ZOMI

大模型迎来跳水降价潮


huggingface.co/deepseek-ai

不挣钱了　就是交个朋友

为行业发展做贡献！

关于本内容

1. 大模型降价潮：最新 LLM 大模型降价对比介绍
2. 降价技术分析：MOE – KVCache – 低精度 – MLA 推理提速
3. 看大模型降价趋势：百模厂商的冲击 && 产业思考

I. 大模型降价潮

百模大战：你降价、我免费

大模型 API 降价时间线

4月25日 — **DeepSeek**

API付费服务定价为1元/百万输入tokens，2元/百万输出tokens。

5月6日 — **DeepSeek**

发布全球最强开源MoE模型，参数更低，能力更强，成本是4月25日发布的API定价1元/百万输入tokens，2元/百万输出tokens。

5月11日 — **智谱AI**

GLM-3 Turbo 模型调用价格由5元/百万tokens调整为1元/百万tokens。相当于一万条小红书文案（以350字计）将仅需约 1 元钱。

5月13日 — **OpenAI**

GPT-4o发布，宣布免费使用。同时，API调用价格减半。

5月15日 — **火山引擎**

2024春季火山引擎FORCE原动力大会上，谭待宣布豆包通用模型pro-32k定价只有0.0008元/千tokens。

5月21日 上午 — **阿里云**

宣布通义千问商业化模型降价，其中核心模型Qwen-Long降幅97%，降价后仅为0.0005元/千token，降价后1元最多=200万tokens。

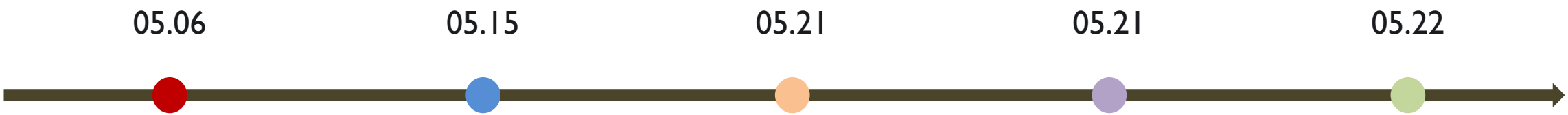
5月21日 下午 — **百度智能云**

宣布文心大模型两大主力模型ERNIE Speed、ERNIE Lite全面免费。

5月22日 — **腾讯云**

宣布全面下调大模型价格，混元-lite、元器-API分发（腾讯全域）免费，混元-standard 0.0045元/千输入tokens，0.005元/千输出tokens。

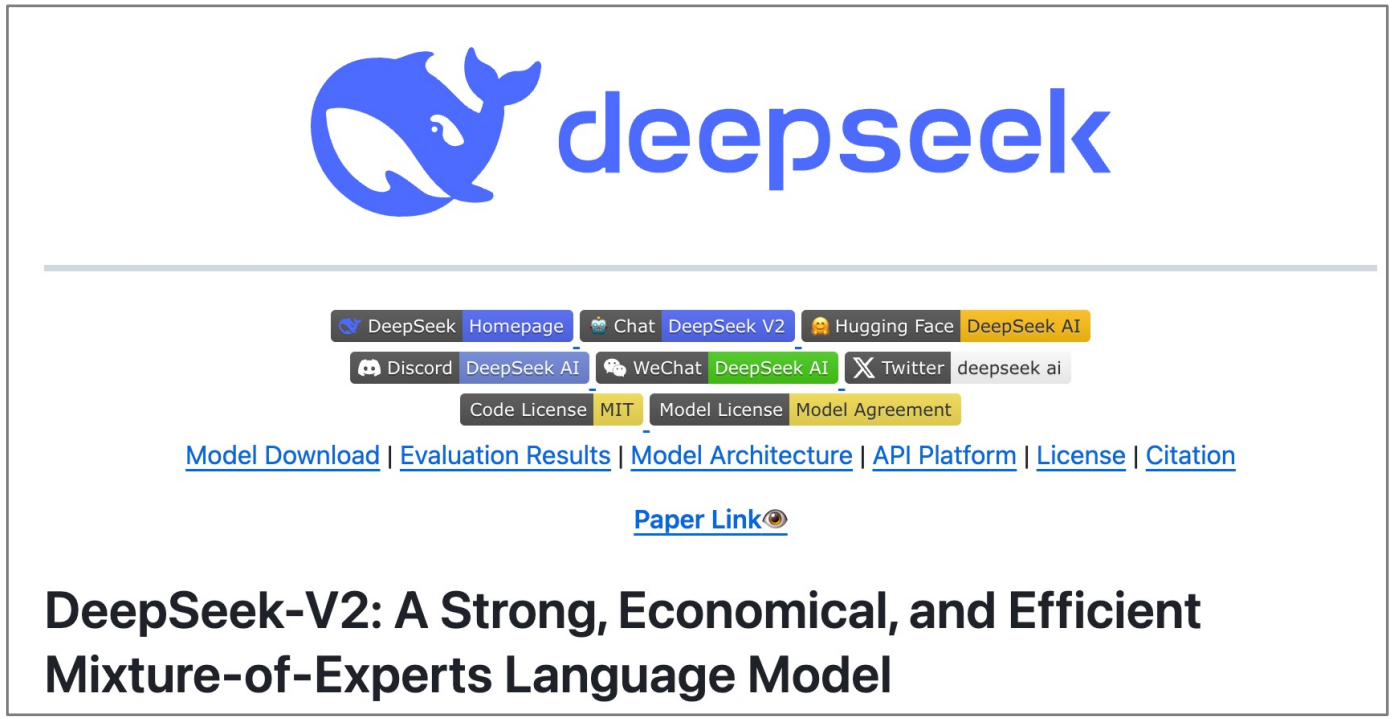
5 月份大模型价格战 TimeLine



	幻方量化	字节	阿里	百度	腾讯
十亿		• 豆包 Lite：输入 ￥0.0008，输出 ￥0.001	• Qwen-long：输入 ￥0.005，输出 ￥0.002	• Ernie Lite/Ernie Speed：免费	• 混元 Lite：免费
百亿	• DeepSpeek 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

第一波降价：以分￥0.0X 计价

- 5月6日，幻方量化宣布旗下深度求索（DeepSeek）正式开源第二代 MoE DeepSeek-V2，API 输入 0.01元/万 Tokens、输出 0.02 元/万 Tokens，~GPT-4-Turbo 1/70；



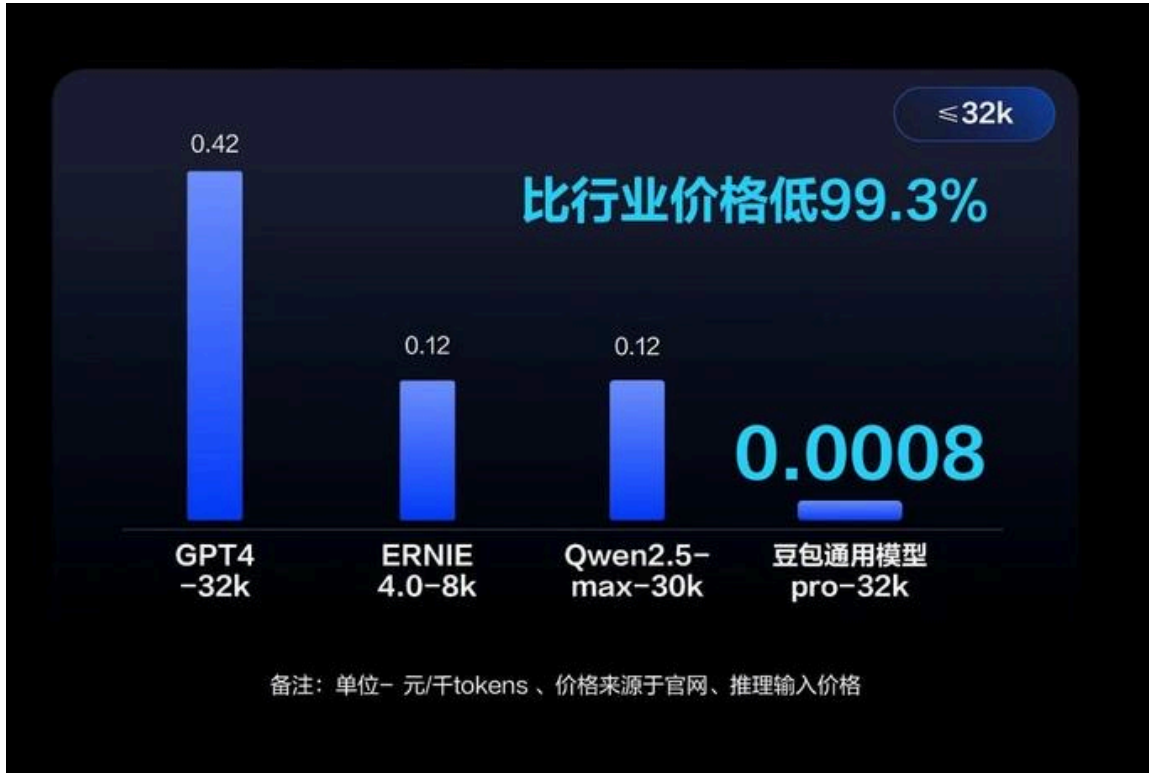
Model	API Pirce / 1M Tokens	
	Input $	Output $
DeepSeek-V2	0.14	0.28
GPT-4-Turbo-1106	10.00	30.00
GPT-4-0613	30.00	60.00
GPT-3.5	1.50	2.00
Gemini 1.5 Pro	7.00	21.00
Claude 3 Opus	15.00	75.00
Claude 3 Sonnet	3.00	15.00
Claude 3 Haiku	0.25	1.25
abab-6.5（MiniMax）	4.14	4.14
abab-6.5s（MiniMax）	1.38	1.38
ERNIE-4.0（文心一言）	16.56	16.56
GLM-4（智谱清言）	13.80	13.80
Moonshot-v1（月之暗面）	3.32	3.32
Qwen1.5 72B（通义千问）	2.76	2.76
LLaMA 3 70B	3.78	11.34
Mixtral 8x22B	2.00	6.00

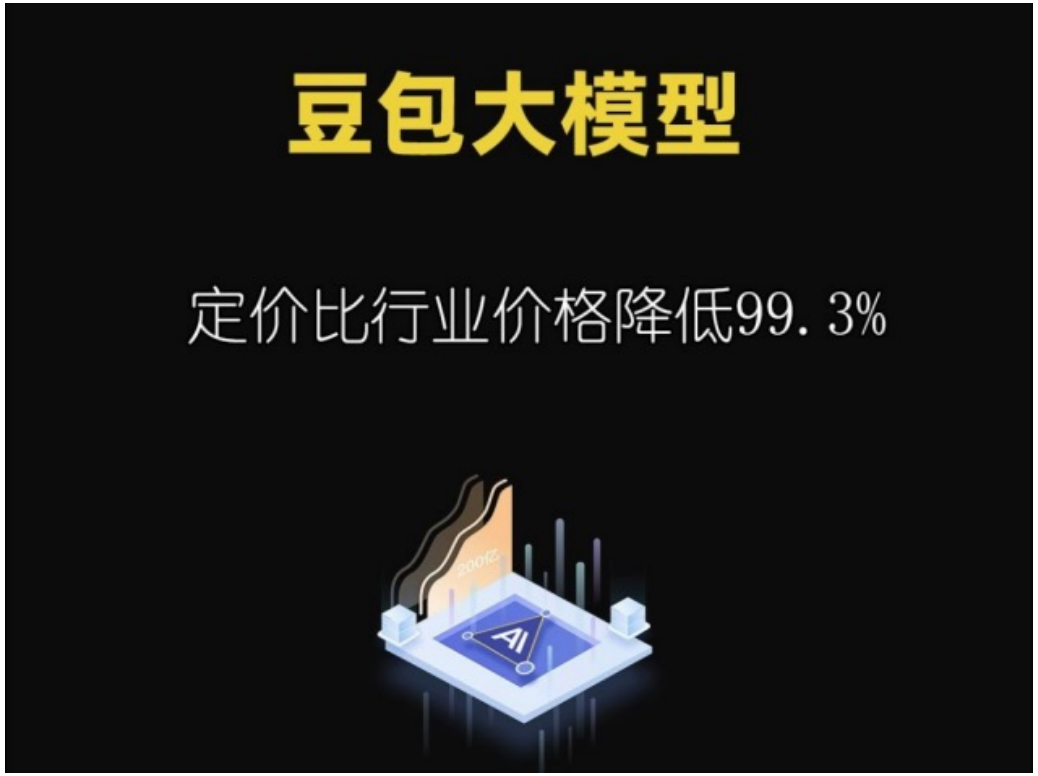
第一波降价：以分￥0.0X 计价

- 5月11日，智谱大模型跟进，官宣新价格：入门级产品GLM-3 Turbo API 从0.05元 / 万Tokens降至0.01元 / 万Tokens，降幅达 80%；

第二波降价：以厘 ￥0.00X 计价

- 5月15日，字节宣布豆包大模型正式开启对外服务，豆包通用大模型 pro-32k 企业市场推理输入价格仅为 0.008 元/万 tokens。





第二波降价：以厘 ￥0.00X 计价

- 5月21日，阿里云宣布通义千问主力模型 Qwen-Long API输入价格从 0.2元/万 tokens 降至 0.005 元/万 tokens，降幅达 97%。降价后，通义千问的价格为GPT-4 1/400。

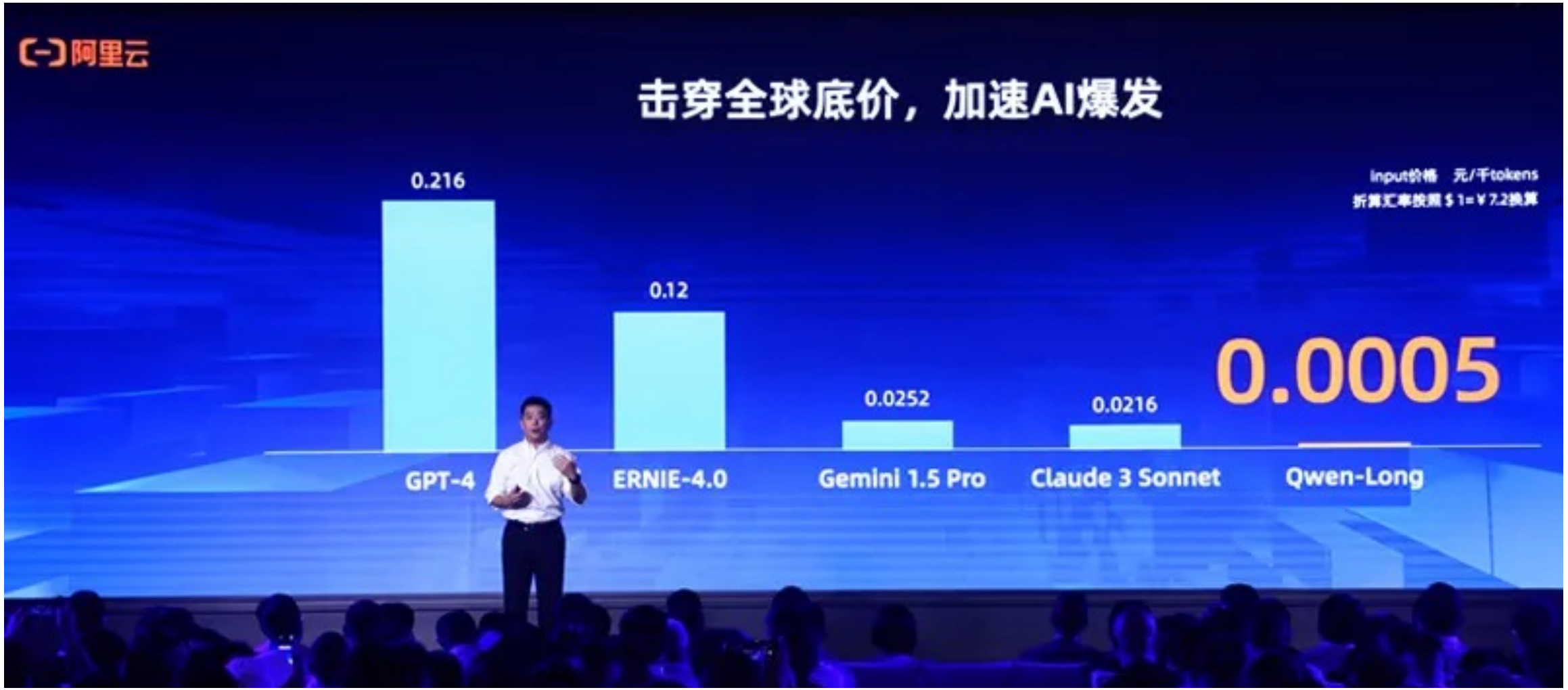
通义千问2024.05降价公告

2024年5月21日生效

模型规格		input 价格（元/千 tokens）		降幅	output 价格（元/千 tokens）		降幅
		降价前	降价后		降价前	降价后	
通义千问商业化模型	Qwen-Turbo	0.008	0.002	75%↓	0.008	0.006	25%↓
	Qwen-Plus	0.02	0.004	80%↓	0.02	0.012	40%↓
	Qwen-Long	**0.02**	**0.0005**	**97%↓**	**0.02**	**0.002**	**90%↓**
	Qwen-Max	0.12	0.04	67%↓	0.12	0.12	

模型规格		input 价格（元/千 tokens）		降幅	output 价格（元/千 tokens）		降幅
		降价前	降价后		降价前	降价后	
通义千问开源模型	Qwen1.5-7B	0.006	0.001	83%↓	0.006	0.002	67%↓
	Qwen1.5-14B	0.008	0.002	75%↓	0.008	0.004	50%↓
	Qwen1.5-32B		0.0035	七天限时免费		0.007	七天限时免费
	Qwen1.5-72B	0.02	0.005	75%↓	0.02	0.01	50%↓
	Qwen1.5-110B		0.007	七天限时免费		0.014	七天限时免费

第二波降价：以厘 ￥0.00X 计价



第三波降价：免费

- 5月21日，百度宣布其文心大模型 Ernie 两大主力模型 ERNIE Speed、 ERNIE Lite 全面免费。

文心大模型两大主力模型全面免费，立即生效！

模型名称	上下文长度	输入	输出
ERNIE Speed	8K、128K	免费	免费
ERNIE Lite	8K、128K	免费	免费

模型规格	调整前价格 元/千tokens		调整后价格 元/千tokens	
	输入	输出	输入	输出
混元-lite	0.008	0.008	★免费	★免费
混元-standard	0.01	0.01	0.0045 ↓下降55%	0.005 ↓下降50%
混元-standard-256k	0.12	0.12	0.015 ↓下降87.5%	0.06 ↓下降50%
混元-pro	0.1	0.1	0.03 ↓下降70%	0.1
元器-API分发（其他场景）	100万免费tokens		★1亿免费tokens	
元器-API分发（腾讯全域）	全量免费支持			

第三波降价：免费

- 5月22日，科大讯飞宣布其大模型 讯飞星火Lite API永久免费开放；
- 腾讯云主力模型 混元-lite 提升 API 输入输出总长度，并且也改为全面免费。

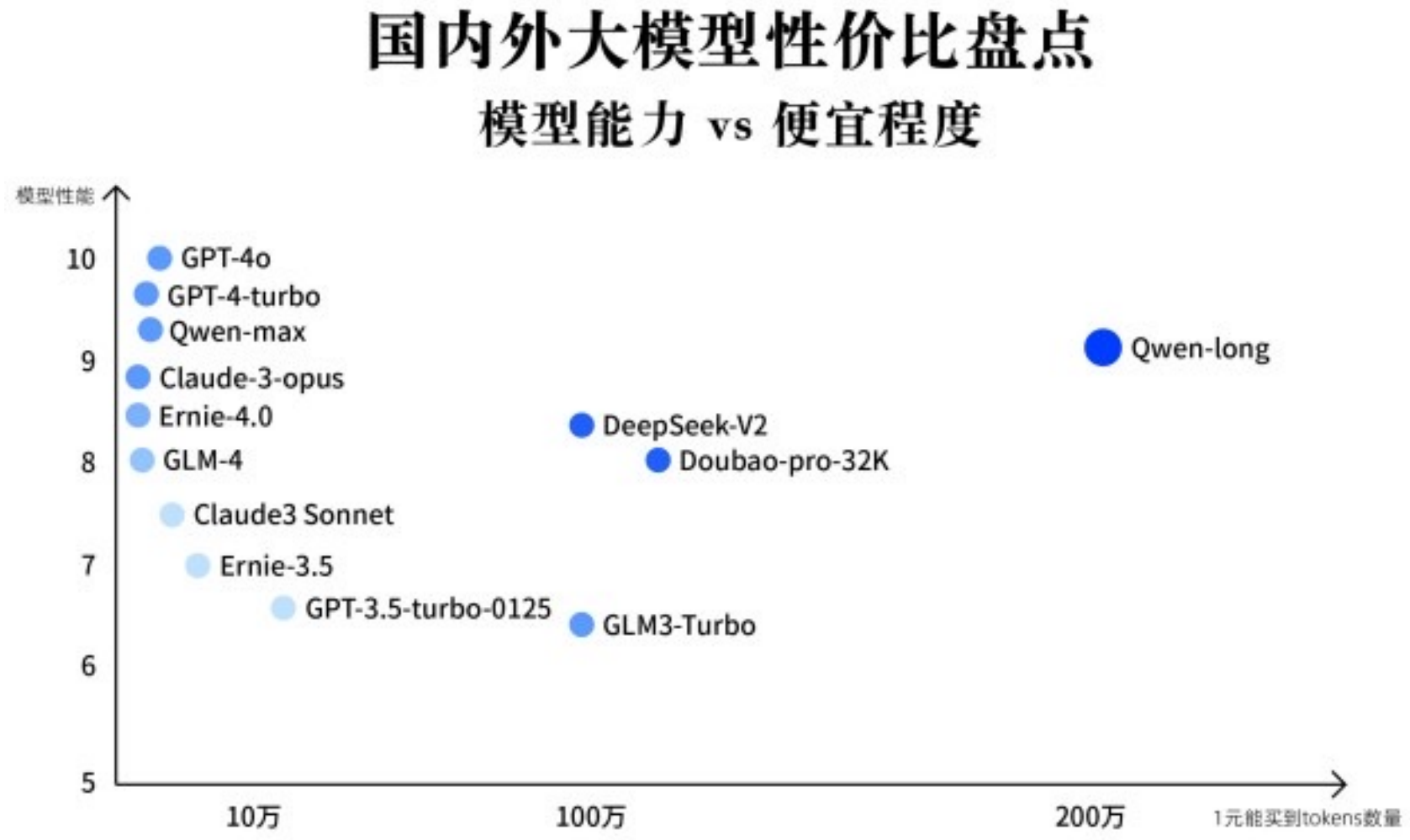
在免费额度用完后，按如下价格进行计费，每月1 – 3日系统会推送上个月账单并自动完成结算和扣费。

产品名	单位	刊例价
hunyuan–pro	每 1000 token	输入：0.03元 输出：0.10元
hunyuan–standard	每 1000 token	输入：0.0045元 输出：0.005元
hunyuan–standard–256k	每 1000 token	输入：0.015元 输出：0.06元
hunyuan–lite	每 1000 token	输入：免费 输出：免费
hunyuan–embedding	每 1000 token	0.0007元
腾讯元器智能体	每 1000 token	0.10元

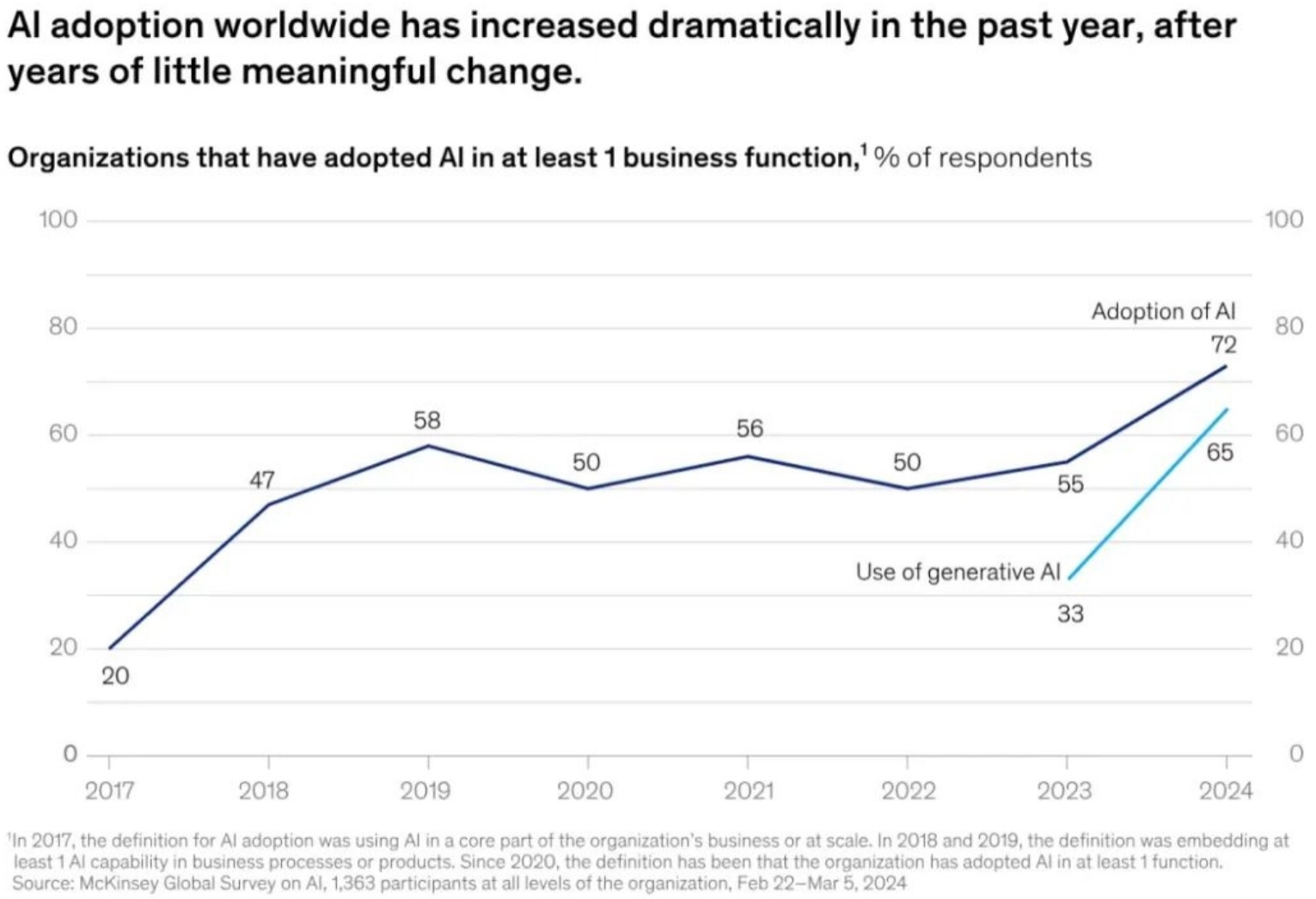
大模型价格一览

大模型名称	输入价格 元/千tokens	输出价格 元/千tokens	宣布价格或宣布下调价格的日期	所属公司
DeepSeek-V2	0.001	0.002	5 月 6 日	深度求索
GLM-3-Turbo	0.001	0.001	5 月 11 日	智谱 AI
豆包通用模型 pro-32k	0.0008	0.0008	5 月 15 日	字节跳动
通义千问 Qwen-Max	0.04	0.12	5 月 21 日	阿里云
通义千问 Qwen-Plus	0.004	0.002	5 月 21 日	阿里云
通义千问 Qwen-Long	0.0005	0.002	5 月 21 日	阿里云
文心一言 ERNIE Speed	免费	免费	5 月 21 日	百度
文心一言 ERNIE Lite	免费	免费	5 月 21 日	百度
讯飞星火 spark Lite	免费	免费	5 月 22 日	科大讯飞
讯飞星火 Spark3.5 Max	0.021-0.03	0.021-0.03	5 月 22 日	科大讯飞
混元-lite	免费	免费	5 月 22 日	腾讯云
混元-standard	0.0045	0.005	5 月 22 日	腾讯云
混元 -standard-256k	0.015	0.06	5 月 22 日	腾讯云
混元-pro	0.03	0.1	5 月 22 日	腾讯云

大模型价格一览


国内外大模型性价比盘点
模型能力 vs 便宜程度
模型性能
10
9
8
7
6
5
GPT-4o
GPT-4-turbo
Qwen-max
Claude-3-opus
Ernie-4.0
GLM-4
Claude3 Sonnet
Ernie-3.5
GPT-3.5-turbo-0125
GLM3-Turbo
DeepSeek-V2
Doubao-pro-32K
Qwen-long
10万
100万
200万
1元能买到tokens数量

大模型客户增长

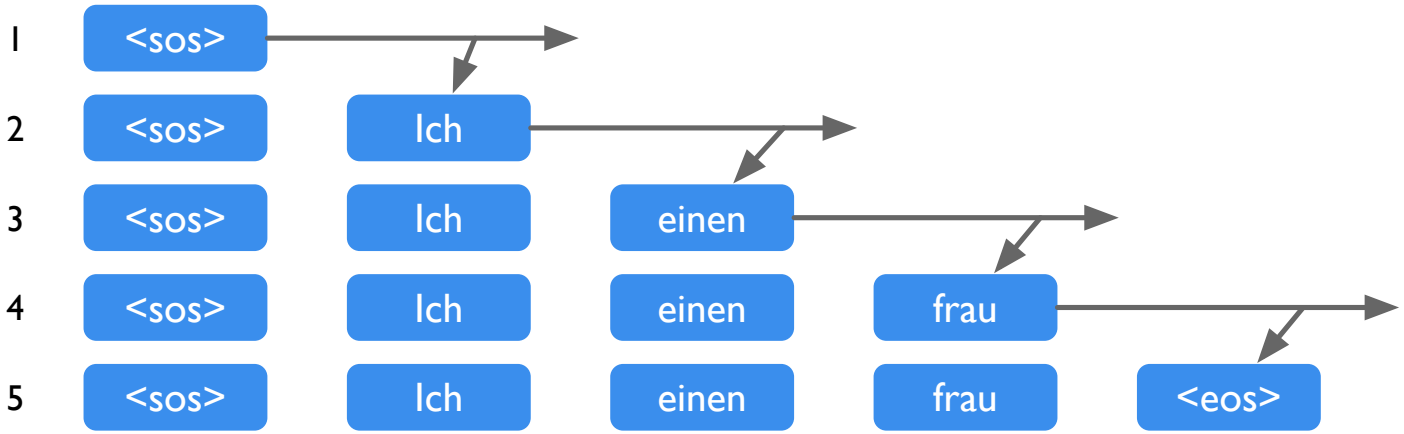

AI adoption worldwide has increased dramatically in the past year, after years of little meaningful change.
Organizations that have adopted AI in at least 1 business function,[1] % of respondents
100
80
60
40
20
0
Adoption of AI
72
58
56
50
50
47
55
65
Use of generative AI
33
20
2017
2018
2019
2020
2021
2022
2023
2024
[1]In 2017, the definition for AI adoption was using AI in a core part of the organization's business or at scale. In 2018 and 2019, the definition was embedding at least 1 AI capability in business processes or products. Since 2020, the definition has been that the organization has adopted AI in at least 1 function.
Source: McKinsey Global Survey on AI, 1,363 participants at all levels of the organization, Feb 22–Mar 5, 2024

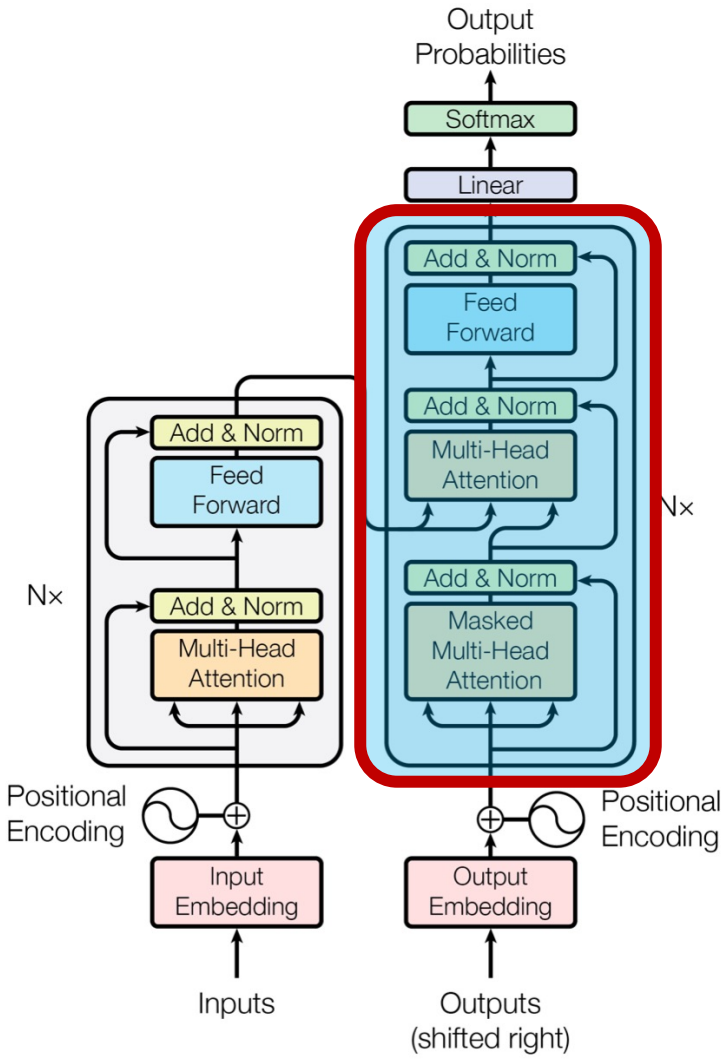
2. Tokens 词元

Transformer: Masked Multi Head Attention

- 输入和输出资源消耗不同，导致输入计费针对用户向大模型提交请求数据进行计费，输出针对模型返回给用户输出结果进行计费。

Inference



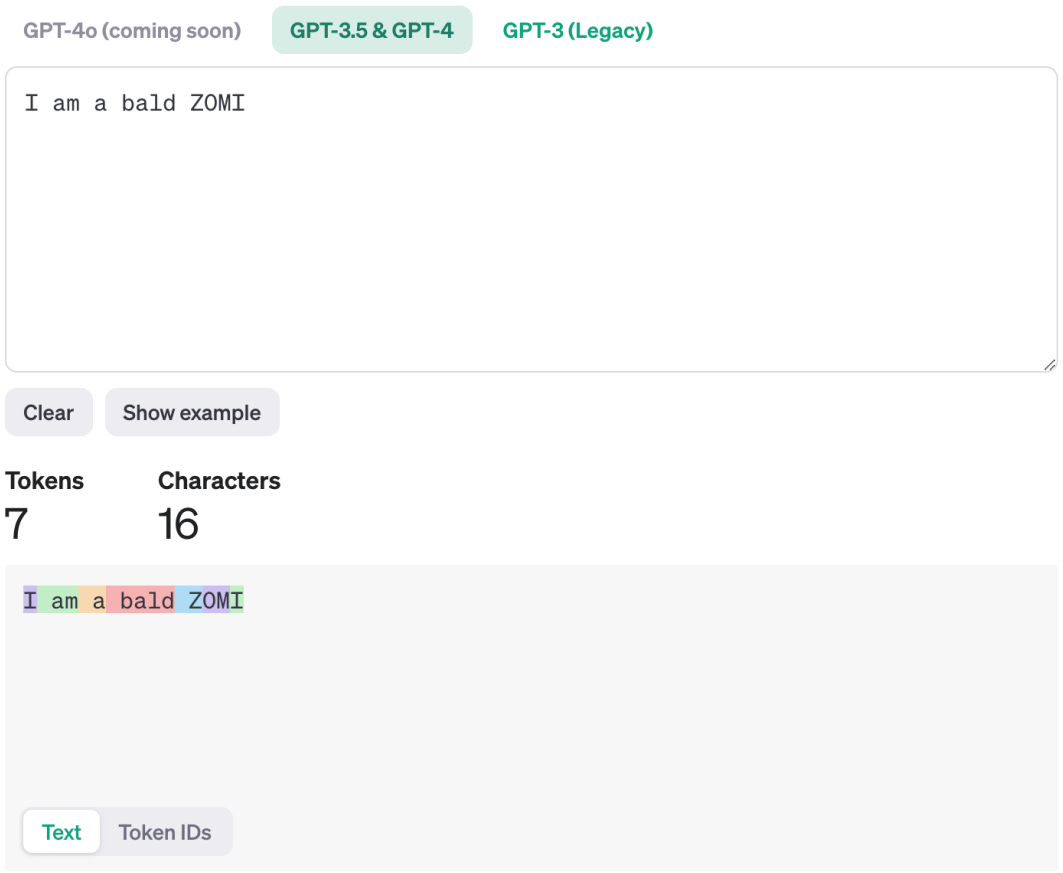


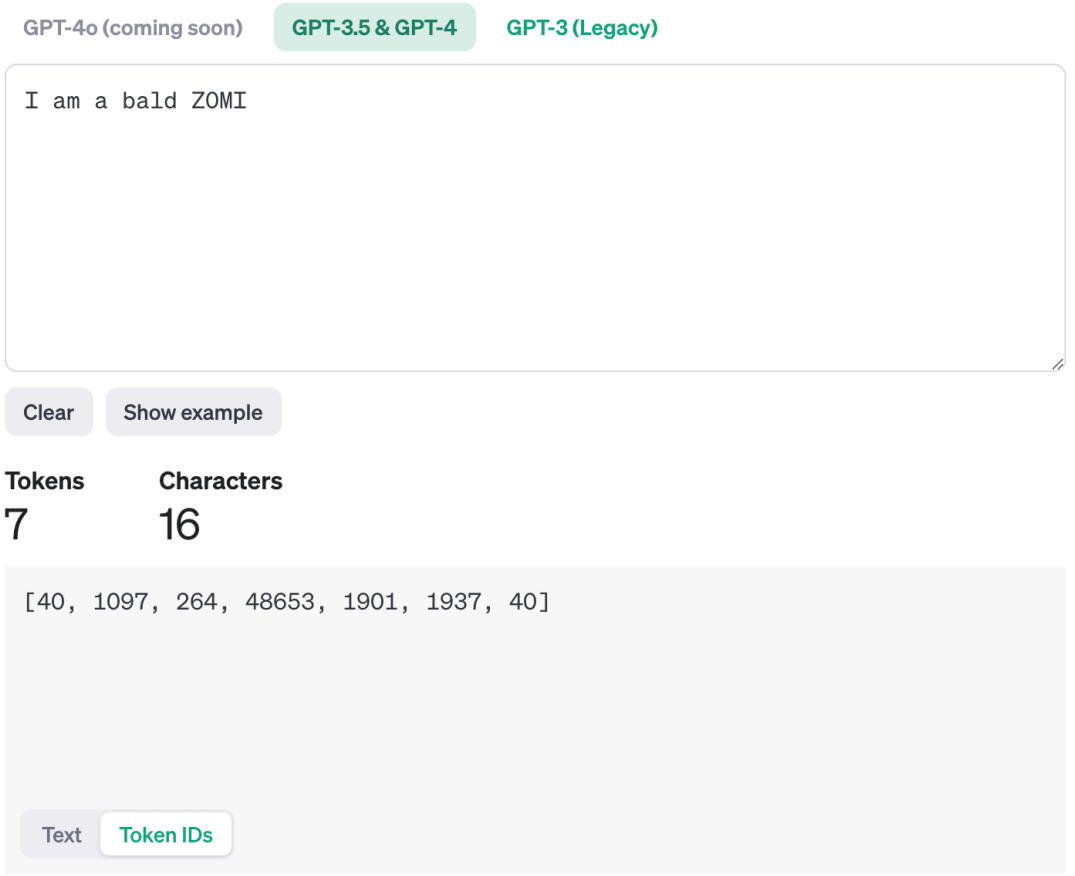
Tokens 词元

- 理解降价前，需要了解 LLM API 定价核心单位 —— Token 词元。
  1. Token 是 LLMs 用于处理和生成语言的文本或代码的基本单元。LLM API 发送 Prompt 或者 Query时，将输入转化为 Token 作为大模型输入，并且输出的张量也会转换为 Token 再进行输出。
  2. 对于英文文本，1 个 Token 大约为 4 个字符或 0.75 个单词。对于中文，1 个 Token 大约为 1.45 个中文单词。标点符号、空格和特殊字符通常被计为单独的 token。

what is token?

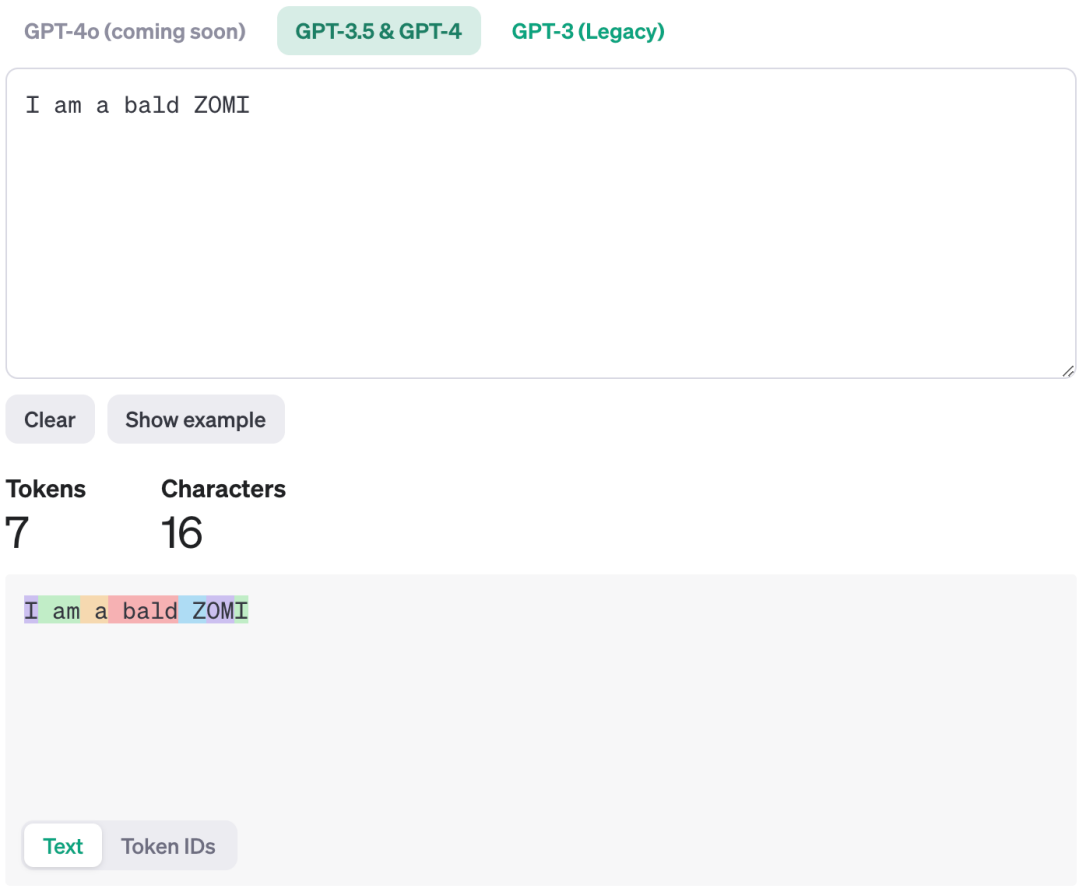
- https://platform.openai.com/tokenizer 1 个标记大约为 4 个字符或 0.75 个单词





what is token?

- https://platform.openai.com/tokenizer 1 个标记大约为 0.75 个英文单词， 为 1.45 个中文单词



Token 形象化理解

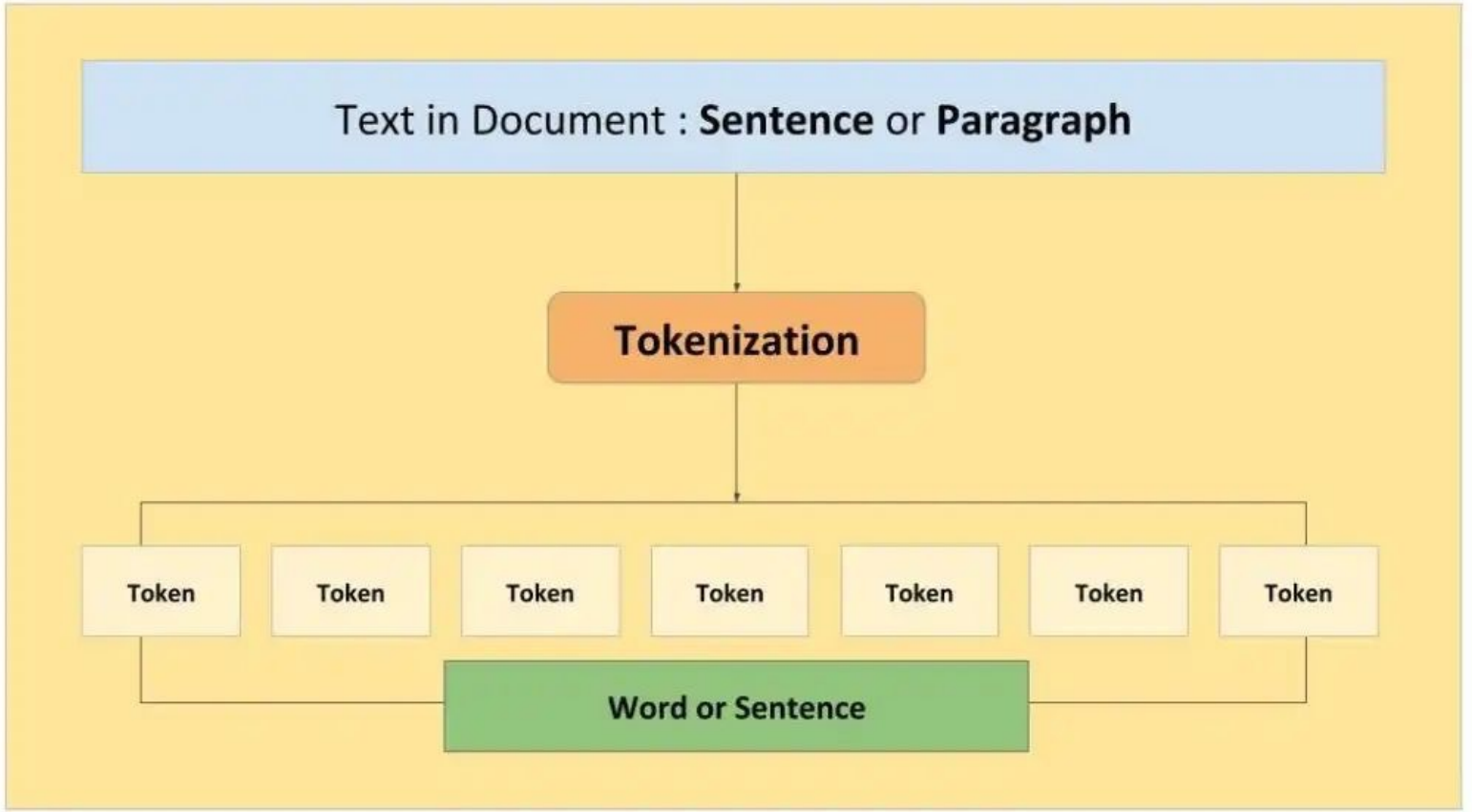
- 1 token ~= 4 chars in English
- 1 token ~= ¾ words
- 100 tokens ~= 75 words

或者

- 1-2 句子 ~= 30 tokens
- 1 段落 ~= 100 tokens
- 1,500 单词 ~= 2048 tokens

tokenization 的含义

- 将文本划分为不同 token 过程为 tokenization。tokenization 捕获文本含义和语法结构，将文本分割成由 Token 的组成。

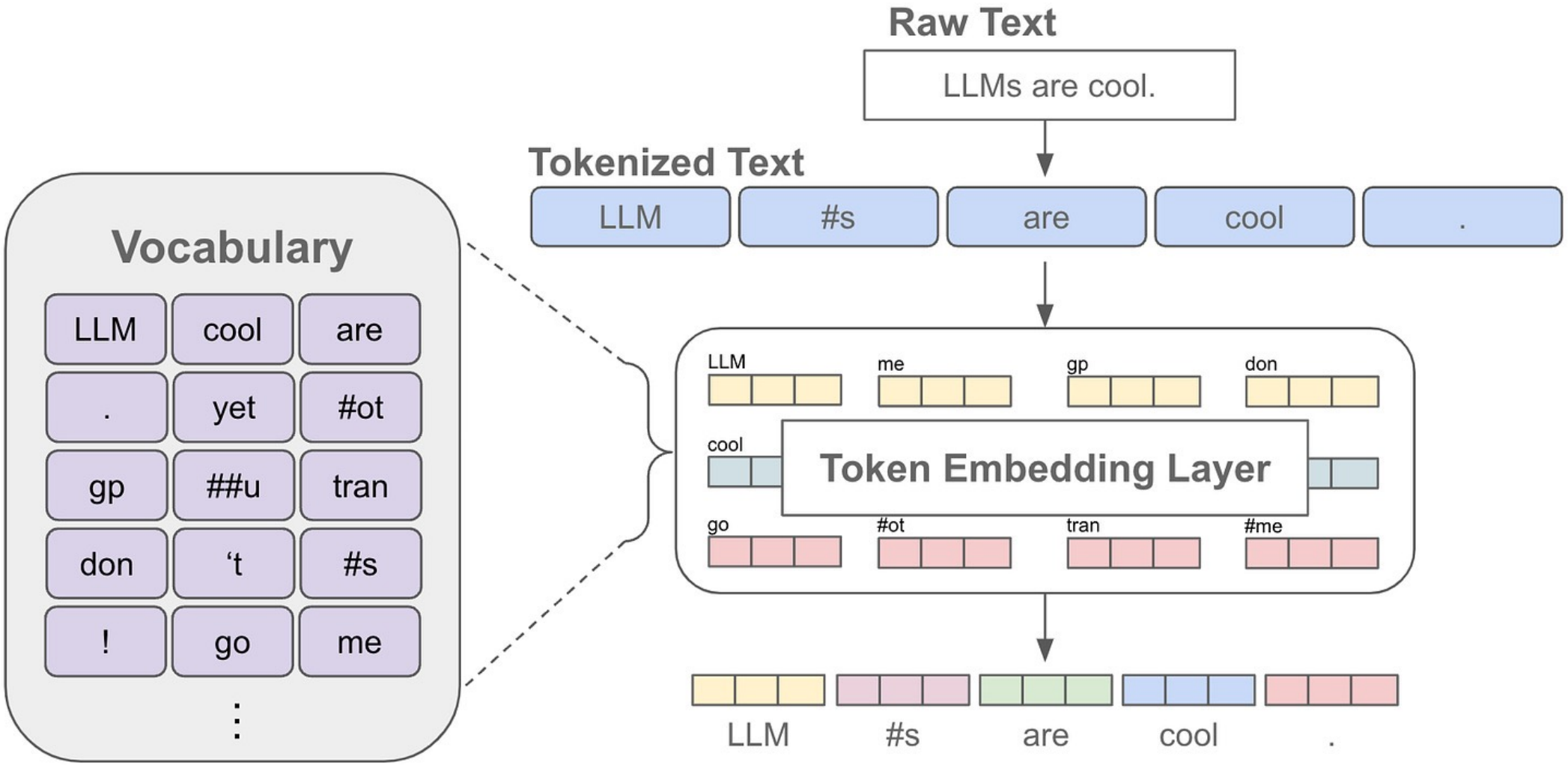


Tokenization 算法

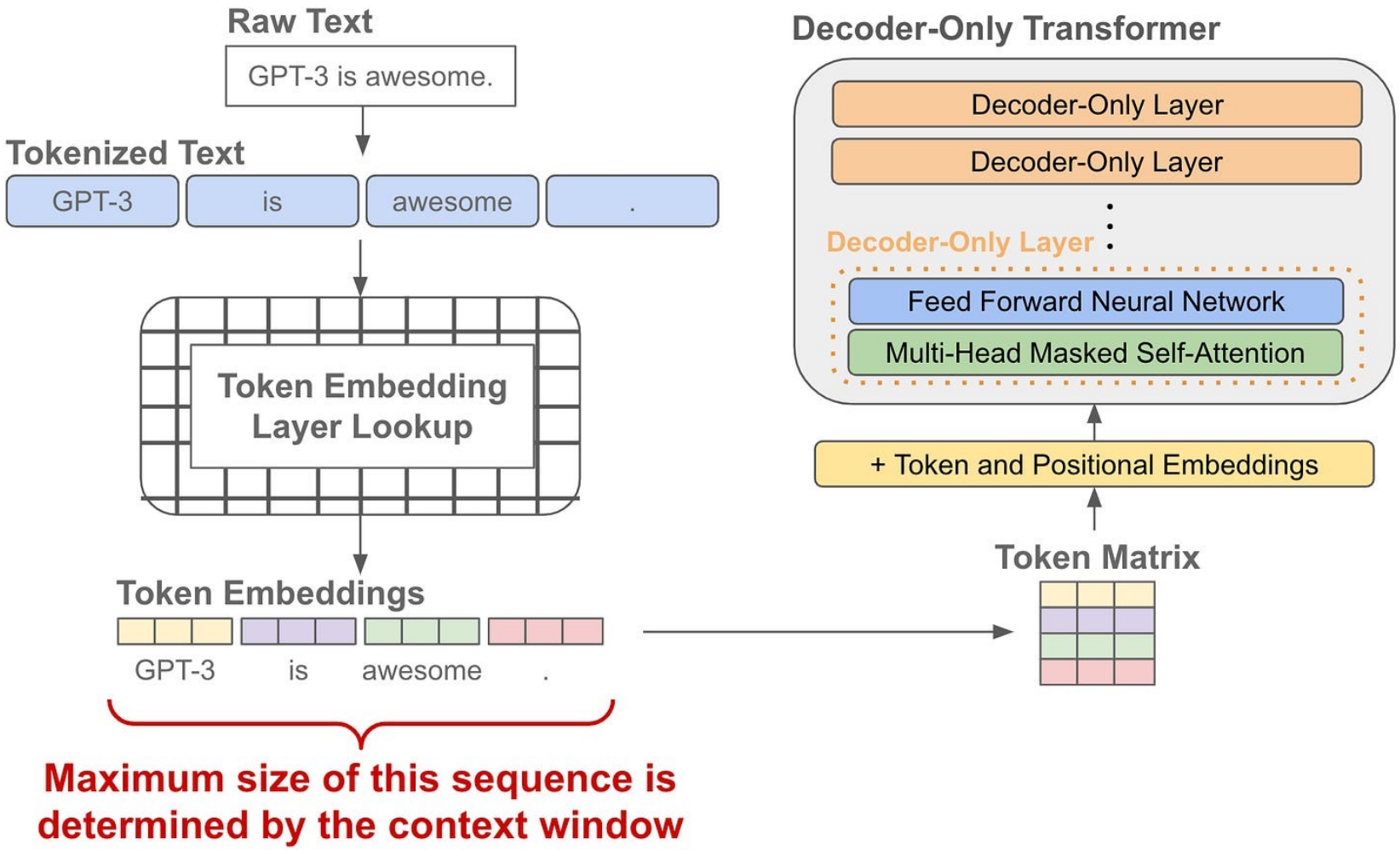
- LLM 已经扩展了处理多语言和多模式输入的能力。为了适应这些数据的多样性，已经开发了专门的tokenization方法。通过利用特定语言的token或子词技术，多语言标记在一个模型中处理多种语言。多模态标记将文本与其他模式(如图像或音频)结合起来，使用融合或连接等技术来有效地表示不同的数据源。

分词方法	典型模型
BPE	GPT, GPT-2, GPT-J, GPT-Neo, RoBERTa, BART, LLaMA, ChatGLM-6B, Baichuan
WordPiece	BERT, DistilBERT， MobileBERT
Unigram	AlBERT, T5, mBART, XLNet

Tokenized in LLM



Tokenized in LLM



Tokenized in LLM

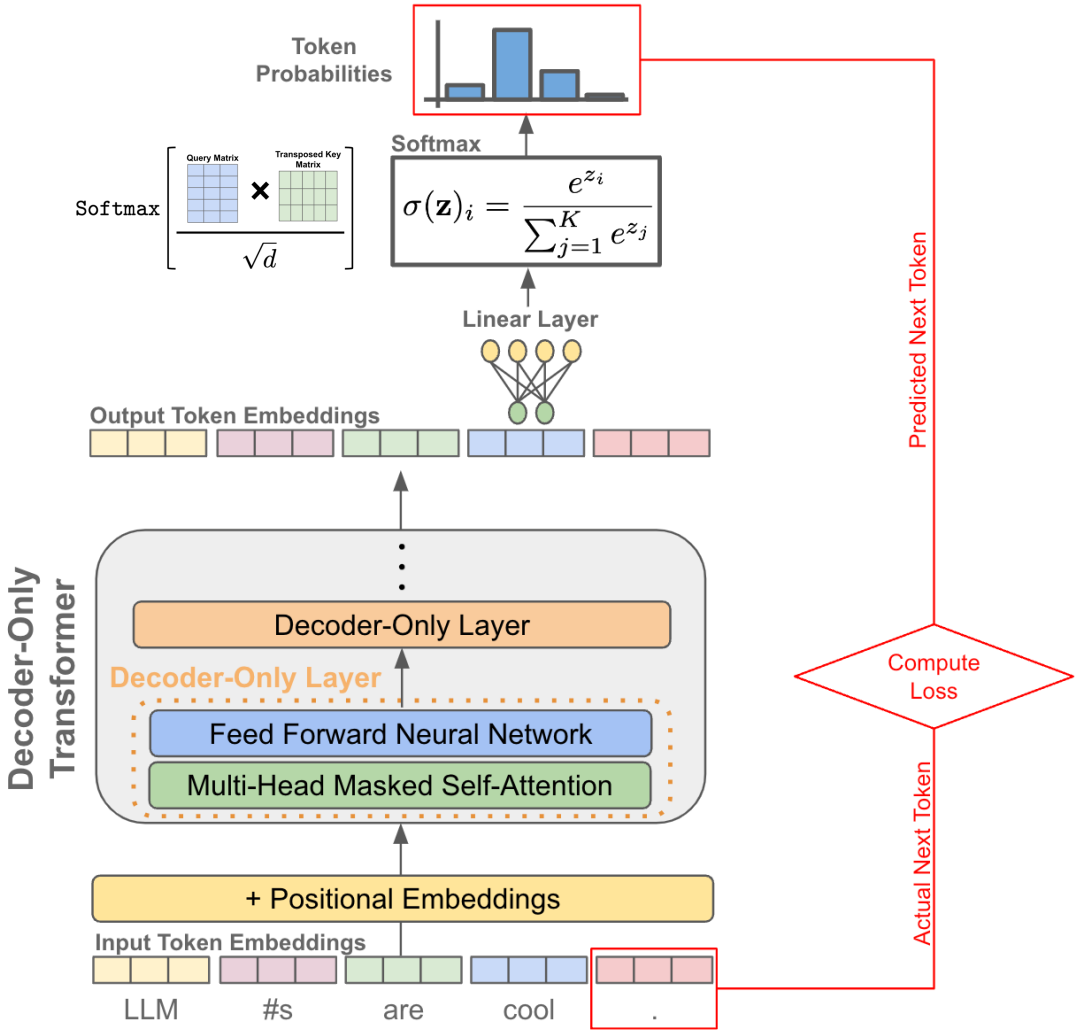
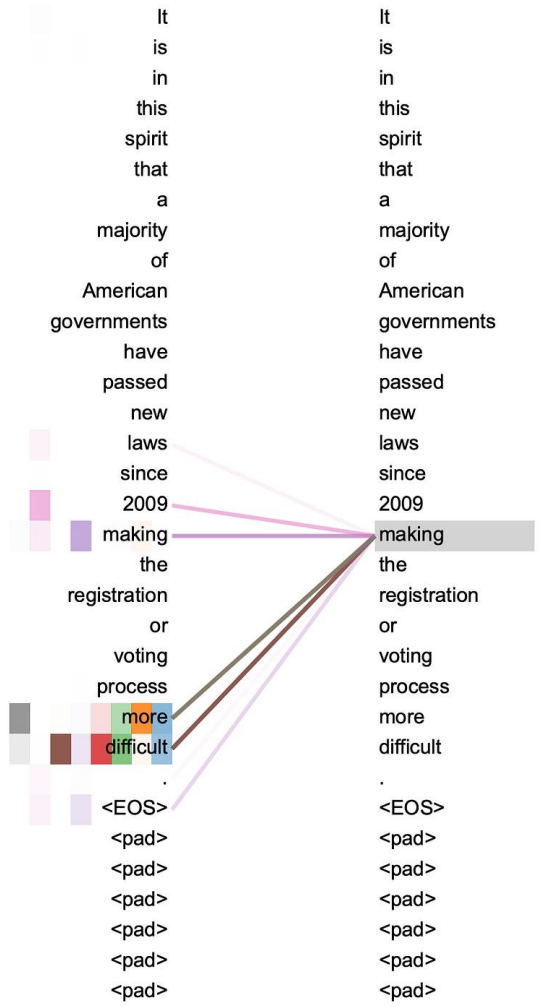
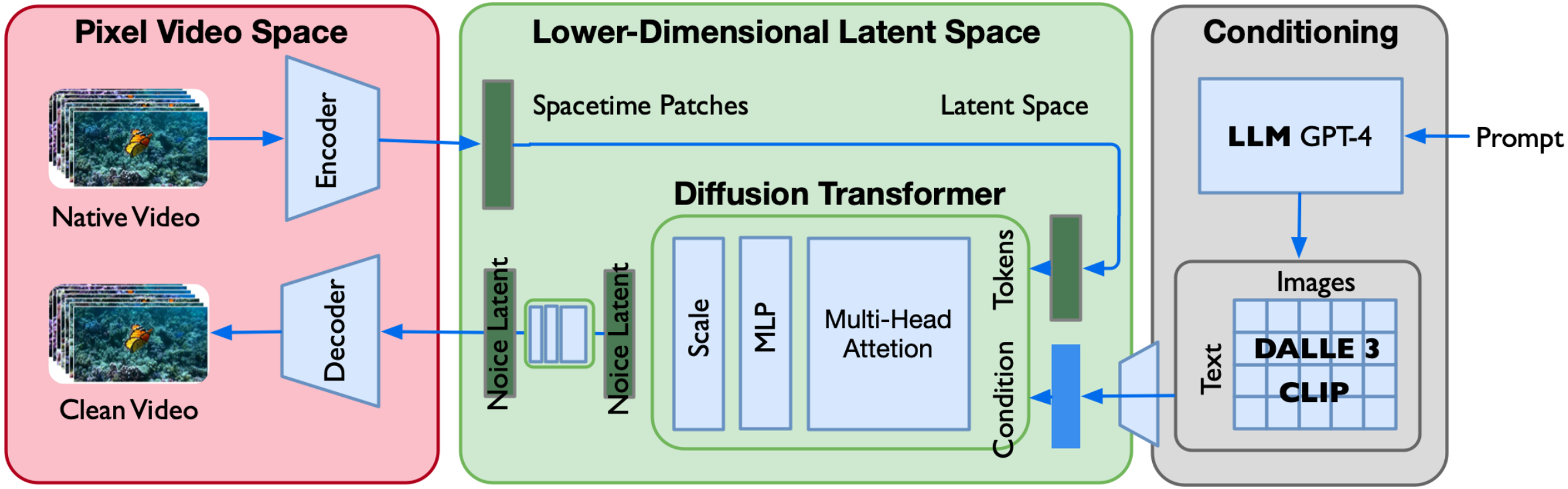


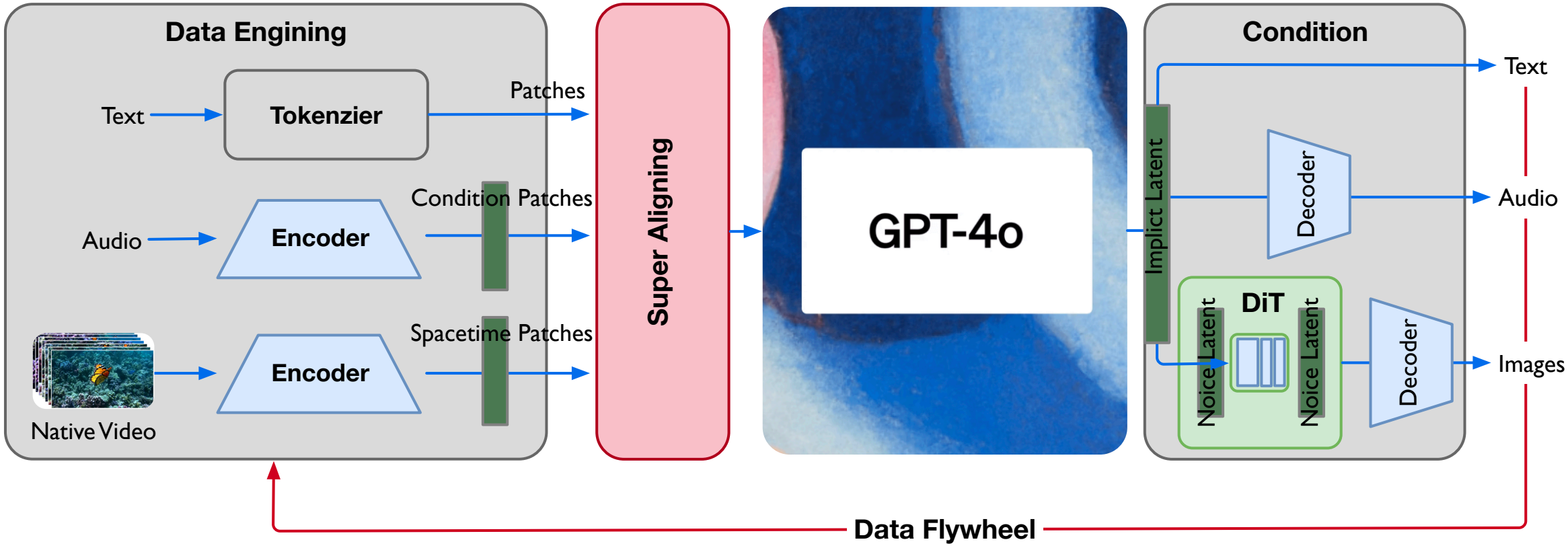
Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb 'making', completing the phrase 'making...more difficult'. Attentions here shown only for the word 'making'. Different colors represent different heads. Best viewed in color.



Tokenized in MLM



Tokenized in MLM



3. 降价技术分析

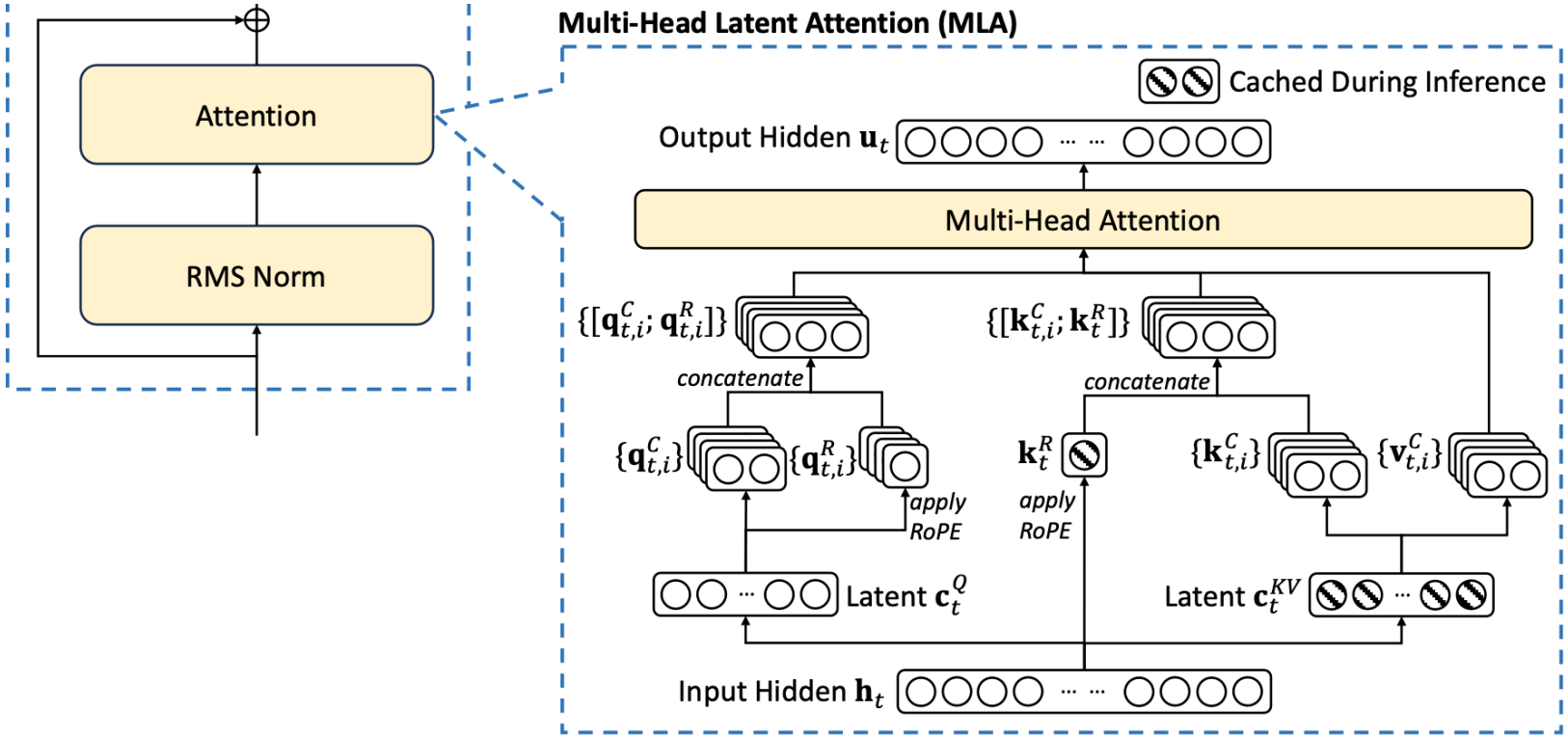
降价技术分析

- 算力未升级情况下，大模型价格战，得益于架构创新、推理优化、系统升级以及推理集群计算架构改进等多方面努力，让大模型推理成本实现了显著的下降。

- 目前，算法框架的创新主要沿着两条路径发展：轻量化和线性化。

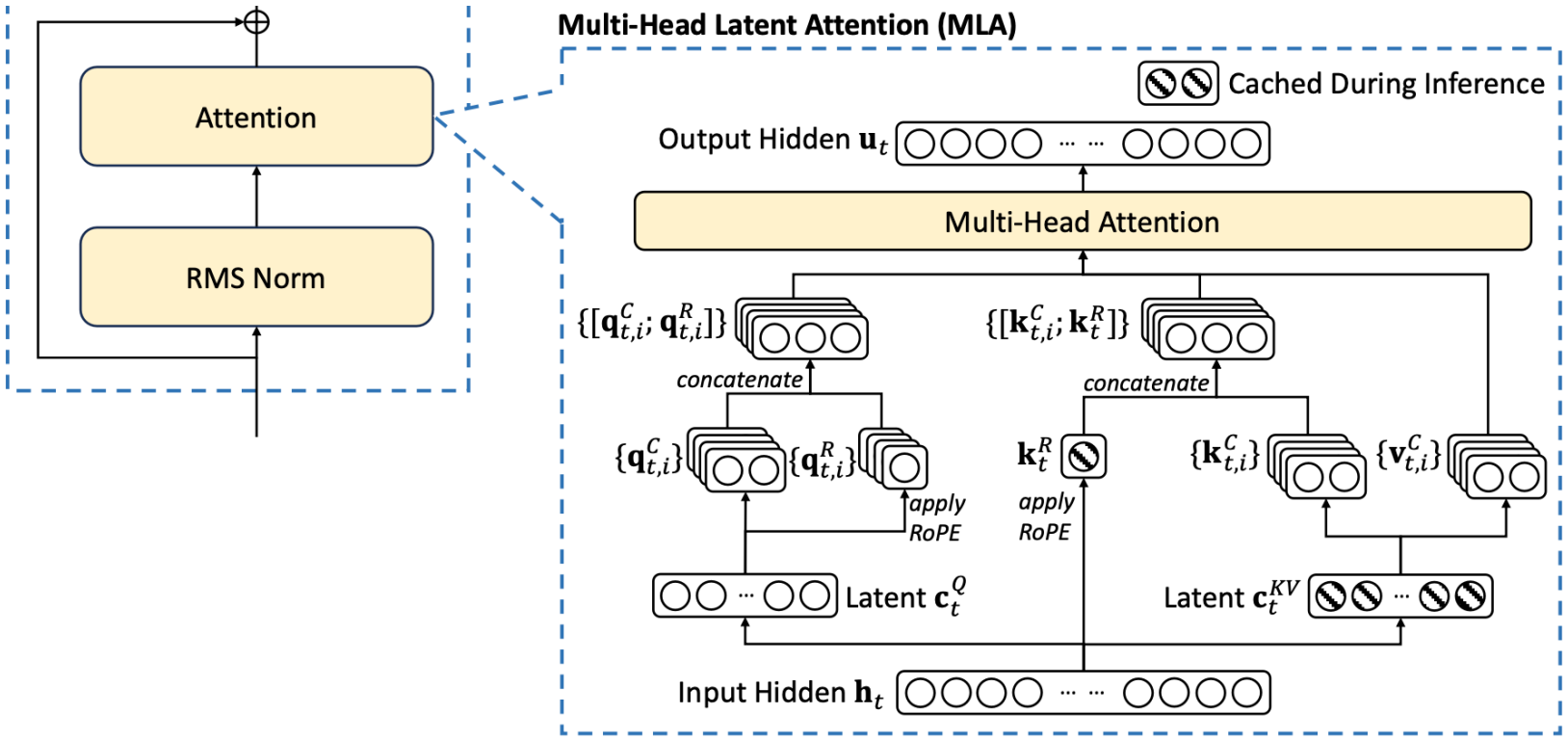
Multi-Head Latent Attention（MLA）

- MHA KV 被映射到为和 Q 相同维度里，推理时需要保存大量 KV。MLA 本质是：KV Cache 低秩分解。将 KV 映射到较小维度里，再映射回 Q 相同维度，保证了推理过程中只需要保存较小维度的KV值。



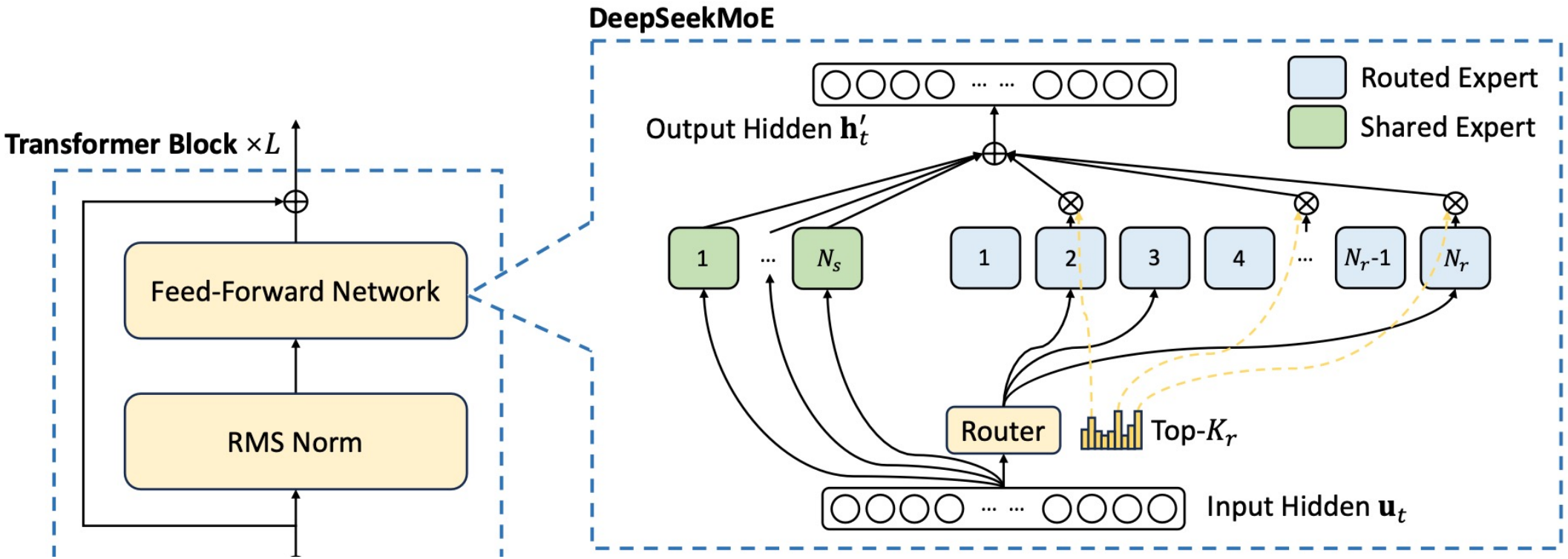
Multi-Head Latent Attention（MLA）

1. RoPE 会给每个不同位置 token QK 向量执行矩阵变换，这这会打乱 MLA 中矩阵融合操作，DeepSeek-V2 单独在一个较小维度 QK 执行 RoPE，相当于在 MLA 保留 MHA 计算。

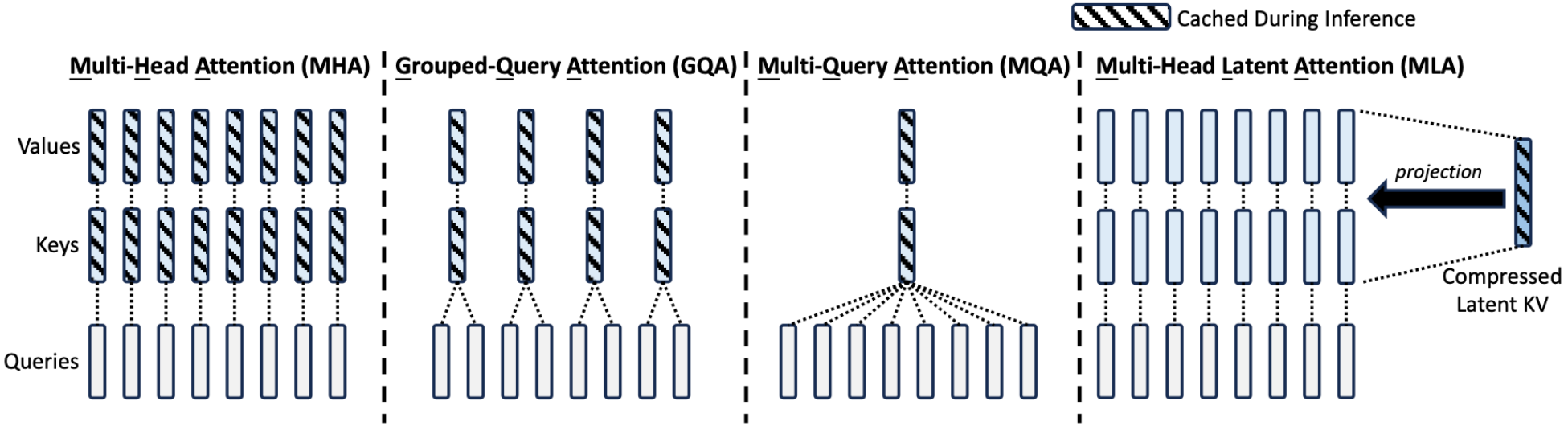


Multi-Head Latent Attention（MLA）

- 专家分割：DeepSeek 将专家分割更细粒度，同时包含共享专家 SE 和路由专家 RE。SE 用于减少路由专家之间知识冗余，RE 负责特定任务或数据。有助于提高模型整体效率和性能。

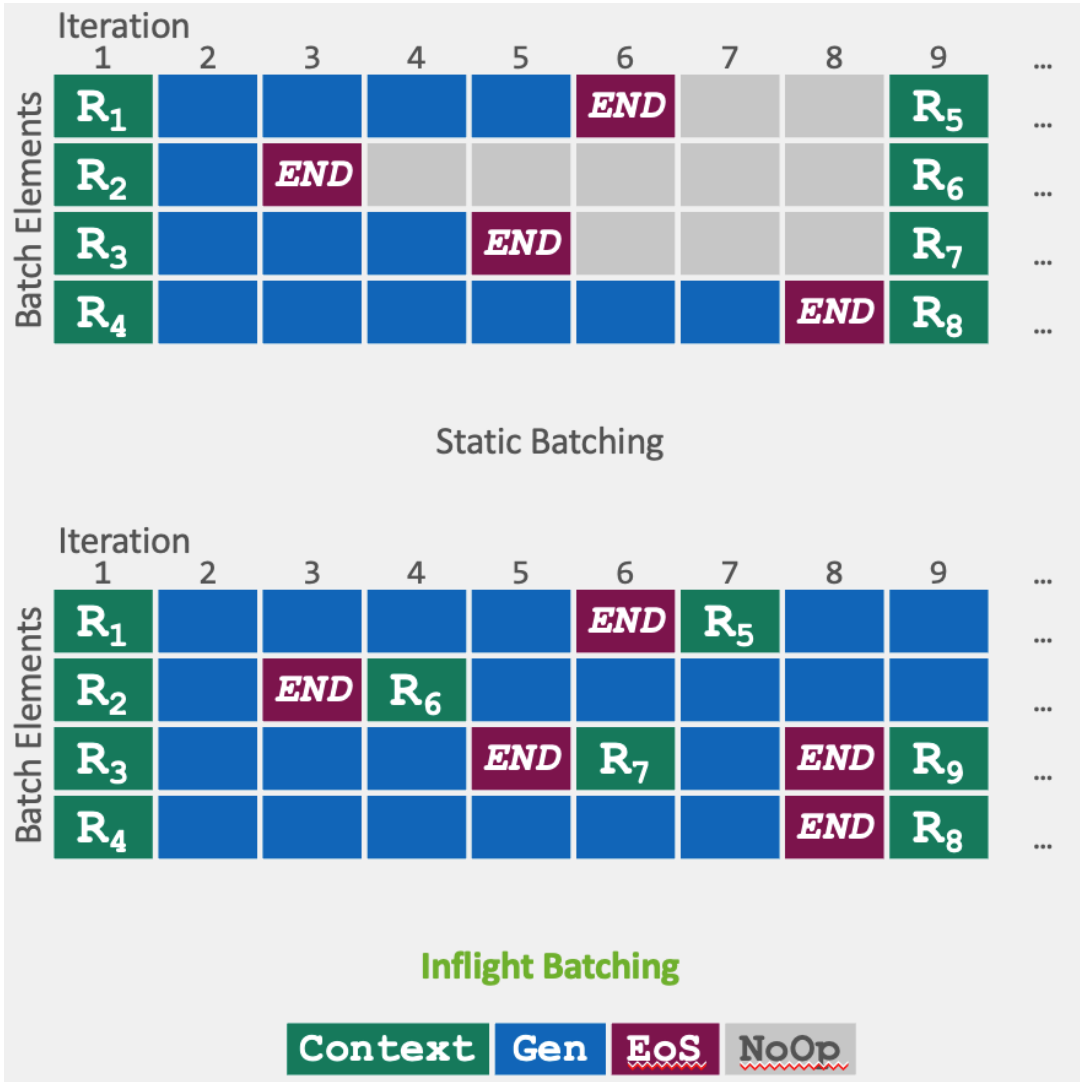


Multi-Head Latent Attention (MLA)



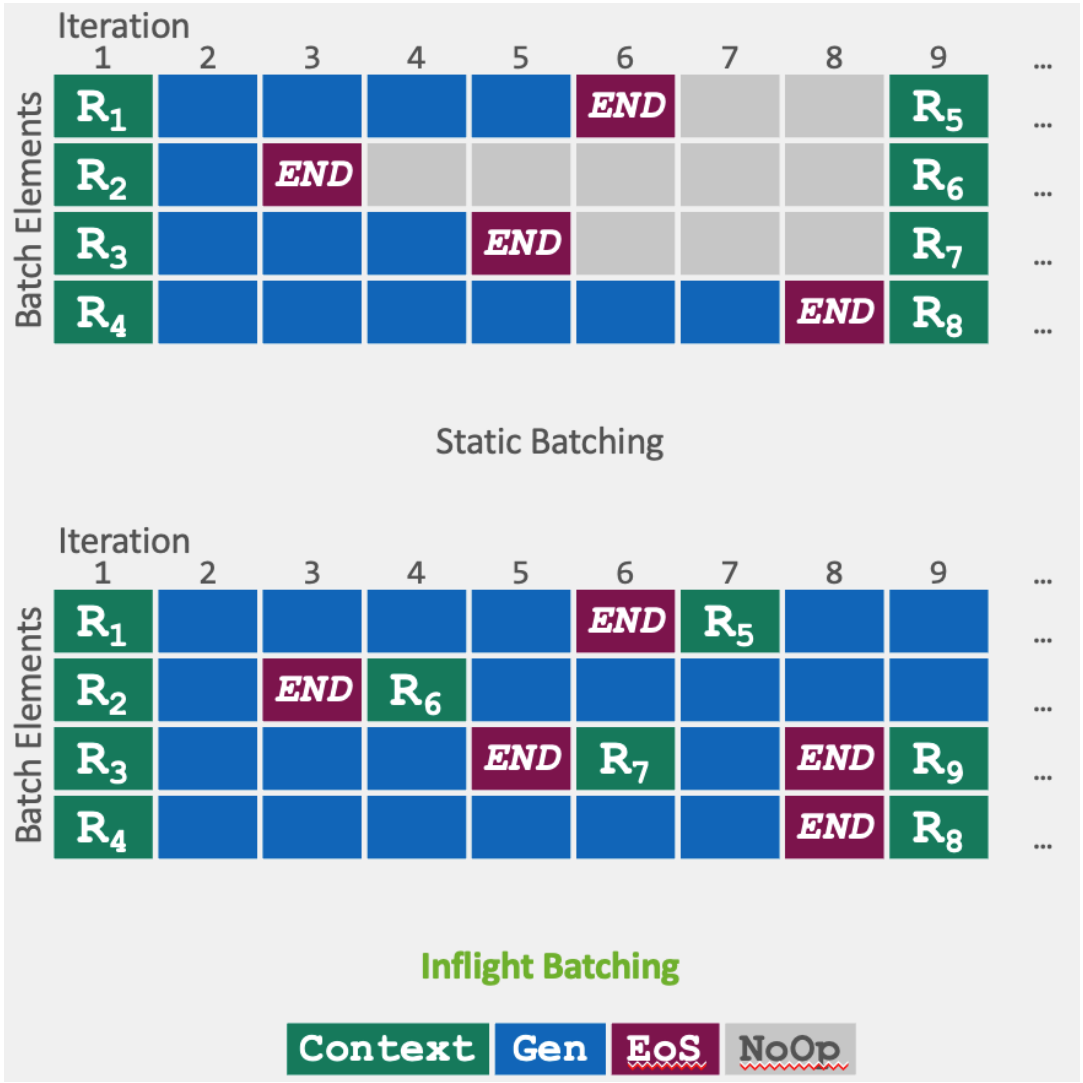
Attention Mechanism	KV Cache per Token (# Element)	Capability
Multi-Head Attention (MHA)	$2n_h d_h l$	Strong
Grouped-Query Attention (GQA)	$2n_g d_h l$	Moderate
Multi-Query Attention (MQA)	$2d_h l$	Weak
MLA (Ours)	$(d_c + d_h^R)l \approx \frac{9}{2} d_h l$	Stronger

Naive batching / static batching



- Batching 是提高推理性能常用做法，但在 LLM 推理场景中，一个 batch 中每个 request 输出长度无法预测。
- 按照 static batching，batch 时延取决于 request 中输出最长的。
- 因此，虽然输出较短 request 已经结束，但是并未释放计算资源，其时延与输出最长的那个 sample/request 时延相同。

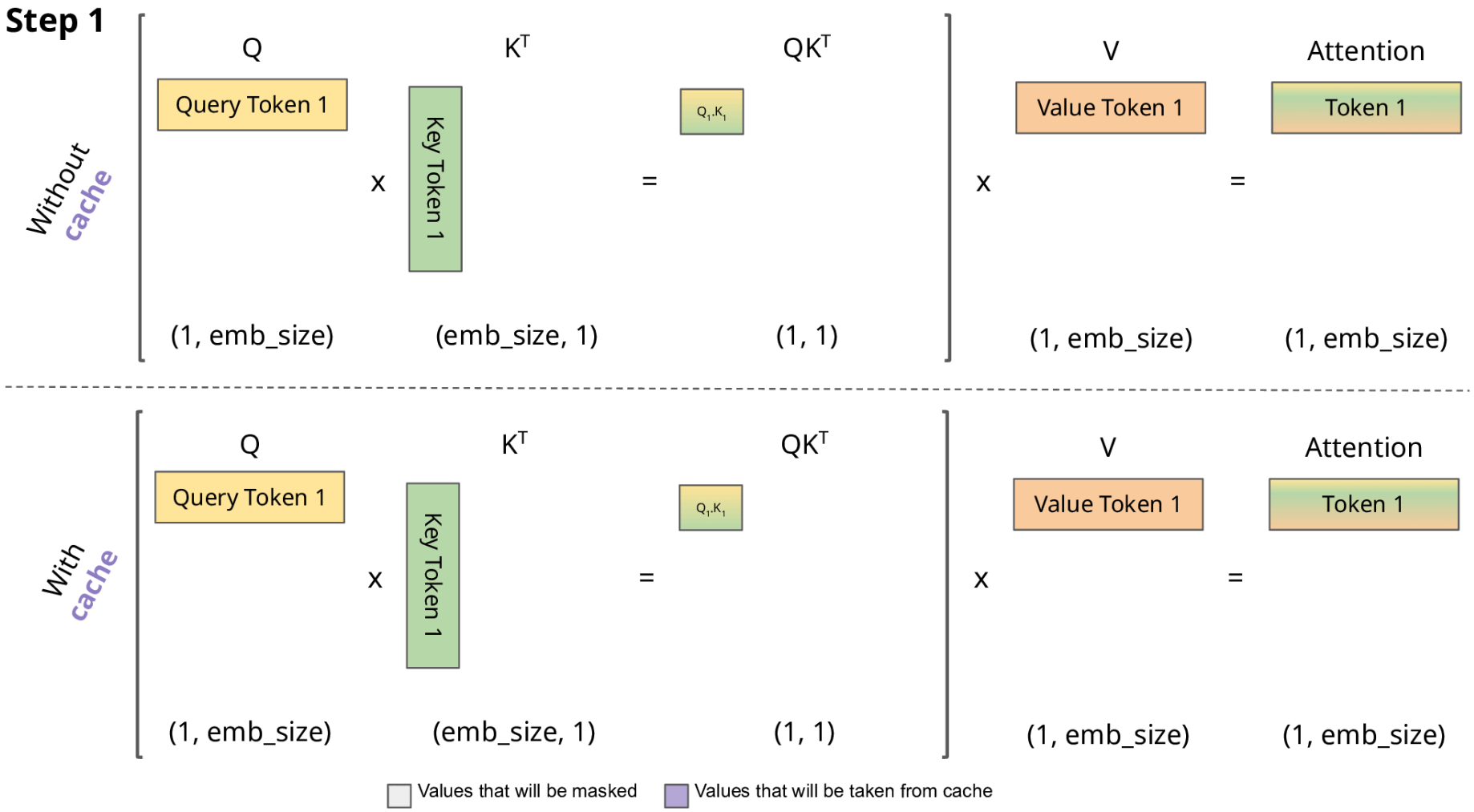
In-Flight Batching/Continuous Batching



- In-flight batching 在已经结束 request 处插入新的 request。实现迭代级调度，其中批量大小由每次迭代决定。不但减少 request 延时，避免资源浪费问题，同时提升整个系统吞吐。
- pre-padding 需要计算，并且与生成阶段计算模式不同，所以不能轻易与 Token 生成进行 batching。In-flight batching 通过超参数管理等待预填充请求和等待序列结束标记的比例。

- Orca: A Distributed Serving System for Transformer-Based Generative Models

KV Cache


Step 1
Without cache
Q
K^T
QK^T
V
Attention
Query Token 1
Key Token 1
Q1.K1
Value Token 1
Token 1
x
=
x
=
(1, emb_size)
(emb_size, 1)
(1, 1)
(1, emb_size)
(1, emb_size)
With cache
Q
K^T
QK^T
V
Attention
Query Token 1
Key Token 1
Q1.K1
Value Token 1
Token 1
x
=
x
=
(1, emb_size)
(emb_size, 1)
(1, 1)
(1, emb_size)
(1, emb_size)
Values that will be masked
Values that will be taken from cache

4. 看大模型趋势

本轮价格战思考

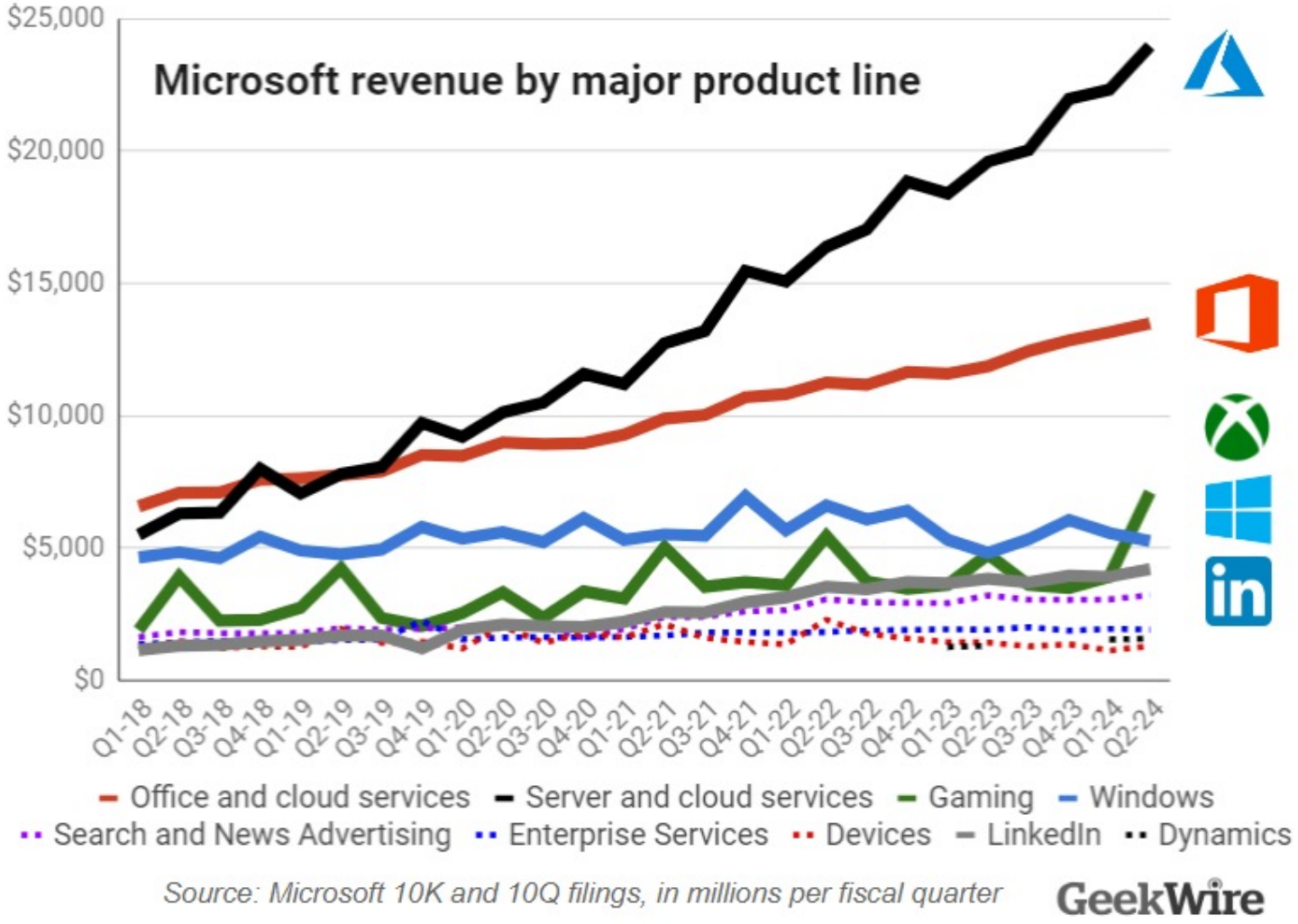
1. **为什么会降价？** （爆发原因）
2. **为什么现在降价？** （24 年年中）
3. **为什么降价这么多？** （小规模大模型接近免费）
4. **是否还会继续降价？** （继续降价的空间）
5. **对行业的影响？** （算力芯片厂商、应用、大模型提供商）

降价潮跟随国外

	Older models	New models
GPT-4 Turbo	GPT-4 8K Input: $0.03 Output: $0.06 GPT-4 32K Input: $0.06 Output: $0.012	GPT-4 Turbo 128K Input: $0.01 Output: $0.03
GPT-3.5 Turbo	GPT-3.5 Turbo 4K Input: $0.0015 Output: $0.002 GPT-3.5 Turbo 16K Input: $0.003 Output: $0.004	GPT-3.5 Turbo 16K Input: $0.001 Output: $0.002
GPT-3.5 Turbo fine-tuning	GPT-3.5 Turbo 4K fine-tuning Training: $0.008 Input: $0.012 Output: $0.016	GPT-3.5 Turbo 4K and 16K fine-tuning Training: $0.008 Input: $0.003 Output: $0.006

- OpenAI 过去一年降价4 次，模型序列 seq 更长，精度更高
- 推理价格下降国外已经成为趋势，国内降价晚于国外
- 但是国内降价幅度大，引起行业地震

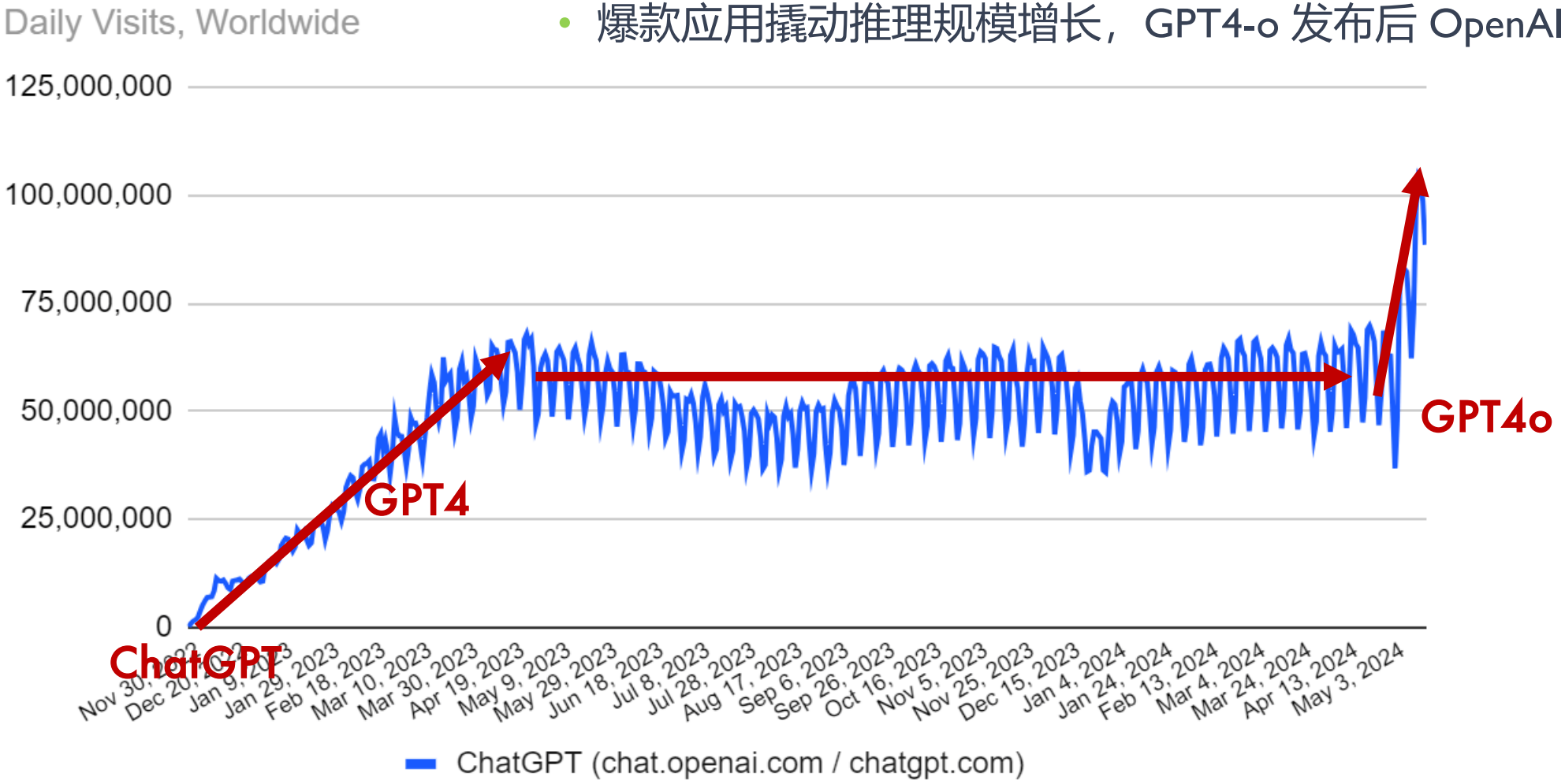
AIGC 对云厂商增长助力



- 国外云巨头加持 AI 后处于高增长，AIGC 对云收入贡献比例稳步提升
- 国内云反而增速放缓，AIGC 对云厂商贡献减少
- 阿里云、腾讯云、百度云、字节火山云都推出对应大模型角逐，华为云？

价格持续下跌需要爆款应用

- 爆款应用撬动推理规模增长，GPT4-o 发布后 OpenAI 访问量激增



对催生应用的商业思考

- 技术进步不是游戏规则改变主要原因，商业决策才是本次降价的根本原因：

1. 美国云重回 20% 高增长，AIGC 对营收贡献占比提升，国内云增速<10%，希望降价提升营收；
2. 在 C 端投放 ROI 越来越低，AI Chat Robot 形态弊端明显：无法下沉到 C 端规模化；
3. 国内大模型应用投资规模落后美国，落地缺少行业突破，希望通过降价助推标杆应用爆发；
4. 需要开发者去创造造应用、产品，然后帮助 L1/L2 厂商，去触达更多潜在需求用户；

百模大战对模型思考

- 24 年国内外开源闭源大爆发，开源 or 闭源之争：

1. 大模型性能同质化严重，开源闭源都在对标 GPT4 等产品，精度在低 1~3% 徘徊；
2. 开源大模型（DeepSpeekV2）赶超部分闭源，第二梯队倾向于开源模型减少试错成本；
3. 降幅大模型参数规模集中在 ~B、~10B 规模，>100B 以及万亿价格持平，卖家不可能亏太多；

对技术发展和计算思考

- 大模型和云上算力绑定会越来越深，推理算力会因为技术发展持续减少：

1. 量化、MOE、KVCache等技术会让推理的成本持续下跌；
2. 小规模大模型的出现，也让大模型产品矩阵更加丰富，可以提供不同价格组合；
3. 推理营由单 token 价格和吞吐组成，随单 tokens 价格下降，吞吐将成为推理服务核心指标；

对技术发展和计算思考

- 小规模大模型将会下沉到更多推理芯片上：

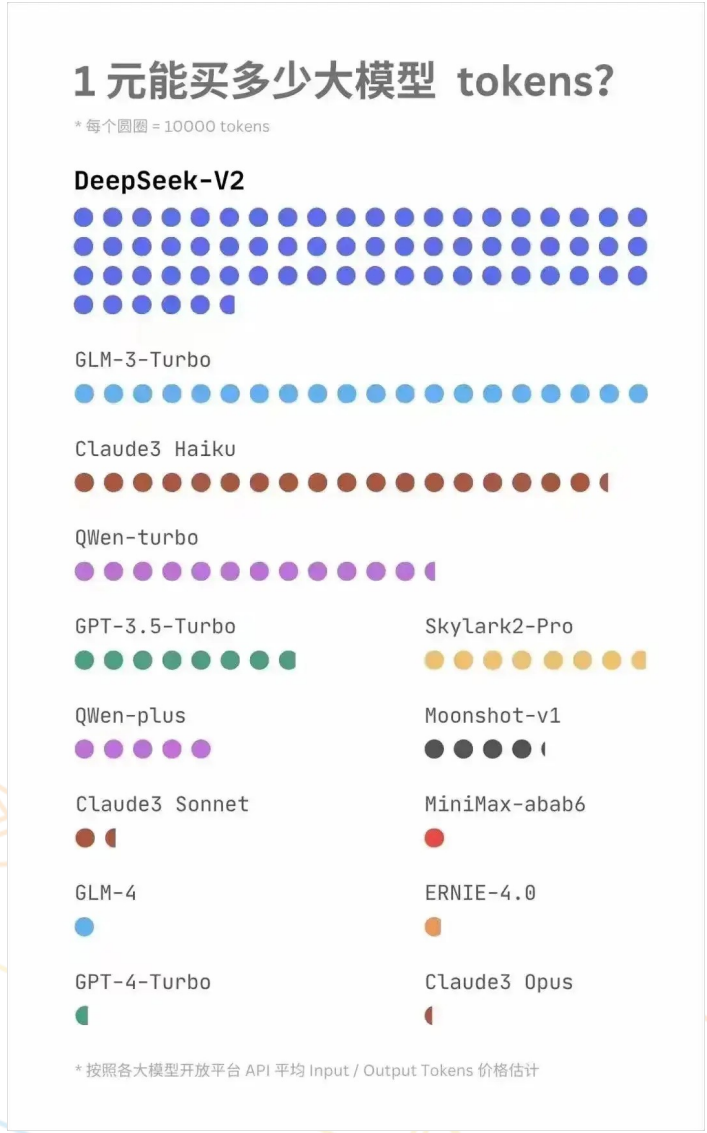
1. 推理极致性价比降加速 LO 行业厂商洗牌，百模大战已经进入下半场，可剩玩家不多；
2. 推理芯片算力提供商不会像训练服务器剩下 NV 和 Ascend，极致性价比之王将会诞生；
3. 推理场景短期对很多厂商是机遇，也是死亡的角逐；

对技术发展和计算思考

1. 对于普通开发者/用户，类似于 2013、2014年流量套餐时代，超出套餐外流量 0.01元/KB ，现今把大模型推理价格下调，对开发者/用户/消费者来说最好 AI 时代。
2. 对于AI公司/百模大战参战商，残酷的淘汰赛开始，没有庞大计算集群资源算力，没有雄厚资本支撑，生存更难。
3. 原来昂贵、遥不可及的算力成本，降低到一种近乎于不要钱形态，对于这个社会、这个行业的冲击，会有多恐怖?

如果你要创业?

- 如果我要建一个 AI 公司，叫什么名字好?



月之暗面

无问苍穹

至暗时刻

无线光年

阶跃星辰

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

参考引用

1. https://cameronrwolfe.substack.com/p/language-model-training-and-inference
2. https://cameronrwolfe.substack.com/p/decoder-only-transformers-the-workhorse
3. https://towardsdatascience.com/transformers-explained-visually-part-2-how-it-works-step-by-step-b49fa4a64f34
4. https://cloud.tencent.com/developer/article/2327739