.8	77.6	72.6	84.2
GPT-4(官网)	75.32	78.6	71.6	87.8	74.8	72.8	66.4	71.8	76.4	72	81
Claude3-Opus	74.47	76.8	61.2	84.5	74.0	77.4	66.6	74	75.8	70.8	83.6
Baichuan3	73.32	74.8	68.6	74.8	74.4	82	64.6	70.8	73.6	71.4	78.2
GLM-4	72.58	71	64.6	72.6	74.6	78	64.2	73.6	75.2	71	81.0
通义千问2.1	72.45	71	59.6	53.3	75.6	81	71.8	78.8	75.4	75.6	82.4
腾讯Hunyuan-pro	72.12	71	62.6	63.6	74.0	78.4	66.4	73.6	75.4	70.6	85.6
文心一言4.0	71.9	70.6	64.6	73.8	72.4	77	60.2	71	74.8	71.4	83.2
MoonShot(Kimi)	70.42	71.4	56.2	61.6	70.6	79.8	68.6	72.6	65.8	72.2	85.4
从容大模型V1.5	70.35	75.8	57.8	54.1	71.0	80.6	64.2	72	74.8	70.2	83.0
MiniMax-abab6.1	70.18	71.4	56.4	65.2	68.6	75.8	66.2	68.8	73.8	67.2	88.4
山海大模型	69.51	68.2	54.6	58.1	69.2	78.2	68.2	73.4	74.8	72.6	77.8
阶跃星辰step-1-32k	68.69	64.8	55.2	60.3	69.2	79.2	61.6	70.8	73.2	68.8	83.8
qwen-1.5-72b-chat	68.07	68.2	52.8	51.7	69.0	77.4	67	72.6	74.2	71	76.8
360gpt-pro	66.6	63.2	55.2	42.2	68.8	74.4	63.6	69.8	73.8	70.6	84.4
qwen-1.5-14b-chat	63.51	62.6	45.4	34.7	65.4	72.4	63.4	69	72.8	70.6	78.8

➤ 整体比较而言，国内大模型与GPT-4（官网）尚存在明显差距，但个别能力上已展现出优势。

➤ 理科能力差距明显。除通义千问2.1的工具使用能力较高，其它国内大模型的得分均低于GPT-4（官网）。

➤ 文科能力差异缩小。GPT-4在语义理解中保持优势，国内大模型在其他能力上表现出色。

✓ 通义千问2.1：在知识百科、长文本、角色扮演和生成与创作等具有突出优势，得分较GPT-4高。

✓ Kimi：kimi在长文本能力上表现出色，与GPT-4的较量中得分占优。

2.1 先文后理：国内大模型长文本能力部分赶超GPT-4

图表：国内外最新大模型的长文本能力比较

- 大模型的长文本能力已经成为了重要的竞争力，国内长文本能力赶超了部分国外大模型。
- 国内大模型厂商布局长文本能力。长文本能力被视为大模型竞争的重要性能指标。通义千问、360的智脑、文心一言开放长文本能力。
- 国内头部企业展现出较强的竞争力。根据基准测评报告，通义千问凭借71.8分成为全球大模型中唯一超过70分的大模型，而kimichat和山海大模型则占据第四和第五的席位。在前32名中，有20个席位被国内大模型占据。

模型	类型	长文本能力	模型	类型	长文本能力
通义千问2.1	国内前三	71.8	GPT3.5-Turbo-0125	海外模型	61.2
GPT-4-Turbo-0125	海外模型	69.8	文心一言4.0	其它	60.2
MoonShot(kimichat)	国内前三	68.6	云雀大模型	其它	59.8
山海大模型	国内前三	68.2	Llama-3-70B-Instruct(poe)	海外模型	59
GPT-4-Turbo-0409	海外模型	67	qwen-1.5-7B-Chat	其它	58
qwen-1.5-72b-chat	其它	67	Gemini-Pro	海外模型	56.8
Claude3-Opus	海外模型	66.6	Llama-3-70B-Instruct(千帆)	海外模型	55.8
GPT-4(官网)	海外模型	66.4	Llama-3-8B-Instruct	海外模型	54.4
Hunyuan-pro-32K-0423	其它	66.4	讯飞星火V3.5	其它	54.2
MiniMax-abab6.1	其它	66.2	ChatGLM3-6B	其它	49.2
Baichuan3	其它	64.6	XVERSE-13B-L	其它	44.4
GLM-4	其它	64.2	Gemma-7b-it	海外模型	44
从容大模型V1.5	其它	64.2	Chinese-Alpaca2-13B	其它	12.2
360gpt-pro	其它	63.6	Baichuan2-13B-Chat-v2	其它	3.4
qwen-1.5-14b-chat	其它	63.4	Llama2-7B-Chat	海外模型	0.8
阶跃星辰step-1-32k	其它	61.6	Llama2-13B-Chat	海外模型	0.2

2.1 先文后理：国内大模型长文本能力部分赶超GPT-4

图表：国内外主流大模型的长文本能力比较

- ▶ **Kimi率先在国内掀起大模型长文本竞赛。**3月18日，月之暗面宣布在大模型长上下文窗口技术上取得突破：Kimi智能助手从支持20万字直接到支持200万字超长无损上下文，并于即日起开启产品“内测”。受此消息影响，百度、阿里等国产大模型龙头厂商陆续宣布大模型迭代进展，打响百万级长文本上下文竞赛。3月22日晚间，阿里的通义千问官宣升级，向所有人免费开放1000万字的长文档处理功能；360紧随其后，360智脑正式内测500万字长文本处理功能，即将入驻360AI浏览器；百度在4月将对文心一言进行升级，届时也将开放长文本能力，文本范围在200万-500万。
- ▶ **对比海外龙头：**1) 5月谷歌发布的最新Gemini 1.5 pro、1.5 flash分别支持200万（理论极限1000万）、100万（开发人员可尝试注册200万）token；2) 6月，Claude 3.5 Sonnet支持20万token；3) GPT-4 Turbo支持12.8万token。

公司/机构/团队	模型/产品名称	上下文Tokens	对应汉字
OpenAI	GPT-3.5	4-16k	3.2k-12.8k
	GPT-4	8-32k	6.4k-25.6k
	GPT-4 Turbo	128k	102.4k
Google	Gemini 1.0 pro	32k	25.6k
	Gemini 1.5 pro	10000k	8000k
	Gemini 1.5 flash	2000k	1600k
Anthropic	Claude	100k	80k
	Claude2	100k	80k
	Claude3	200k	160k
	Claude 3.5 Sonnet	200k	160k
Meta	LLaMA	2k	1.6k
	LLaMA2	4k	3.2k
	Llama 2 Long	32k	25.6k
IDEAS NCBR、Google DeepMind等	Loag LLaMA	256k	204.8k
Moonshot	Kimi Chat		200w
阿里	通义千文		1000w
百度	文心一言		200-500w
360	360智脑		500w
字节	豆包通用模型pro	128k	102.4k
智谱	GLM-3-Turbo	128k	102.4k

注：统计时间截至2024年6月30日；对应汉字未标红的数据按照1 token=0.8个汉字测算；Gemini 1.5 pro、flash为理论极限值。

02

国内大模型逆袭之路：聚焦长文本，降价迭代提升竞争力

2.1 先文后理：聚焦长文本，国产大模型已有赶超GPT之势

2.2 长文本的三大难题和解决之道

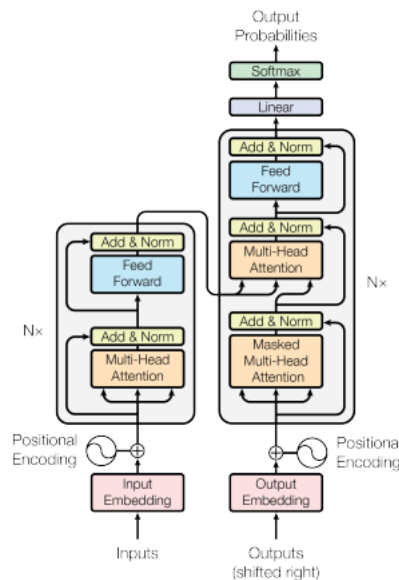
2.3 商业上降价，加速迭代卷出未来

2.2 长文本的三大难题和解决之道

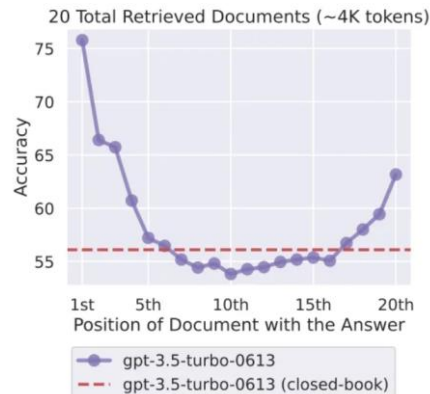
➤ **Transformer**架构的大模型在长文本中存在注意力机制计算复杂度、上下文记忆以及最长文本约束难题：

- ✓ **(1) 注意力机制计算复杂度：**注意力机制的计算时间复杂度、空间复杂度都是序列长度的二次方 $O(L^2)$ ，随着序列变长，存在计算速度变慢和内存增长的问题，模型的训练和推理均存在负担。
- ✓ **(2) 上下文记忆：**LLM缺乏显性的内存机制，仅依靠KV缓存来存储以前所有token之间的信息，在结束了一次查询调用操作后，除非将历史上所有token的信息重新加载到KV缓存中，否则transformer不会保存该次查询的信息。因而，每次调用时，模型只拥有较短的上下文工作内存，而不具备长期记忆。这虽然在并行计算方面有优势，但在对长期记忆有要求的聊天机器人场景中存在应用困难。
- ✓ **(3) 最长文本约束：**在训练时，由于GPU内存的限制，工程师通常会确定最长的超参数长度（如1K、2K、4K等，表示任意批训练中样本序列长度的限制），随着序列变长，模型性能可能下降。在推理过程中，Transformer在访问长文本中间时性能也会显著下降，LLM服务商通常会限制用户prompt上下文长度，以保证模型性能的稳定。

图表：Transformer模型一般架构



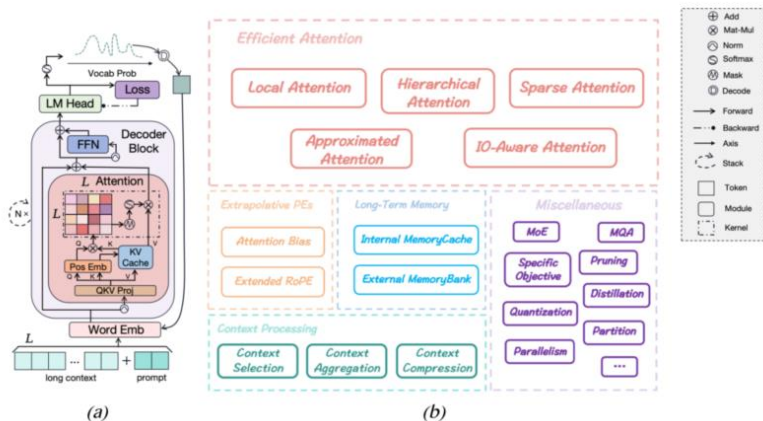
图表：访问长文本中间性能下降



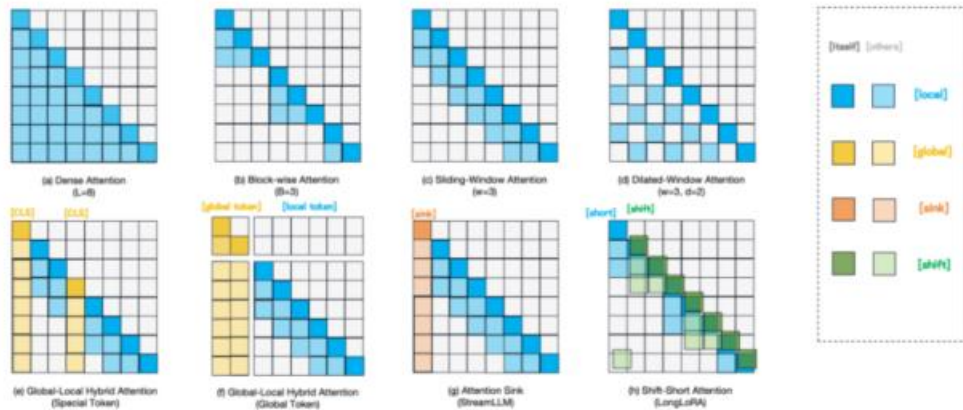
2.2 长文本的三大难题和解决之道

- 突破长文本限制的方法尚未收敛。
- 目前，突破长文本的方式主要涉及优化注意力机制、长期记忆力机制、上下文处理以及位置编码外展等。
- ✓ **优化注意力机制**：这类方法的重点是实现有效的注意机制与降低计算成本，甚至实现线性时间的复杂性，从而可以增加训练前阶段的上下文长度以及推理过程中的有效上下文长度边界。可以分成五种不同的策略，每种策略都有一个特定的重点：**局部注意**、**分层注意**、**稀疏注意**、**近似注意**和**IO-觉察注意**。例如，局部注意是将每个token的注意仅限于其相邻的token，而不是每个token与其他所有的token都计算attention。

图表：长文本问题的解决之道



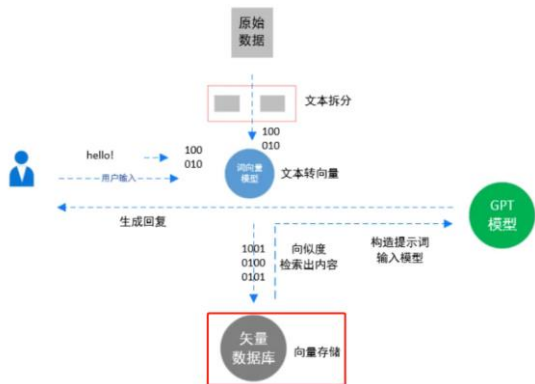
图表：局部注意力机制原理图



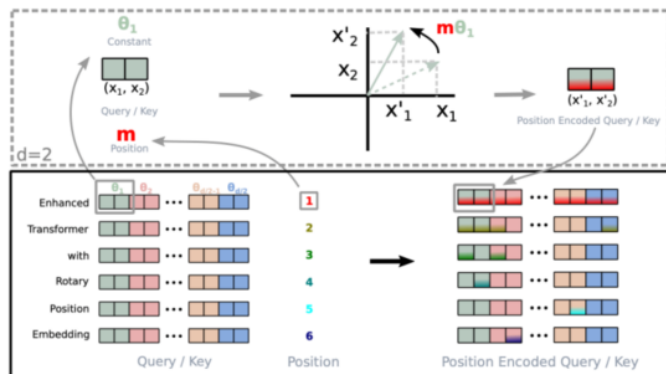
2.2 长文本的三大难题和解决之道

- ✓ **长期记忆力机制**：通过引入额外的记忆模块来扩展上下文窗口。1) 内部记忆力缓存：利用递归等方法，将长文本分成固定长度的片段流，此上下文信息是从先前片段的缓存或提炼出的信息中获得的；2) 外部知识库：在推理过程中，模型可从外部知识库中读取内容以丰富其上下文输入，并从用户的响应中写入这些知识库以刷新其长期记忆，例如RAG技术。
- ✓ **位置编码外展**：通过设计可推广的位置嵌入方法来处理不同长度的输入，实现更长的推理长度，例如使用注意力偏差和扩展的RoPE（旋转位置编码），实现上下文长度外推。
- ✓ **上下文处理**：在预处理和后处理阶段优化长文本处理，将预训练的LLM视为黑盒或灰盒模型，并通过多次调用模型来处理长上下文输入，确保每次调用都遵守 L_{max} 限制。这些方法并没有增强LLM处理长上下文的固有能力，只是利用模型的上下文学习能力。例如将上下文分段、将分段的信息融合以及压缩上下文的大小。
- ✓ **其他方法**：包括特定的优化目标、MoE、并行计算、压缩权重等其他方法，以提高LLMs在各个阶段的性能。

图表：RAG技术原理图



图表：旋转位置嵌入 (RoPE) 原理图



02

国内大模型逆袭之路：聚焦长文本，降价迭代提升竞争力

2.1 先文后理：聚焦长文本，国产大模型已有赶超GPT之势

2.2 长文本的三大难题和解决之道

2.3 商业上降价，加速迭代卷出未来

2.3.1 降价抢占API调用量，撬动大模型“飞轮迭代”

- 随着技术进步和市场竞争，大模型训练&推理成本降低，国内大模型厂商纷纷降价，以吸引用户和提高市场份额。
- ✓ 国内大模型降价从初创公司开启，由云厂商加速，更多初创公司并未降价。根据智东西微信公众号统计，5月6日，深度求索开源MoE模型DeepSeek-V2，百万tokens仅需1元。5月15日，字节宣布旗下通用模型豆包pro-128k版模型推理输入价格定价比行业价格低95.8%，豆包pro-32k模型推理输入降至比行业价格低99.3%。之后阿里、百度、科大讯飞、腾讯等云厂商或科技企业相继加入降价队伍。对初创公司而言，除了深度求索、智谱AI之外，包括百川智能、月之暗面、零一万物在内的几家头部大模型初创公司并未加入降价行列。
- ✓ 高幅度降价以轻量级、入门级模型为主，主力模型、旗舰模型降幅相比较小。
- ✓ 我们认为，降价不等于恶性竞争和模型缺陷，更多的是在技术支持下商业逻辑的打磨与模型能力的完善。云厂商降价行为更活跃，与其具备更完善的云算力基础设施息息相关，通过降价的方式积累更多的用户，进一步抢占市场份额。

图表：国内外主流大模型价格情况

公司	模型型号	降价/定价时间	降价前 (元/百万token)		降价后/初次定价 (元/百万token)		降价幅度	
			输入	输出	输入	输出	输入	输出
OpenAI	gpt-4o	5月13日首次定价			36	109		
	gpt-4-turbo	-			73	218		
	gpt-4	-			218	436		
Google	gemini 1.5 pro	5月15日首次定价			51/25 (128k上下文)	153/76 (128k上下文)		
	gemini 1.5 flash	5月15日首次定价			5/3 (128k上下文)	1527/8 (128k上下文)		
Anthropic	Claude 3.5 Sonnet	6月21日首次定价			22	109		
深度求索	DeepSeek V2	5月6日首次定价			1	2		
智谱AI	GLM-4-0520	6月5日首次定价			100	100		
	GLM-3 Turbo	5月11日	5	5	1	1	80%	80%
字节跳动	豆包-pro-128k				5	9		
	豆包-pro-4k/32k	5月15日首次定价			0.8	2		
	豆包-lite-128k				0.8	1		
	豆包-lite-4k/32k				0.3	0.6		

注：统计时间及汇率截至2024年6月28日；加粗的为大模型厂家的旗舰型号。

2.3.1 降价抢占API调用量，撬动大模型“飞轮迭代”

图表：国内外主流大模型价格情况（续表）

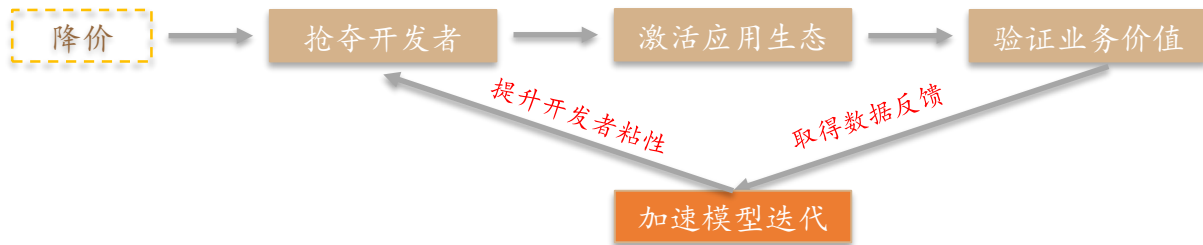
公司	模型型号	降价/定价时间	降价前（元/百万token）		降价后/初次定价（元/百万token）		降价幅度	
			输入	输出	输入	输出	输入	输出
阿里云	Qwen-Max	5月21日	120	120	40	120	67%	
	Qwen-Long		20	20	0.5	2	98%	90%
	Qwen-Plus		20	20	4	12	80%	40%
	Qwen-Turbo		8	8	2	6	75%	25%
百度	ERNIE 4.0系列	5月21日	120	120	120	120		
	ERNIE 3.5系列		12	12	12	12		
	ERNIE Speed 8K/128K		4/16	8/32	0	0	100%	100%
	ERNIE Lite 8K/128K		3/12	6/24	0	0	100%	100%
科大讯飞	Spark 3.5 Max	5月22日	21-30	21-30	21-30	21-30		
	Spark Pro		21-30	21-30	21-30	21-30		
	Spark Lite		18	18	0	0	100%	100%
腾讯云	混元-pro	5月22日	100	100	30	100	70%	
	混元-standard-256k		120	120	15	60	88%	50%
	混元-standard		10	10	4.5	5	55%	50%
	混元-lite		8	8	0	0	100%	100%
未降价的公司旗舰模型定价情况								
百川智能	Baichuan4				100	100		
	Baichuan3-Turbo				12	12		
零一万物	yi-large				20	20		
	yi-large-turbo				12	12		
月之暗面	moonshot-v1-128k				60	60		
	moonshot-v1-32k				24	24		
	moonshot-v1-8k				12	12		

注：统计时间及汇率截至2024年6月28日；加粗的为大型模型厂家的旗舰型号。

2.3.1 降价抢占API调用量，撬动大模型“飞轮迭代”

- 大模型降价是对API调用量的“跑马圈地”。
- ✓ 短期来看大模型性能提升遇到瓶颈，同质化严重，包括OpenAI的用户增速陷入了低迷期，降价是吸引更多开发者参与进来最直接的方法。
- ✓ 智东西微信公众号6月20日报道，从阿里、百度、腾讯、字节等大模型厂家的后台数据来看，在宣布降价后，各家主力模型在最近一个月调用量均大幅上扬——**增长数倍到十几倍都有**，甚至有个别客户如字节调用量在降价两周内增长**5000倍**。智谱AI、科大讯飞、DeepSeek等大模型厂商虽未公开相关数据，但调用量可以预测也是水涨船高。
- 我们认为，调用量提升有望**撬动大模型“飞轮迭代”，打磨出更好的模型**。
- ✓ 调用量提升能够**抢夺更多的开发者→激活更多的应用场景与生态→验证大模型业务价值→加速迭代打磨出更好的模型→增强开发者粘性**。对于大模型企业而言，只有吸引越多的开发者，才能缔造更繁荣的应用生态、催生更多的应用创新。应用生态越完善，使用场景越多，用户规模越大，生成的新数据会反哺大模型性能提升。
- ✓ 在这过程中，开发者是核心角色。既可能通过反复调用模型打造出应用生态并提供了模型优化建议，而且可能在开发出应用后与大模型生态捆绑，从而增强大模型粘性。

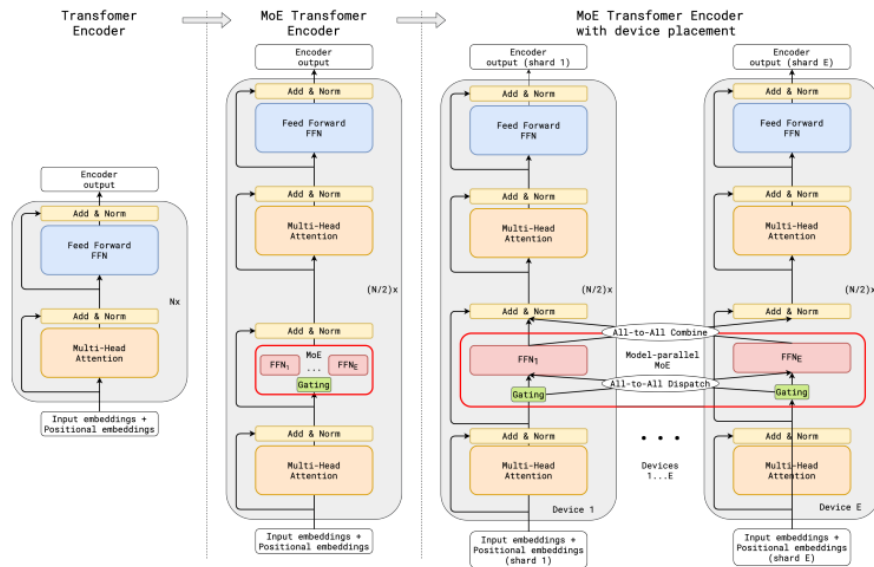
图表：大模型降价提升调用量有望撬动大模型“飞轮迭代”



2.3.2 降价的背后是训练&推理成本的下降

- **训练成本下降：改善大模型训练工程与提高训练效率与算力利用率。**
- **大模型工程改善技术：**在Scaling Law驱使下，大模型迫切变大，却面临着训练成本的难题。目前，国内外主流大模型GPT-4、Gemini、Mistral、天工3.0、DeepSeek V2等均采用了**MoE架构**，在模型扩展性与训练成本直接达到了较好的平衡。
- ✓ **训练成本指数级增长是Dense模型扩展的重要难题。**根据seminianalysis，训练一个1万亿参数模型的成本约为3亿美元，如果在1.25万个HGX/DGX系统中使用10万个A100，则需要大约3个月的时间进行训练；训练10万亿参数模型则需要将近300亿美元，即使在1.25万个HGX/DGX系统中使用100万个A100，也需要两年多的时间。
- ✓ **MoE技术原理：**与传统大模型架构相比，MoE架构在数据流转过程中集成了一个专家网络层，该层的核心由门控网络和一组专家模型构成。数据进入MoE时会先通过门控网络分割数据，将每组数据分配给一个或多个专家，最终输出由所有专家结果的加权融合。
- ✓ **效果：**与传统的Dense模型相比，**MoE能够在远少于前者所需的计算资源下进行有效的预训练**，计算效率更高、速度更快，进而使得模型规模得到显著扩大，让**万亿参数成为可能**，获得更好的AI性能。2022年，Google采用MoE架构设计的Switch Transformers模型参数达到1.57万亿，与此前的T5模型相比，在相同的计算资源下获得高达7倍的模型预训练速度提升，并实现了4倍的模型加速。

图表：MoE一般基础架构



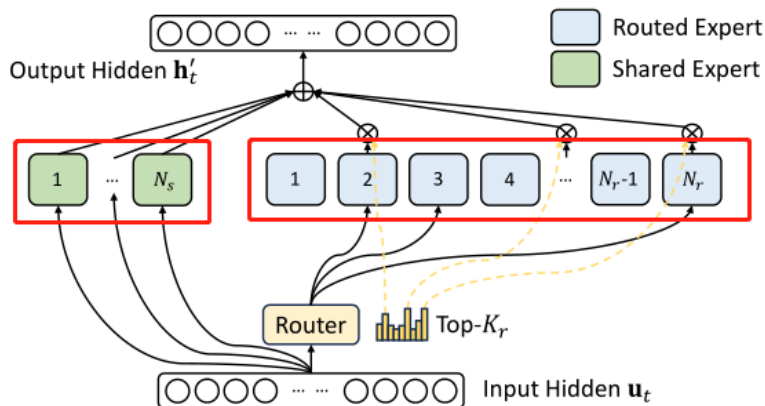
图表：Dense模型随着参数增长成本指数级扩张

Chinchilla Optimal LLM Training Cost					
Model	Size (# Parameters)	Tokens	GPU	Optimal Training Compute Cost	
1 Billion Parameter	1 Billion	20 Billion	A100	\$	294
10 Billion Parameter	10 Billion	205 Billion	A100	\$	29,866
67 Billion Parameter	67 Billion	1,500 Billion	A100	\$	1,398,601
175 Billion Parameter	175 Billion	3,700 Billion	A100	\$	9,348,291
280 Billion Parameter	280 Billion	5,900 Billion	A100	\$	24,038,462
520 Billion Parameter	520 Billion	11,000 Billion	A100	\$	83,284,771
1 Trillion Parameter	1,000 Billion	21,200 Billion	A100	\$	308,372,183
10 Trillion Parameter	10,000 Billion	216,200 Billion	A100	\$	28,935,185,185

2.3.2 降价的背后是训练&推理成本的下降

- ✓ **MoE存在进一步改进空间。**例如，国内DeepSeek V2采用DeepSeekMoE混合专家架构，通过细粒度的专家分割（更多的专家）和共享专家隔离（专家之间共享知识，减少知识冗余）等优化措施，实现了比传统MoE架构更低的训练成本。与初代DeepSeek 67B相比，新模型训练成本下降42.5%。
- **提高训练效率与算力利用率：**大模型训练方法与算力基础设施能力的综合比拼。
- ✓ 大模型训练中存在一些的分布式并行范式可以有效优化训练过程，分别为**数据并行**、**流水线并行**和**张量并行**，通过训练任务的拆分、提高内存利用率等多种方式减少训练时间，提高训练效率。例如数据并行中ZeRO（Zero Redundancy Optimizer），思想就是拆分参数、梯度及优化器状态，使得节点保存部分参数、梯度及优化器状态，可以减少内存占用并提升一定的通信速率。
- ✓ 算力提供方可以通过一些基础工具提高算力利用率。例如，英伟达提供了Nsight System (nsys) 工具查看模型运行时的CPU/GPU执行情况，可以帮助发现GPU运行时的一些问题。

图表：MoE一般基础架构



图表：ZeRO方法减少了内存占用、提升通信速率

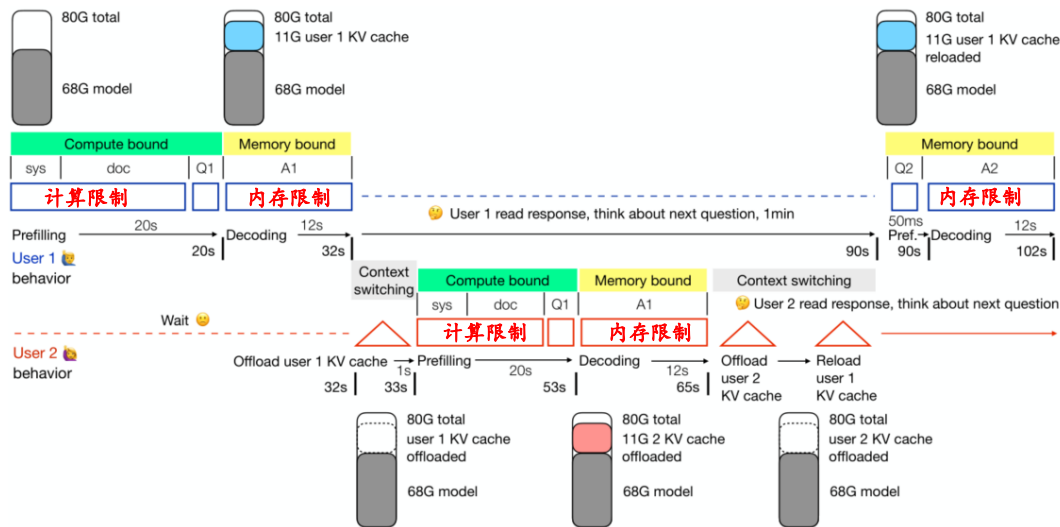
M is the number of parameters, N is the number of devices.

	Optimizer States (12M)	Gradients (2M)	Model Weights (2M)	Memory Cost	Communication Cost
Data Parallelism	Replicated	Replicated	Replicated	$16M$	all-reduce(2M)
ZeRO Stage 1	Partitioned	Replicated	Replicated	$4M + \frac{12M}{N}$	all-reduce(2M)
ZeRO Stage 2	Partitioned	Partitioned	Replicated	$2M + \frac{14M}{N}$	all-reduce(2M)
ZeRO Stage 3	Partitioned	Partitioned	Partitioned	$\frac{16M}{N}$	1.5 all-reduce(2M)

2.3.2 降价的背后是训练&推理成本的下降

- **推理成本下降**：目前大模型普遍支持长文本，而GPU HBM、带宽以及PCIe带宽大小有限，大模型落地应用实现规模化推广后，KV缓存给推理成本带来较大挑战，相应的成本优化路线主要围绕压缩KV缓存展开。
- ✓ **拆解多个用户同时请求长文本问题时大模型推理过程**（以左图模型与GPU条件为例），其中并发用户的数量 = (HBM大小-模型权重) / **KV缓存**：
- ✓ 1) **预填充阶段**：用户1向模型上传文件并提问，模型预填充为**KV缓存**，用户2处于等待状态，**响应时间**=上下文FLOP/GPU算力，主要受GPU计算能力的限制。
- ✓ 2) **解码阶段**：模型逐步向用户1输出token，受存储带宽限制，单位token**响应时间**=内存访问字节数/GPU HBM带宽，其中内存访问字节数=模型权重+**KV缓存**的字节数。
- ✓ 3) **切换用户阶段**：卸载用户1的KV缓存至CPU，同时填充用户2的KV缓存，其中CPU与GPU交流速度由PCIe决定。**响应时间**=(用户1+用户2的**KV缓存**) / PCIe带宽。
- ✓ 4) 重复以上过程。
- 在这过程中，可以看到**KV缓存的大小**同时影响了用户**并发数量、解码响应时间以及切换用户时间**，对模型的**推理成本和推理效率**具有重要影响。

图表：在有限GPU HBM下多个用户长文本推理请求过程图



注：该图基于80GB-A100 NVLink，模型参数大小34B（占用68GB），单用户请求50K上下文长度（占用11GB），并假设算力利用率为较高水平。

2.3.2 降价的背后是训练&推理成本的下降

➤ 围绕KV缓存的公式，压缩KV缓存可以通过transformer架构的层数、自注意力机制的头数、输入的token数以及隐藏层的维度数量的优化方式去实现。

- ✓ 1) transformer架构的层数：主要假设是，有些任务可能不需要全深度计算，**预填充阶段就可以跳过某些层**，这样可以同时减少预填充失败的概率和KV缓存的大小。
- ✓ 2) 自注意力机制的头数：主要假设是，一些头是专门为检索和长上下文相关的能力，那么**可能保留检索头和删除其他头**，从而减少KV缓存大小。
- ✓ 3) 输入的token：主要假设是，如果可以从token的上下文中推断出token的信息，那么**就可以删掉该token**，从而减少KV缓存大小。
- ✓ 4) 隐藏层维度：目前隐藏层已经是128维度了，再优化的空间不大，但还是由KIVI和WKVQuant等量化方法可以在一定程度上压缩KV缓存。

图表：Transformer架构KV缓存计算公式

$$100K \text{ context: } \frac{100000}{\text{seqLen}} \times \frac{60}{\text{layer}} \times \frac{8}{\text{head}} \times \frac{128}{\text{dim}} \times \frac{2}{\text{KV}} \times \frac{2}{\text{bf16}} \text{ bytes} = 22.8GB \quad (1)$$

$$4K \text{ context: } \frac{4000}{\text{seqLen}} \times \frac{60}{\text{layer}} \times \frac{8}{\text{head}} \times \frac{128}{\text{dim}} \times \frac{2}{\text{KV}} \times \frac{2}{\text{bf16}} \text{ bytes} = 0.91GB \quad (2)$$

图表：压缩KV缓存的主要技术路线

	Desc	Improves	Needle?
Layer			
CALM [22]	Early exit based on estimated confidence	C P D S	?
CoLT5 [2]	Conditionally reducing computation on some layer	C P D S	?
LayerSkip [8]	Skipping some layers then verify	C P D S	?
YOCO [24]	Use only one global KV cache	C P D S	✓
Head			
Voita et al. [26]	Head pruning based on gating	C D S	?
GQA [1]	Reusing KV cache for groups of heads	C D S	✓
Retrieval Head [27]	Removing non-retrieval heads	C D S	✓
MLA [7]	Using latent head	C P D S	✓
Token			
H2O [31]	Dropping insignificant tokens after prefilling	C D S	?
FastGen [10]	Identify important tokens during prefilling	C D S	?
DMC [21]	Dynamically merge tokens	C P D S	?
SnapKV [18]	Identify important tokens based on user questions	D	✓
TriForce [23]	Speculative decoding for long-context	D	✓
Hidden			
KIVI [20]	KV cache quantization	C D S	?
WKVQuant [30]	Weight and KV cache quantization	C D S	?

注：Needle的✓符号表示无损压缩；C表示并发数量、P表示预填充、D表示编码、S表示上下文切换。

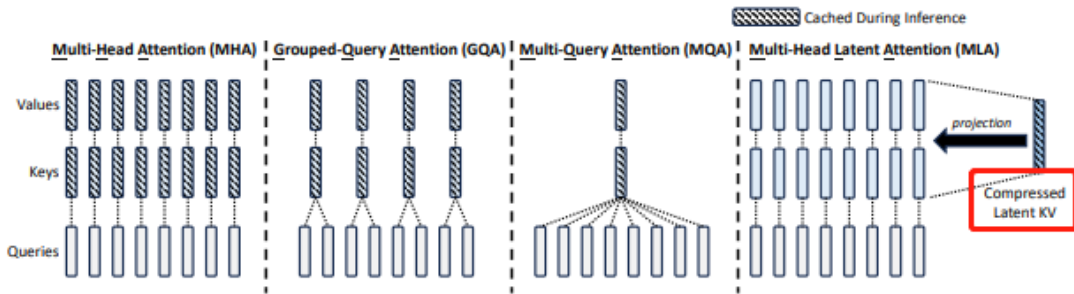
2.3.2 降价的背后是训练&推理成本的下降

➤ 国产大模型DeepSeek V2压缩KV缓存的实践:

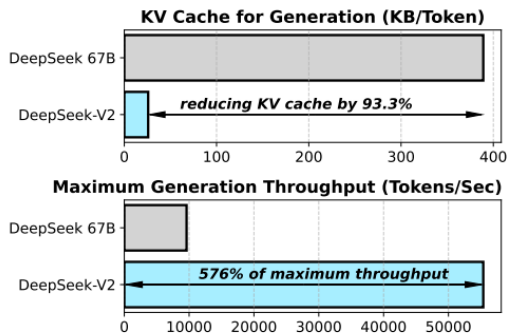
✓ **技术路线:** 采用多头潜在注意力机制**MLA**将KV缓存显著压缩成一个潜在向量来保证有效的推理, 从而提高推理效率。传统MHA方法, 一个Q向量与一对KV向量对应, 而GQA和MQA在压缩KV缓存时, 多个Q向量会对应一组KV向量, 实现缓存的压缩, 但一定程度上会影响模型性能。**MLA使用了低秩KV缓存联合压缩的方法, 通过引入低秩向量将KV压缩到低维空间, 相比MHA显著降低推理过程中缓存的大小, 并取得更好的性能。**

✓ 根据DeepSeek-V2的技术论文, DeepSeek-V2相对上代模型**KV缓存减少93.3%**, 最大吞吐量提升**576%**。简单估计, DeepSeek-V2以236B总参数、21B激活, 大致达到70B~110B Dense的模型能力, 同时消耗的显存(KV Cache)只有同级别Dense模型的1/5~1/100, 每token成本大幅降低。实际部署在8卡H800机器上, 输入吞吐量超过每秒10万tokens, 输出超过每秒5万tokens。

图表: DeepSeek V2通过MLA方法压缩KV缓存



图表: MLA方法减少了93.3%的KV缓存



图表: MLA对KV缓存压缩幅度最大

Attention Mechanism	KV Cache per Token (# Element)	Capability
Multi-Head Attention (MHA)	$2n_q d_h l$	Strong
Grouped-Query Attention (GQA)	$2n_g d_h l$	Moderate
Multi-Query Attention (MQA)	$2d_h l$	Weak
MLA (Ours)	$(d_c + d_h^R)l \approx \frac{9}{2}d_h l$	Stronger

图表: MLA相比MHA模型性能有所提升

Benchmark (Metric)	# Shots	Small MoE w/ MHA	Small MoE w/ MLA	Large MoE w/ MHA	Large MoE w/ MLA
# Activated Params	-	2.5B	2.4B	25.0B	21.5B
# Total Params	-	15.8B	15.7B	250.8B	247.4B
KV Cache per Token (# Element)	-	110.6K	15.6K	860.2K	34.6K
BBH (EM)	3-shot	37.9	39.0	46.6	50.7
MMLU (Acc.)	5-shot	48.7	50.0	57.5	59.0
C-Eval (Acc.)	5-shot	51.6	50.9	57.9	59.2
CMMLU (Acc.)	5-shot	52.3	53.4	60.7	62.5

03

投资建议

3. 投资建议

- ▶ 建议关注（1）国产大模型厂商：科大讯飞、商汤、云从科技、格林深瞳、拓尔思、昆仑万维、创业黑马等。（2）接入头部大模型的应用标的：金山办公、万兴科技、福昕软件、虹软科技、彩讯股份、焦点科技、润达医疗、金证股份、泛微网络、金蝶国际等，同时关注Kimi相关标的。

04

风险提示

4. 风险提示

- ▶ 海外大模型展现闭源趋势，国内大模型技术差距扩大：海外闭源发展以及对国内的技术封闭，可能导致国内大模型与海外龙头主流大模型产生代际差；
- ▶ 国内大模型在整体性能上未能达到商业使用的奇点：国内大模型目前与GPT-4等海外龙头大模型性能之间仍存在性能差异，大规模商业化应用仍需等待时机；
- ▶ 国内大模型在缺乏算力支持的情况下迭代速度放缓：美国对国内AI算力硬件采取严格封锁措施，国内在缺乏先进GPU的情况下，大模型迭代速度可能放缓；
- ▶ 国内大模型技术路线产生分歧，无法引领未来发展方向：国内大模型厂商数量众多，他们或都采取差异化的技术路线维持自身竞争力，但不利于集中力量攻克大模型发展难题。

分析师与研究助理简介

陈涵泊：德邦证券计算机行业首席分析师，上海交通大学信息安全本科，电子与通信工程硕士，曾任职于中信证券研究部、天风证券研究所，多年计算机行业研究经验，具备成熟的计算机研究框架、自上而下产业前瞻视野，云计算领域深入研究。2022-2023年新财富最佳分析师入围（团队），2023年新浪金麒麟最佳分析师第五名（团队）。

王思：德邦证券计算机行业研究助理，湖南大学金融学学士、武汉大学金融学硕士，主要覆盖云计算、工业软件、网安等方向。

投资评级说明

	类别	评级	说明
1. 投资评级的比较和评级标准： 以报告发布后的6个月内的市场表现为比较标准，报告发布日后6个月内的公司股价（或行业指数）的涨跌幅相对同期市场基准指数的涨跌幅；	股票投资评级	买入	相对强于市场表现20%以上；
		增持	相对强于市场表现5%~20%；
		中性	相对市场表现在-5%~+5%之间波动；
		减持	相对弱于市场表现5%以下。
2. 市场基准指数的比较标准： A股市场以上证综指或深证成指为基准；香港市场以恒生指数为基准；美国市场以标普500或纳斯达克综合指数为基准。	行业投资评级	优于大市	预期行业整体回报高于基准指数整体水平10%以上；
		中性	预期行业整体回报介于基准指数整体水平-10%与10%之间；
		弱于大市	预期行业整体回报低于基准指数整体水平10%以下。

免责声明

分析师声明：本人具有中国证券业协会授予的证券投资咨询执业资格，以勤勉的职业态度、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告，本报告所采用的数据和信息均来自市场公开信息，本人对这些信息的准确性或完整性不做任何保证，也不保证所包含的信息和建议不会发生任何变更。报告中的信息和意见仅供参考。本人过去不曾与、现在不与、未来也将不会因本报告中的具体推荐意见或观点而直接或间接接收任何形式的补偿，分析结论不受任何第三方的授意或影响，特此声明。

法律声明：

本报告仅供德邦证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。市场有风险，投资需谨慎。本报告所载的信息、材料及结论只提供特定客户作参考，不构成投资建议，也没有考虑到个别客户特殊的投资目标、财务状况或需要。客户应考虑本报告中的任何意见或建议是否符合其特定状况。在法律许可的情况下，德邦证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。本报告仅向特定客户传送，未经德邦证券研究所书面授权，本研究报告的任何部分均不得以任何方式制作任何形式的拷贝、复印件或复制品，或再次分发给任何其他人，或以任何侵犯本公司版权的其他方式使用。所有本报告中使用的商标、服务标记及标记均为本公司的商标、服务标记及标记。如欲引用或转载本文内容，务必联络德邦证券研究所并获得许可，并需注明出处为德邦证券研究所，且不得对本文进行有悖原意的引用和删改。根据中国证监会核发的经营证券业务许可，德邦证券股份有限公司的经营范围包括证券投资咨询业务。



德邦证券
Topsperty Securities

德邦证券股份有限公司

地 址：上海市中山东二路600号外滩金融中心N1幢9层

电 话：+86 21 68761616 传 真：+86 21 68767880

400-8888-128