

长上下文大语言模型如是说 Thus Spake Long-Context LLM

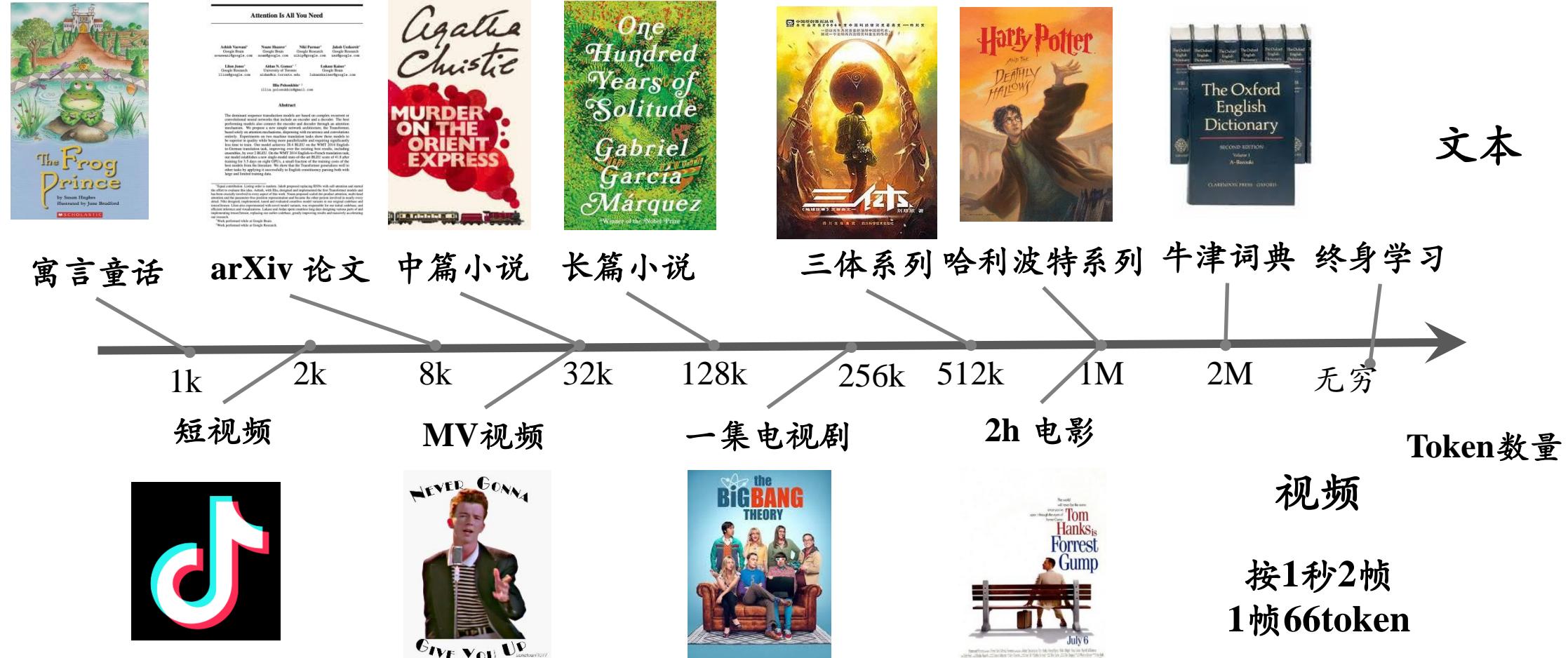
柳潇然

复旦大学自然语言处理实验室

2025年3月27日

长上下文为什么重要?

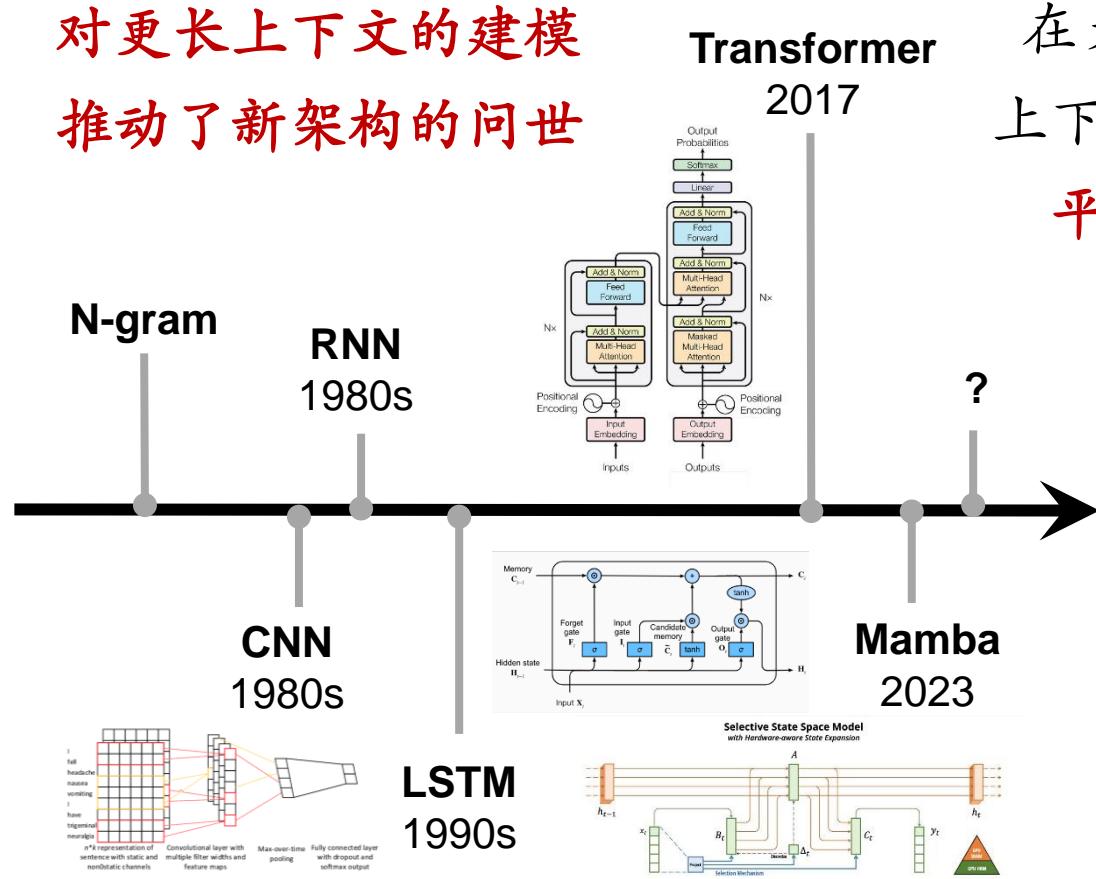
更长的输入输出，或者更长的上下文，给大语言模型带来更多的机遇



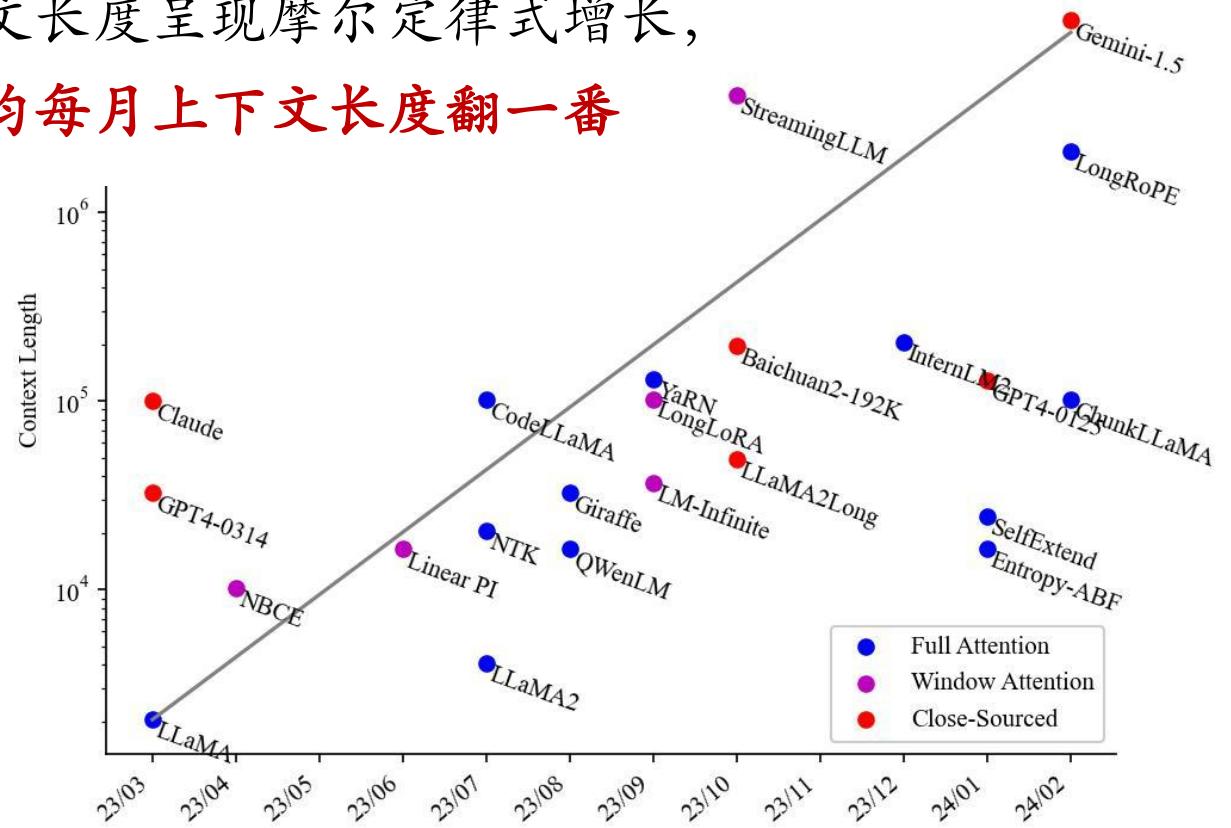
长上下文为什么重要?

长上下文一直是自然语言处理模型，以及大语言模型的核心竞争力

对更长上下文的建模
推动了新架构的问世



在大模型问世的第一年，模型
上下文长度呈现摩尔定律式增长，
平均每月上下文长度翻一番



目录

▶ 架构部分

- ▶ 位置编码与外推；高效缓存；记忆管理；高效注意力、非注意力
- ▶ 我们的工作：ScalingRoPE ICLR24; ReAttention ICLR25; EM2 EMNLP24

▶ 框架部分

- ▶ 三个角度：计算优化、存储优化、分布式；两个场景：训练、推理
- ▶ 我们的工作：IMDC EMNLP24; ParallelTokenizer; CoLLiE EMNLP23

▶ 训练部分

- ▶ 长文预训练、长文后训练；长文提升/激发大模型能力；多模态长文（架构 训练 评测）
- ▶ 我们的工作：LongWanjuan EMNLP24、LongSafety ACL25、Revisit TTS、VideoRoPE ICML25,

▶ 评测部分

- ▶ 长文评测：问答、检索、推理；我们的工作：L-Eval ACL24 Outstanding、DetectiveQA ICLR25
- ▶ 未做回答的问题，总结；同期工作



arXiv



Github



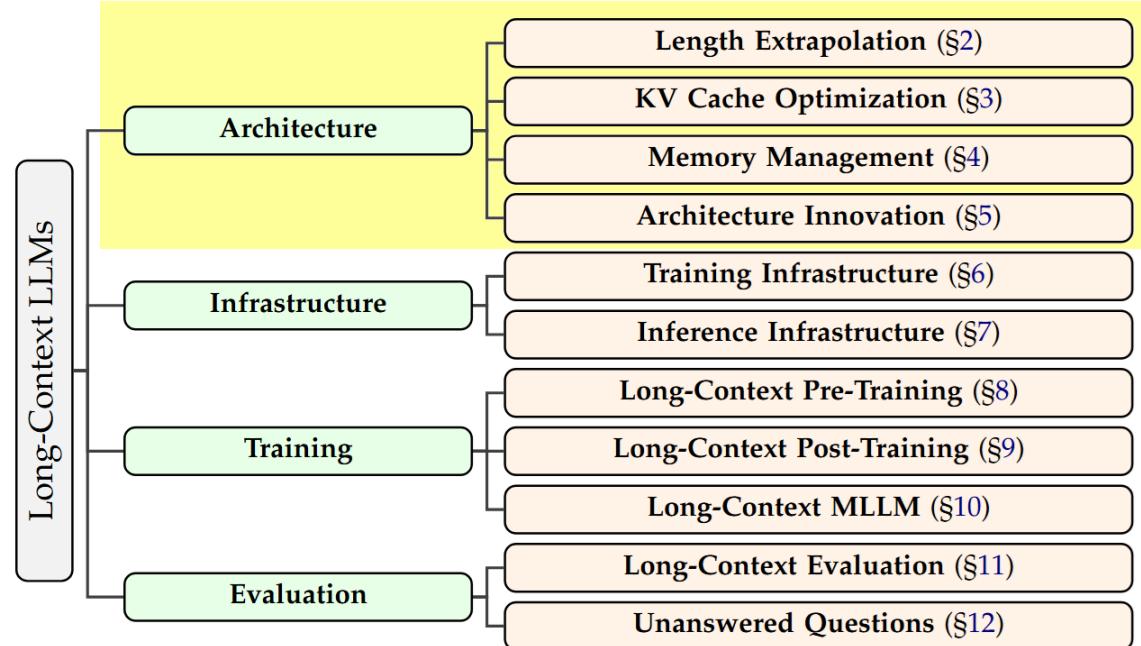
BiliBili



小红书

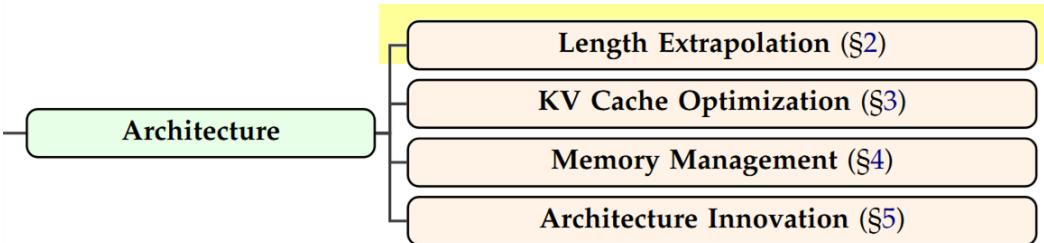
架构部分

- ▶ 对无限长上下文的第一波冲刺！
 - ▶ 位置编码：表示上下文的信息
 - ▶ 长度外推：支持更长的上下文
 - ▶ 高效缓存：限制KV缓存的增长
 - ▶ 记忆管理：对缓存或输入的改与查
 - ▶ 新架构：稀疏注意力、线性注意力、非注意力模型、混合架构

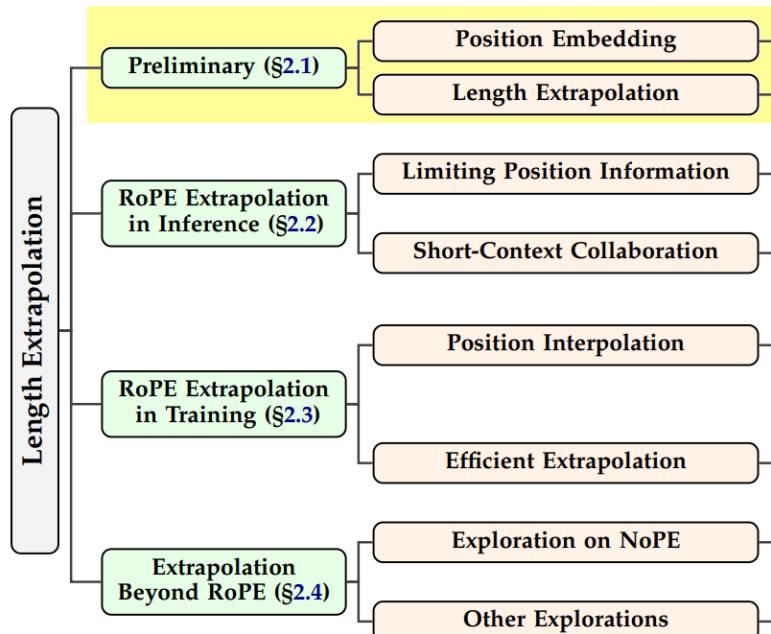


架构部分：位置编码

- ▶ 对无限长上下文的第一波冲刺！
- ▶ 位置编码：表示上下文的信息



$$\theta_n = 10000^{-2n/d} \quad i = 0, \dots, d/2 - 1$$



$$A_{t,s} = \mathbf{x}_t \mathbf{W}_Q \mathbf{R}_{\Theta,t-s}^d \mathbf{W}_K^T \mathbf{x}_s^T = \mathbf{x}_t \mathbf{W}_Q \mathbf{R}_{\Theta,t}^d \left(\mathbf{x}_s \mathbf{W}_K \mathbf{R}_{\Theta,s}^d \right)^T$$

$$\mathbf{R}_{\Theta,t}^d = \begin{pmatrix} \cos t\theta_0 & -\sin t\theta_0 & \cdots & 0 & 0 \\ \sin t\theta_0 & \cos t\theta_0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \cos t\theta_{d/2-1} & -\sin t\theta_{d/2-1} \\ 0 & 0 & \cdots & \sin t\theta_{d/2-1} & \cos t\theta_{d/2-1} \end{pmatrix}$$



arXiv

▶ 旋转位置编码 RoPE

- ▶ 将QK向量两维切分并旋转不同角度
- ▶ 实现了绝对下标与相对距离的统一
- ▶ 能表征无限长文，但存在外推问题

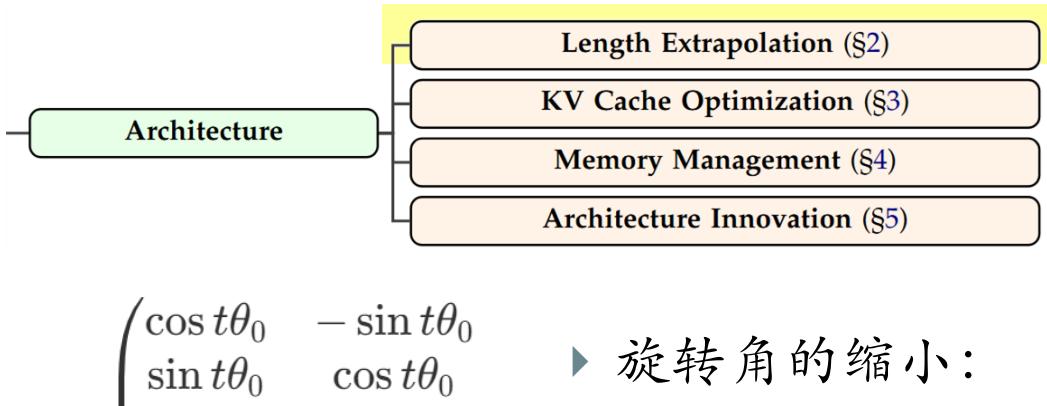
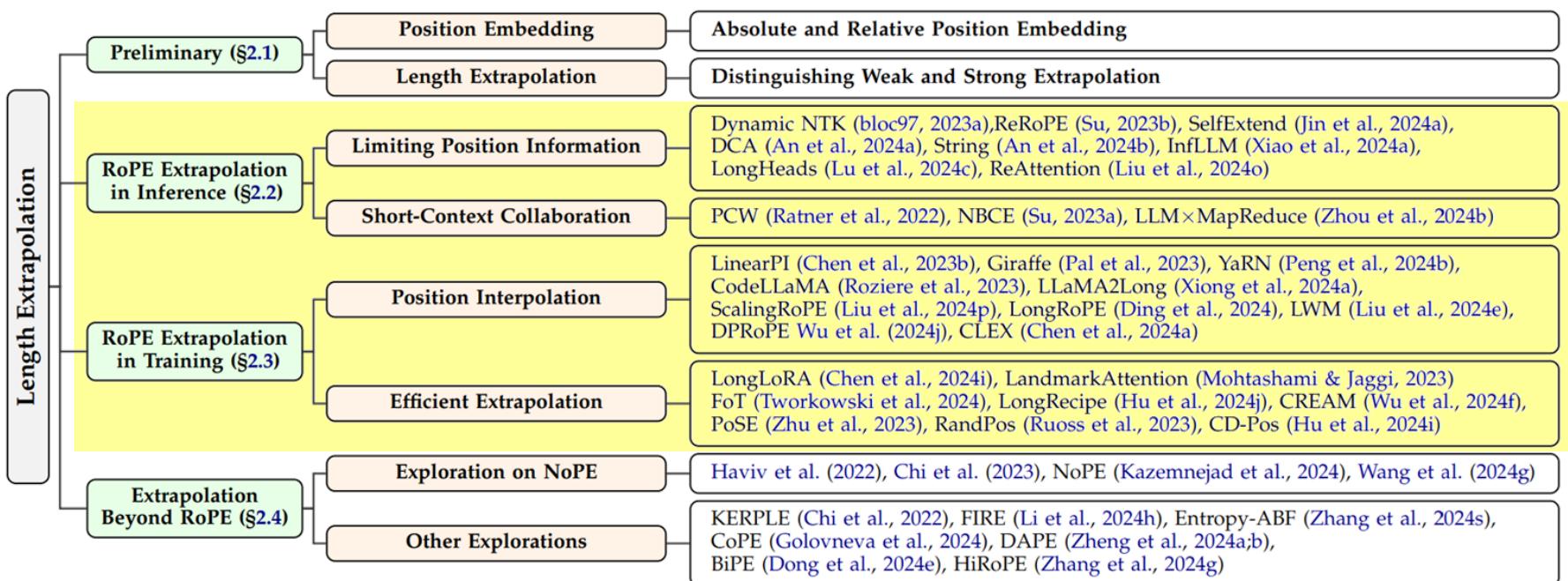


Zhihu

架构部分：长度外推

► 对无限长上下文的第一波冲刺！

- 位置编码：表示上下文的信息
- 长度外推：支持更长的上下文
- 主流的长度外推聚焦RoPE位置信息的内插



$$\begin{pmatrix} \cos t\theta_0 & -\sin t\theta_0 \\ \sin t\theta_0 & \cos t\theta_0 \end{pmatrix}$$

► 旋转角的缩小：
Dynamic NTK

► 位置序号的缩小：
LinearPI、YaRN、
DCA、**LongHeads**

► 应用阶段的考虑

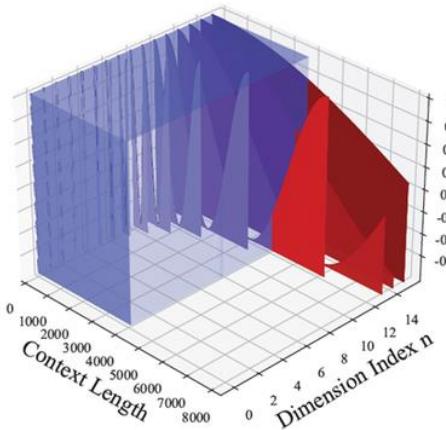
► 训练：高效外推：
PoSE、Landmark

► 推理：以短击长：
PCW、**LongAgent**

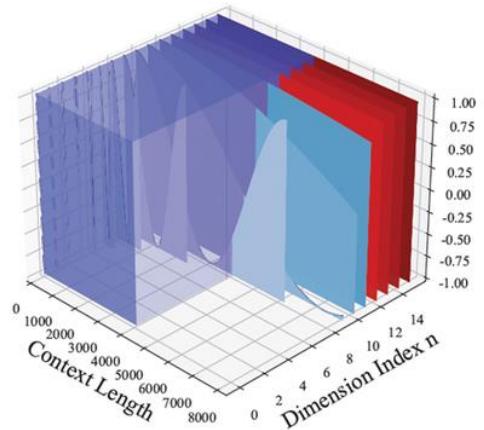
架构部分：ScalingRoPE：RoPE外推的缩放法则

基于位置编码外推法则，将输入长度推广至百万级

外推上限取决于训练阶段见过的位置信息，遇到OOD的就无法外推



(b) RoPE base=10000



(c) RoPE base=1000000

Liu, et al. Scaling laws of rope-based extrapolation, ICLR 2024

结论1. 已知旋转角底数，求最长上下文长度

$$T_{\text{extra}} = \max(T_{\text{tune}}, 2\pi \cdot \beta^{\left| \frac{d}{2} \log_{10000} \frac{T_{\text{train}}}{2\pi} \right|^2})$$

结论2. 已知外推的上界，求底数的最小取值

$$\beta_{\text{extra}} = 10000^{\log_{T_{\text{train}}} \frac{T_{\text{extra}}}{2\pi}}$$

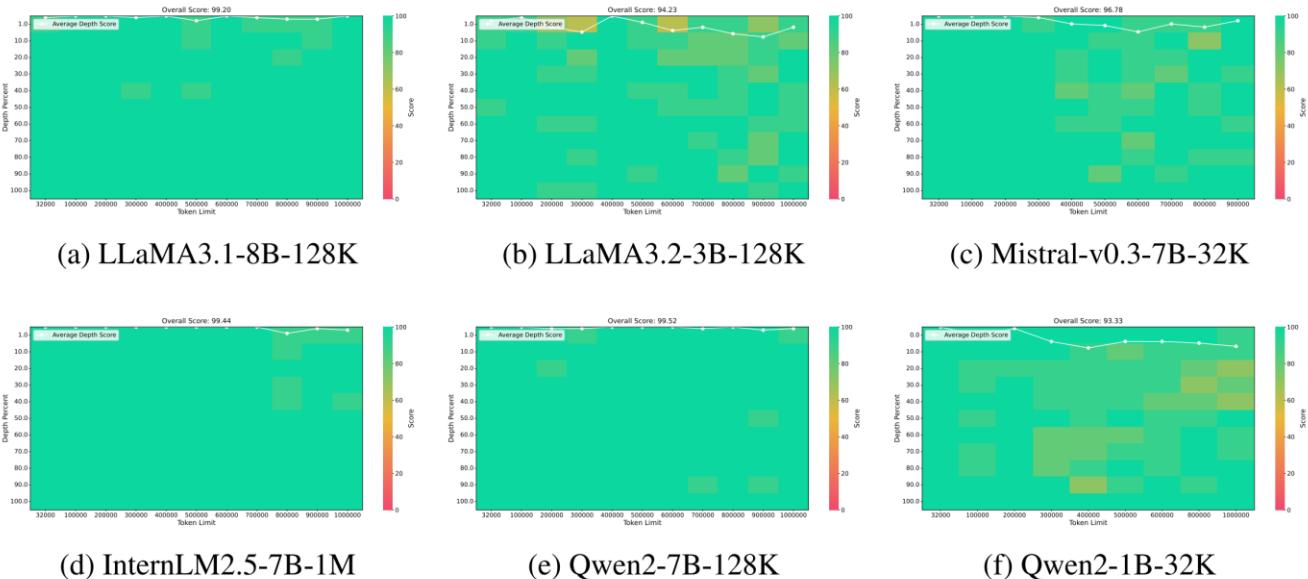
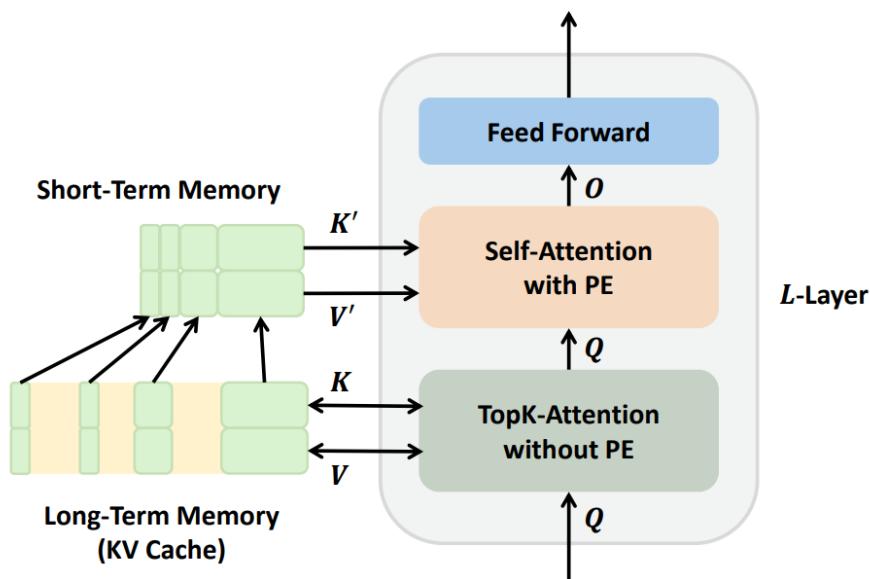
但是，这两个结论并不是外推的全部

架构部分：ReAttention：二次注意力无限外推

针对自注意力外推的三重挑战，即插即用实现无限外推

通过全局筛选后注意力克服位置信息
OOD、注意力熵上升、全上下文感知

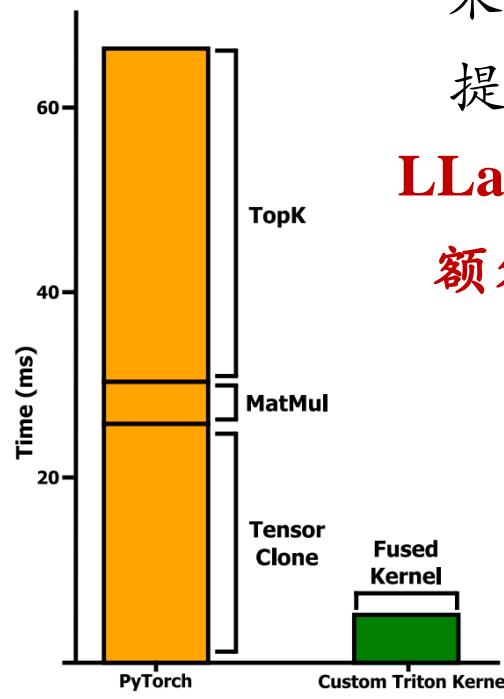
将LLaMA3.1-8B、LLaMA3.2-3B等
主流模型上下文窗口外推至1M+



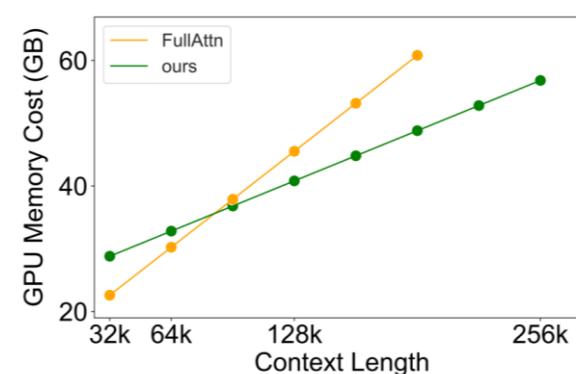
Liu, and Li, et al. ReAttention: Training-Free Infinite Context with Finite Attention Scope, ICLR 2025

架构部分：ReAttention：二次注意力无限外推

不增加计算，降低额外存储，4卡1B模型跑动4M大海捞针



采用FlashAttention的思想
提升TopkAttention的效率
**LLaMA3.1-8B一张A800无需
额外时间开销推理400K长**



不仅在**通用基准上领先同类方法**
还能将1B模型上下文窗口扩至4M
即插即用放大上下文125x

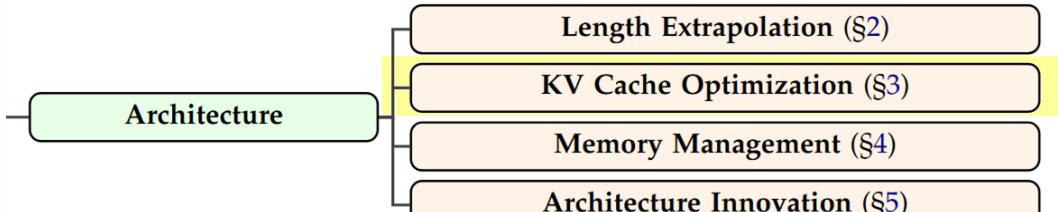
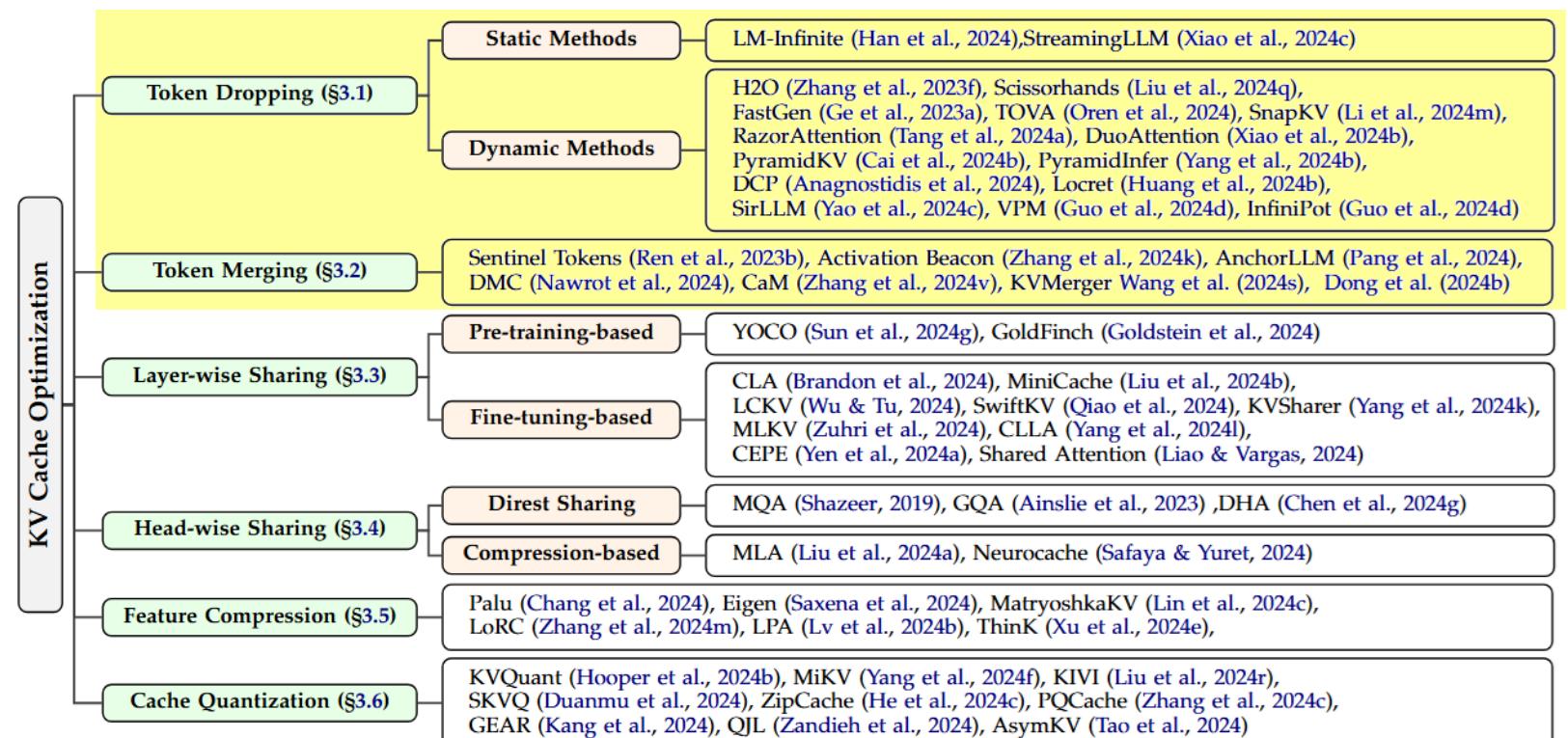


Liu, and Li, et al. ReAttention: Training-Free Infinite Context with Finite Attention Scope, ICLR 2025

架构部分：高效缓存

► 对无限长上下文的第一波冲刺！

- 位置编码：表示上下文的信息
- 长度外推：支持更长的上下文
- 高效缓存：限制KV缓存的增长



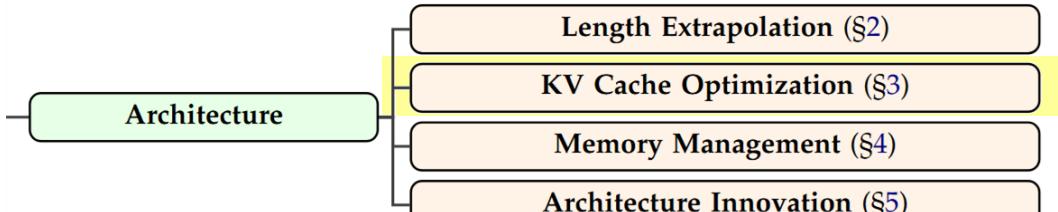
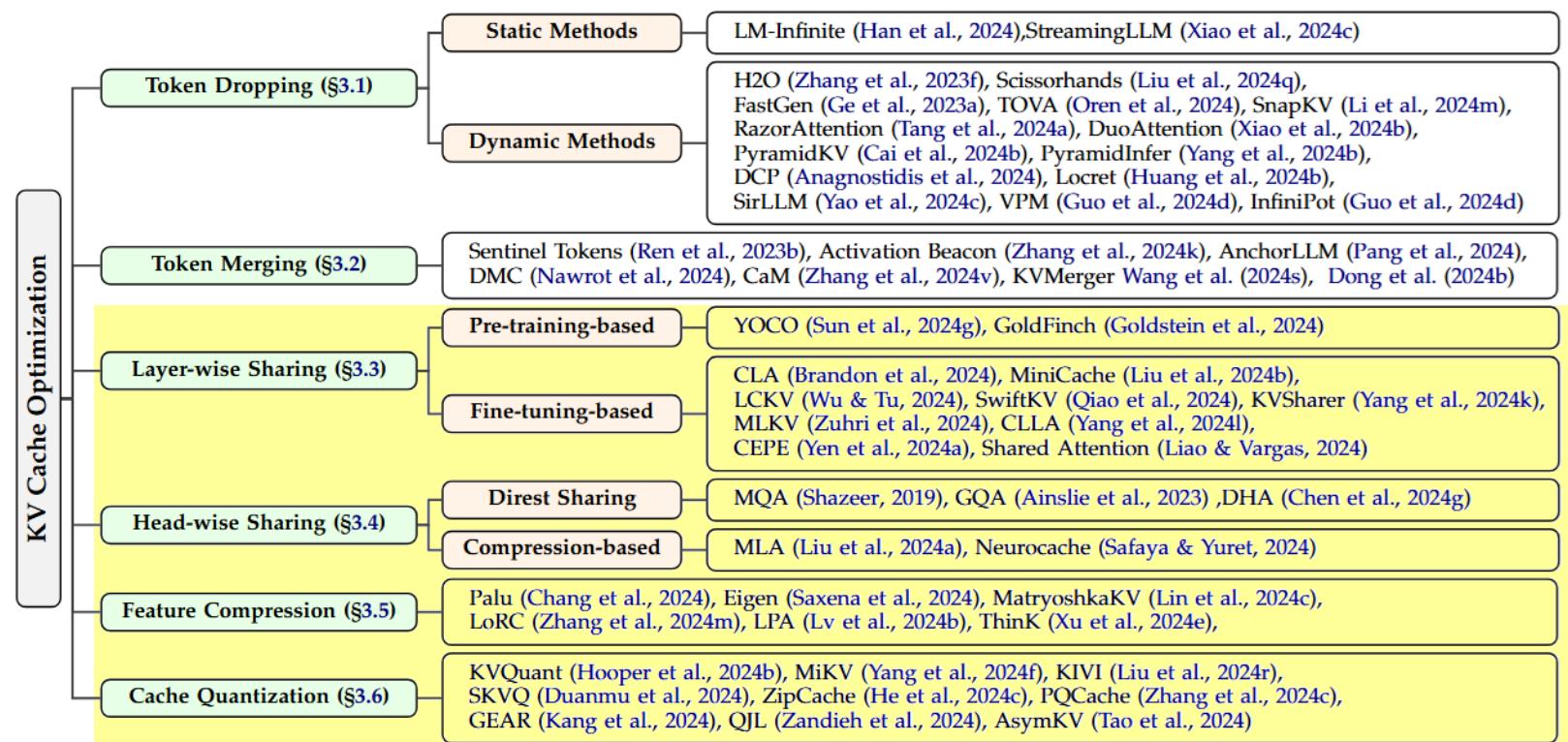
$$\text{sizeof(KV)} = 2 * \text{seq_len} * \text{num_kv_heads} * \text{num_layers} * \text{head_dim} * \text{sizeof(dtype)}$$

- 序列维度探讨最多
- token丢弃
 - 关键是丢弃token准则
- 静态：StreamingLLM
- 过去累积：H2O
- 未来累积：SnapKV
- token压缩

架构部分：高效缓存

► 对无限长上下文的第一波冲刺！

- 位置编码：表示上下文的信息
- 长度外推：支持更长的上下文
- 高效缓存：限制KV缓存的增长



$$\text{sizeof(KV)} = 2 * \text{seq_len} * \text{num_kv_heads} * \text{num_layers} * \text{head_dim} * \text{sizeof(dtype)}$$

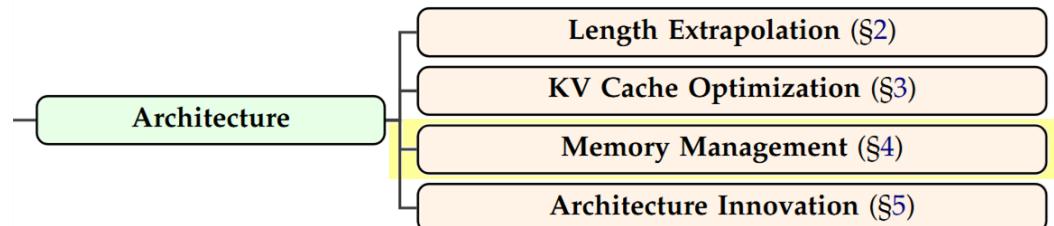
- 每个维度同等重要
- 头和层的压缩
- 基于预训练： YOCO, GQA、MLA
- 基于微调： CEPE, **MHA2MLA**
- 特征和存储的压缩
- 缓存量化： KIVI

架构部分：记忆管理

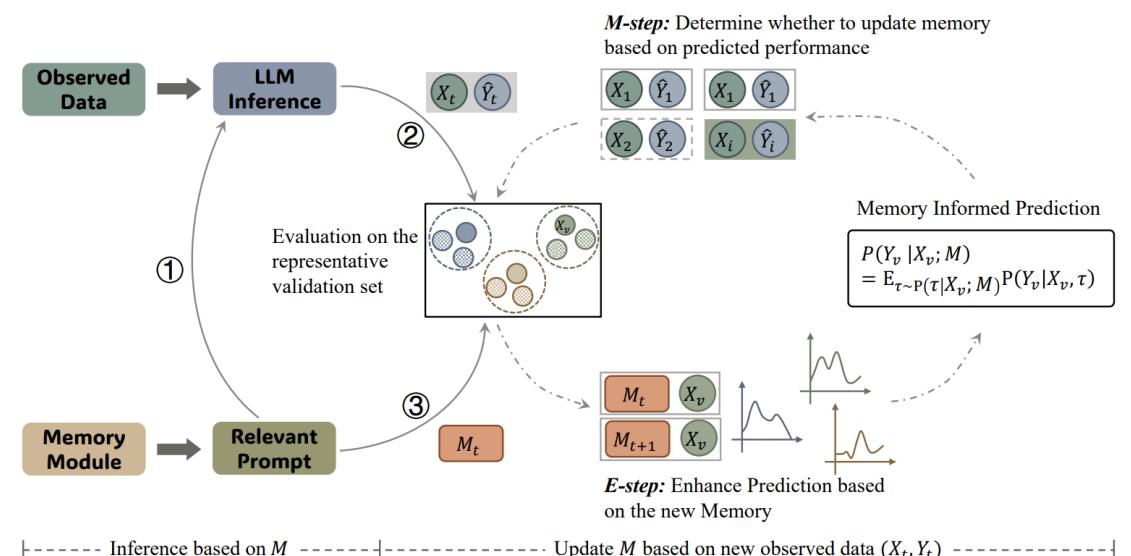
► 对无限长上下文的第一波冲刺！

- 位置编码：表示上下文的信息
- 长度外推：支持更长的上下文
- 高效缓存：限制KV缓存的增长
- 记忆管理：对缓存或输入的改与查

	Cache-Based Memory	Text-Based Memory
Read-Only	§4.1.1 MemTrans (Wu et al., 2022) AutoCompressor (Chevalier et al., 2023) ICAE (Ge et al., 2023b) PromptCache (Gim et al., 2024)	§4.2.1 MemWalker (Chen et al., 2023a) LongRAG (Zhao et al., 2024d) Self-Route (Li et al., 2024r) RAG2.0 (ContextualAI, 2024)
Writable	§4.1.2 Transformer-XL (Dai et al., 2019) RMT (Bulatov et al., 2022) MemoryLLM (Wang et al., 2024n) CAMELoT (He et al., 2024d) Memory ³ (Yang et al., 2024c)	§4.2.2 MemGPT (Packer et al., 2023) LongLLMLingua (Jiang et al., 2023b) RecurrentGPT (Zhou et al., 2023) MemoryBank (Zhong et al., 2024b)



EM算法实现显式文本记忆，动态更新可解释

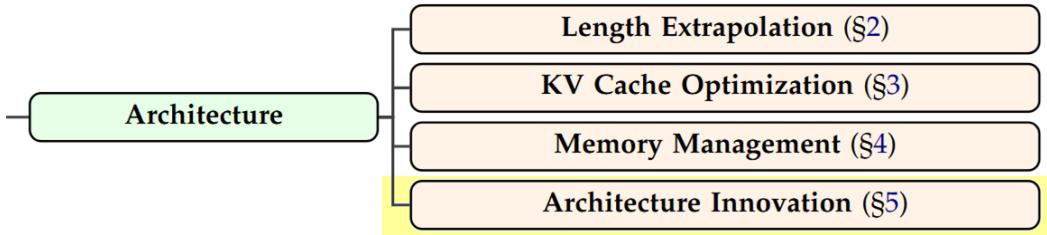


Yin, et al. Explicit memory learning with expectation maximization, EMNLP 2024

架构部分：新架构

▶ 对无限长上下文的第一波冲刺！

- ▶ 位置编码：表示上下文的信息
- ▶ 长度外推：支持更长的上下文
- ▶ 高效缓存：限制KV缓存的增长
- ▶ 记忆管理：对缓存或输入的改与查
- ▶ 稀疏注意力：MInference、NSA、MoBA、XAttention
- ▶ 线性注意力：RetNet、Lightning、DeltaNet、MoM
- ▶ 非注意力：xLSTM、RWKV系列、Mamba系列
- ▶ 混合架构：Jamba系列、MiniMax-01、T1
- ▶ 非自回归：Mercury dLLM、LLaDA、BD3-LM

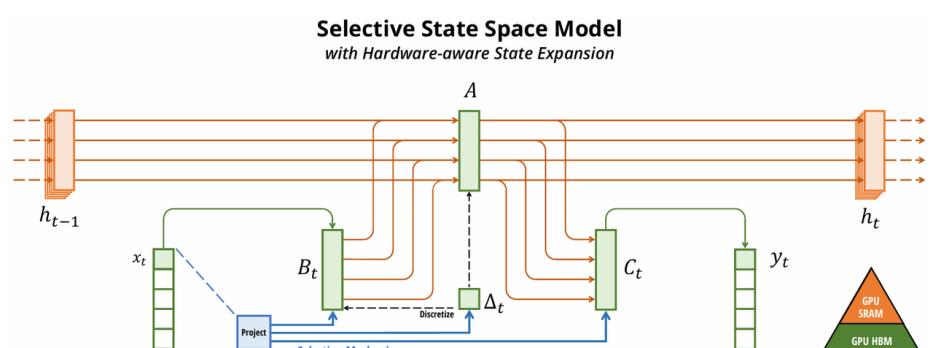
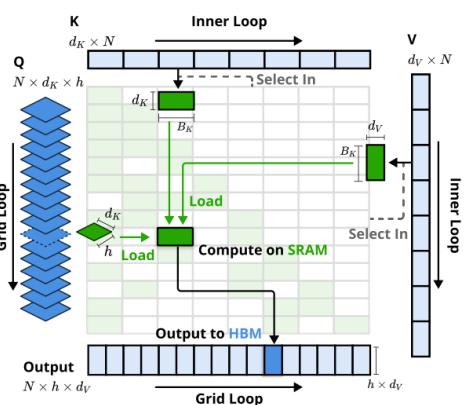
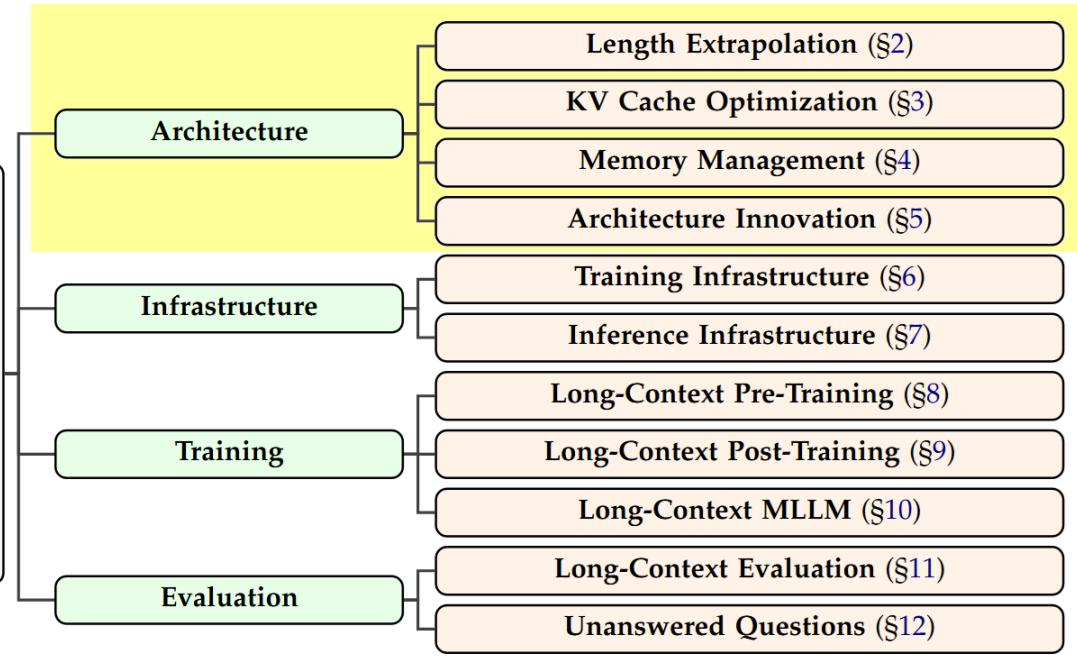
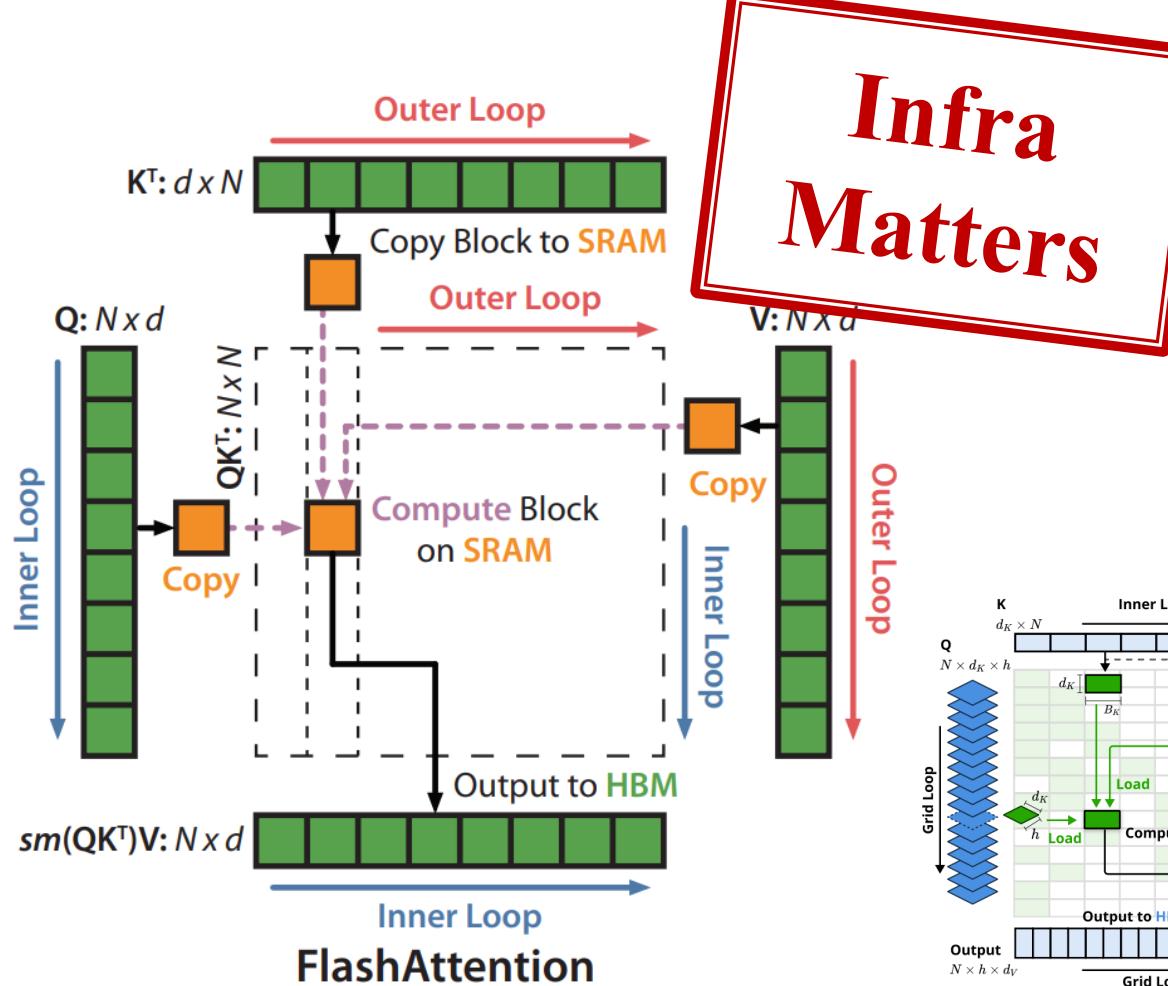


▶ 一些想法

- ▶ 动态检索已成为稀疏注意力的常规操作
- ▶ 非注意力在检索上存在硬伤，但不断有尝试改善
- ▶ 面向长推理场景，**混合架构**和**非自回归**还会继续发力
- ▶ 无论是标准注意力还是高效注意力都离不开**算子优化**

框架部分

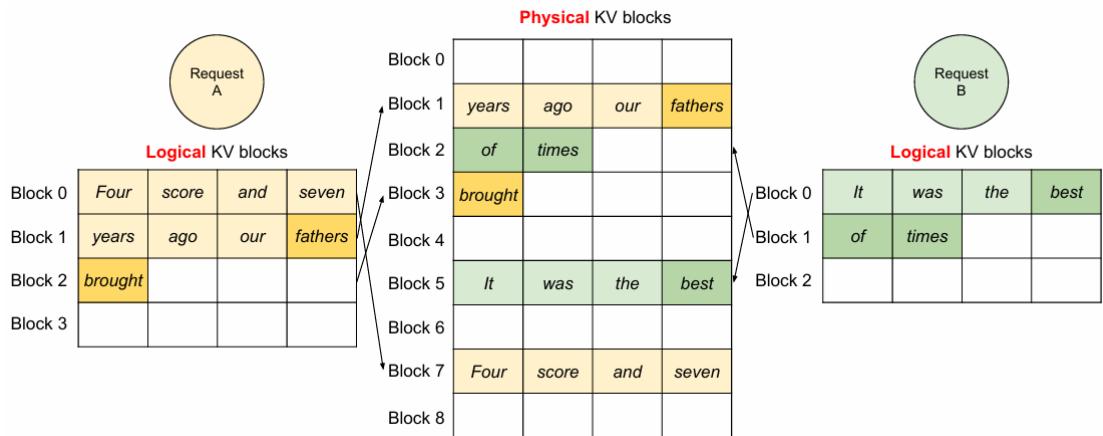
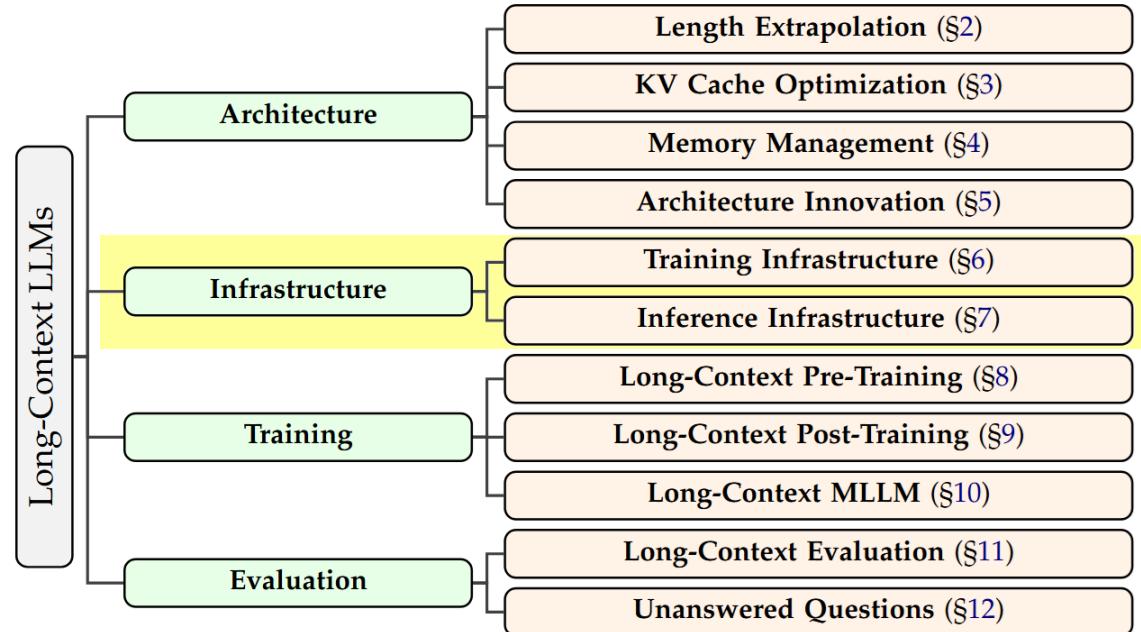
▶ 什么架构都离不开框架 (Infrastructure)



框架部分

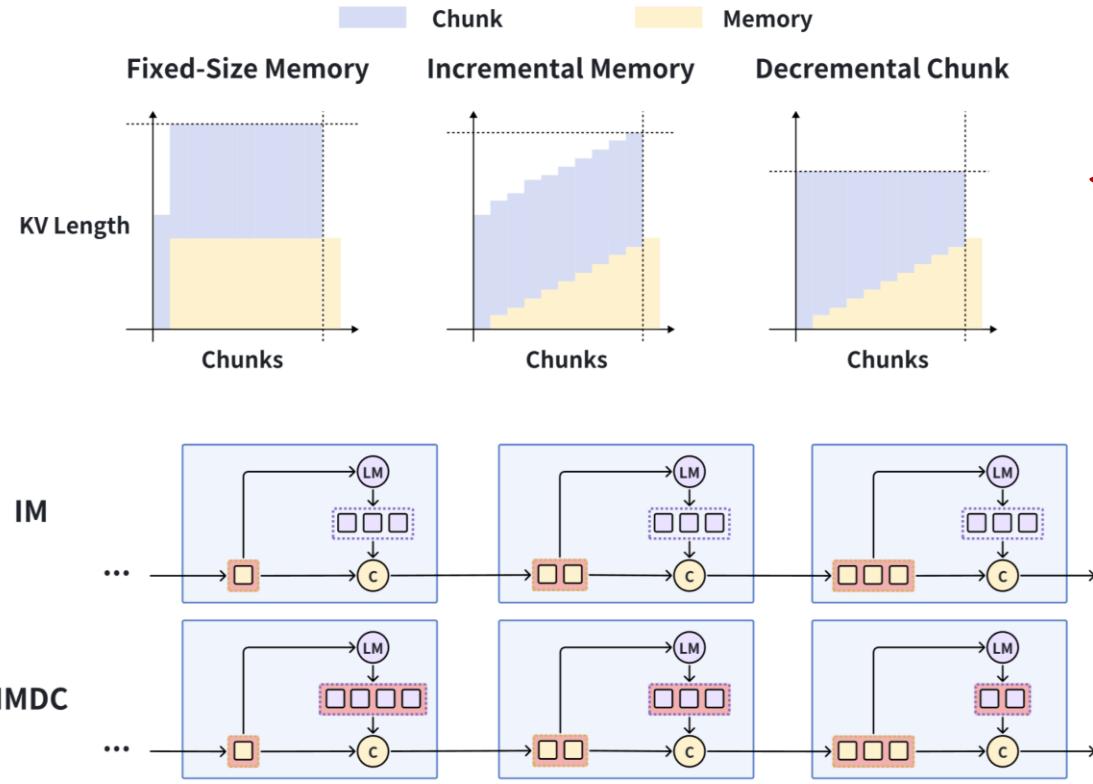
► 理性的沉思与工程的尝试

- 两个场景、三个角度
- 训练框架优化
 - 分布式: RingAttn、DS-Ulysses
 - 计算: 定制化训练算子、调度优化
 - 内存: 重计算、碎片缓解、Offload
- 推理框架优化
 - 分布式: 上下文并行、PD分离
 - 计算: 定制化推理算子、缓存重用
 - 内存: 分块预填充、PagedAttention

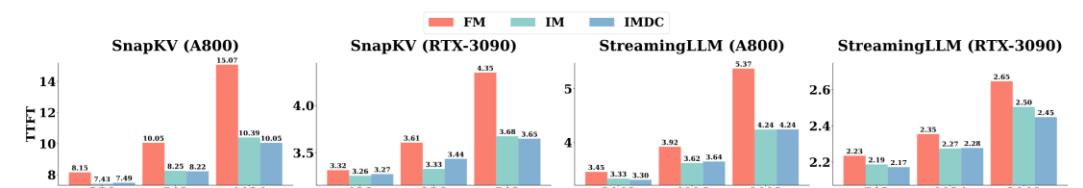


框架部分：IMDC：高效分块预填充策略

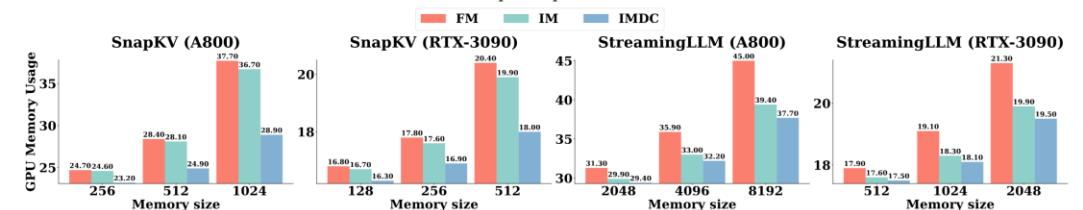
根据显存占用动态调整分块大小，提升计算与存储效率



分块大小随显存占用增大而降低，
提升初始阶段显存占用效率，降低
峰值显存占用，减少预填充时延



(a) TTFT Comparison (Seconds). TTFT (Time To First Token) refers to the time cost associated with the model encoding the input sequence.

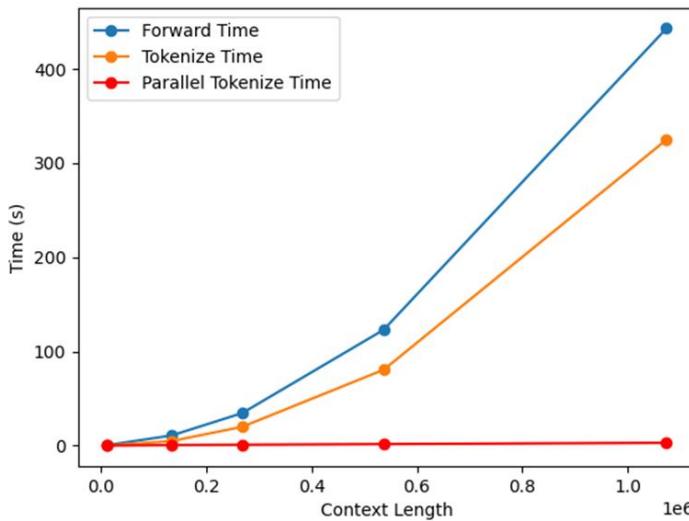


Zeng, et al. Memorize Step by Step: Efficient Long-Context Prefilling with Incremental Memory and Decremental Chunk, EMNLP 2024

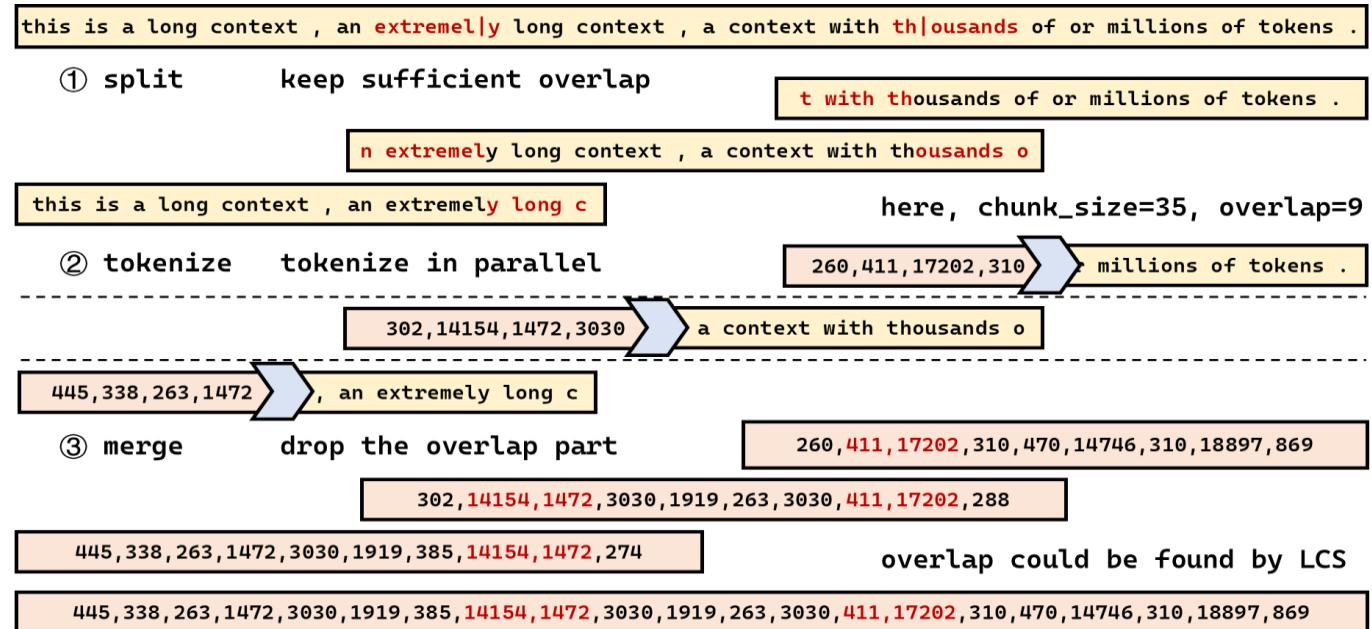
框架部分：ParallelTokenizer：长文并行分词器

首次提出长文分词效率问题，并行分词极大提升训练推理效率

分词和注意力的耗时都随上下文长度显著增加



将文本切分为带重叠的块，**并行分词降低gpu等待时间，提升训练tgs、降低推理时延**



框架部分：高效训练框架

高效微调框架 **CoLLiE**, 高效预训练框架 **InterEvo**



CoLLiE

MOSS的后训练框架，可以帮助您从零开始训练大模型的工具箱，集成了现有的并行策略、高效参数微调方法和高效优化器



Github



arXiv



InternLM的预训练框架，一个轻量的训练框架，实现无需大量依赖关系的大模型预训练，**支持RingAttention/Deepspeed-Ulysses**



Github

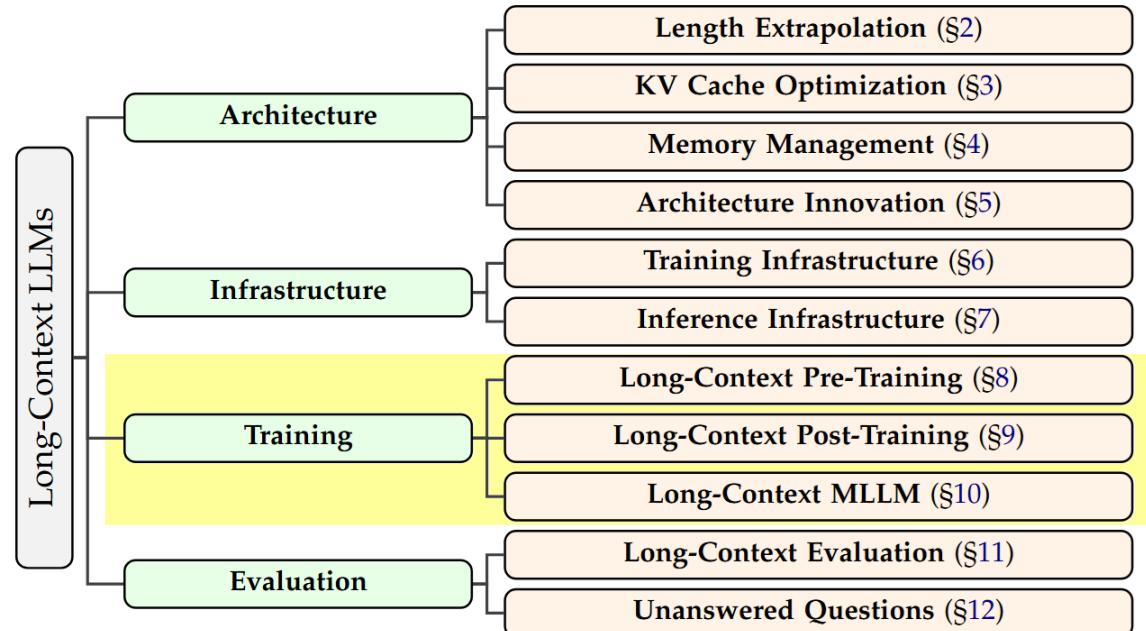


arXiv

Lv & Zhang, et al. CoLLiE: Collaborative Training of Large Language Models in an Efficient Way, EMNLP 2023

训练部分：纯文本训练

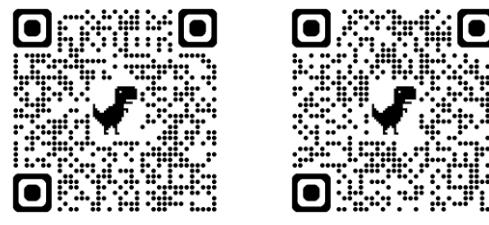
- ▶ 对无限长上下文的第二波冲刺！
- ▶ 预训练：数据质量与数量
 - ▶ 混训、配比、筛选、构造
- ▶ 后训练：任务形式与长度
 - ▶ 长输入短输出、短输入长输出
 - ▶ 后训练之后：长文提升/激发模型能力
 - ▶ Test-Time Training：侧重长上文
 - ▶ Test-Time Scaling：侧重长下文
- ▶ **TTT和TTS是目前以及可预见的一段时间内长文发展的重点方向**



训练部分： LongWanjuan：长文预训练数据质量

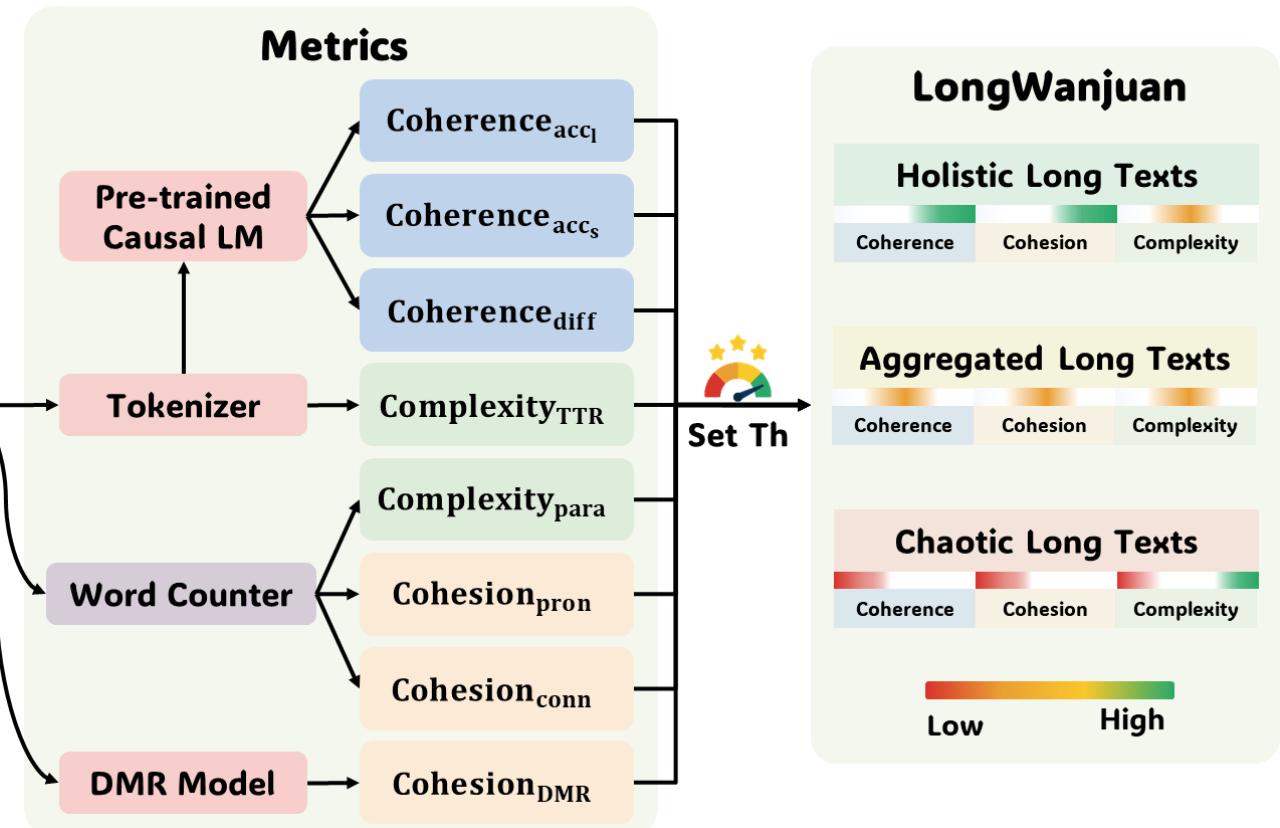
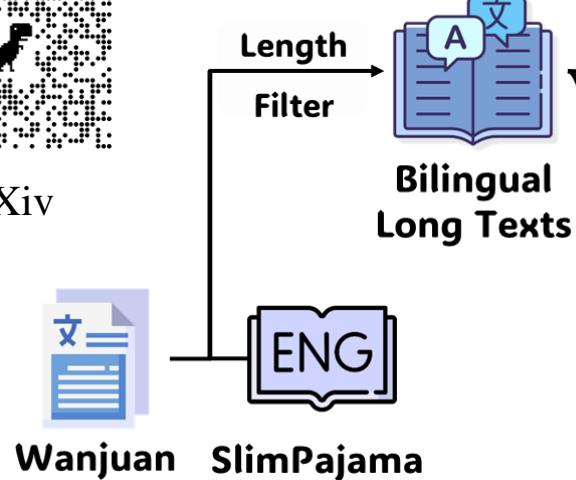
长文质量定义尝试，应用于InternLM2/2.5/3和MOSS预训练

设计模型与规则打分筛选过滤
能实现LongBench/LEval稳定提点



Github

arXiv

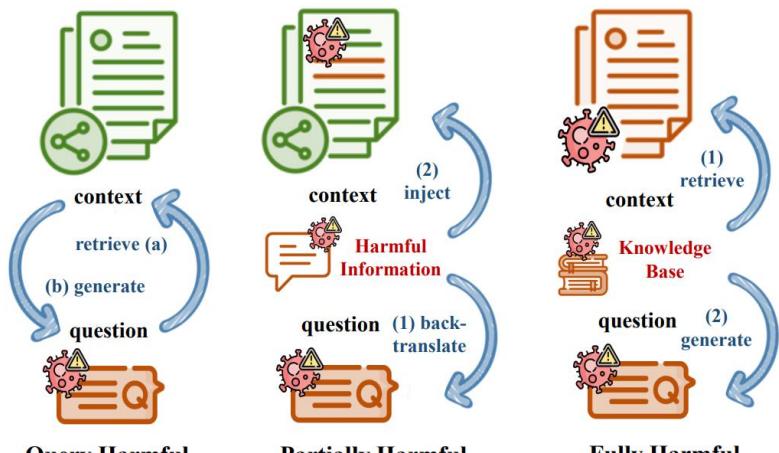


Liu & Lv, et al. LongWanjuan: Towards Systematic Measurement for Long Text Quality, EMNLP 2024

训练部分： LongSafety：长文后训练安全对齐

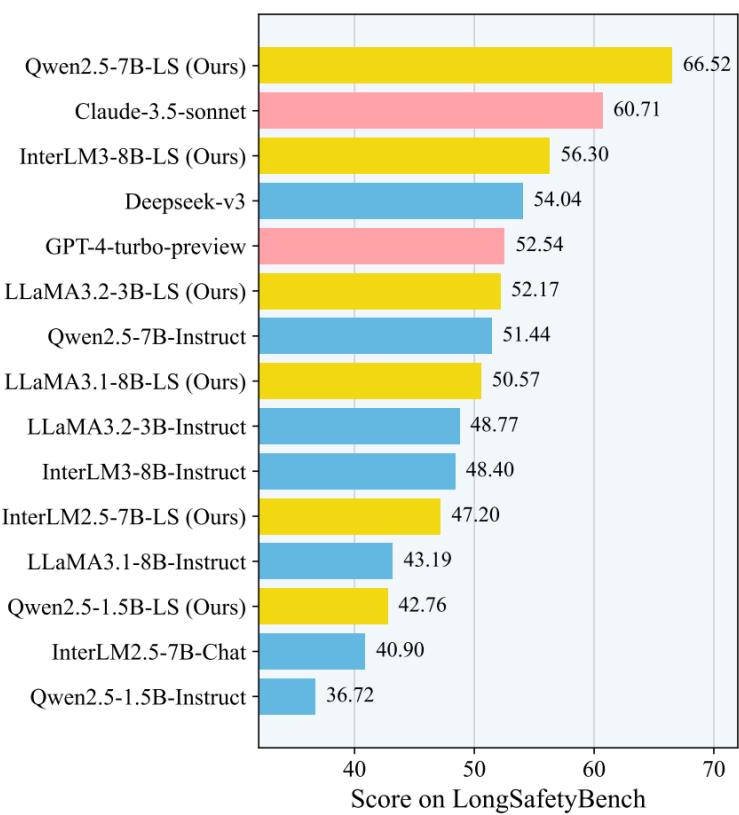
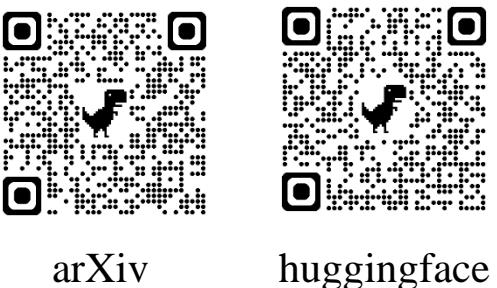
长上下文之罪，首次系统分析大模型长文安全的评估与提升

系统分析长文安全问题
首个聚焦长文安全的训练集



(a) Three long-context safety scenarios with corresponding pipelines

少量微调实现**长短文安全与性能全面提升**
Qwen2.5-7B 长文安全
超越Claude-3.5-Sonnet



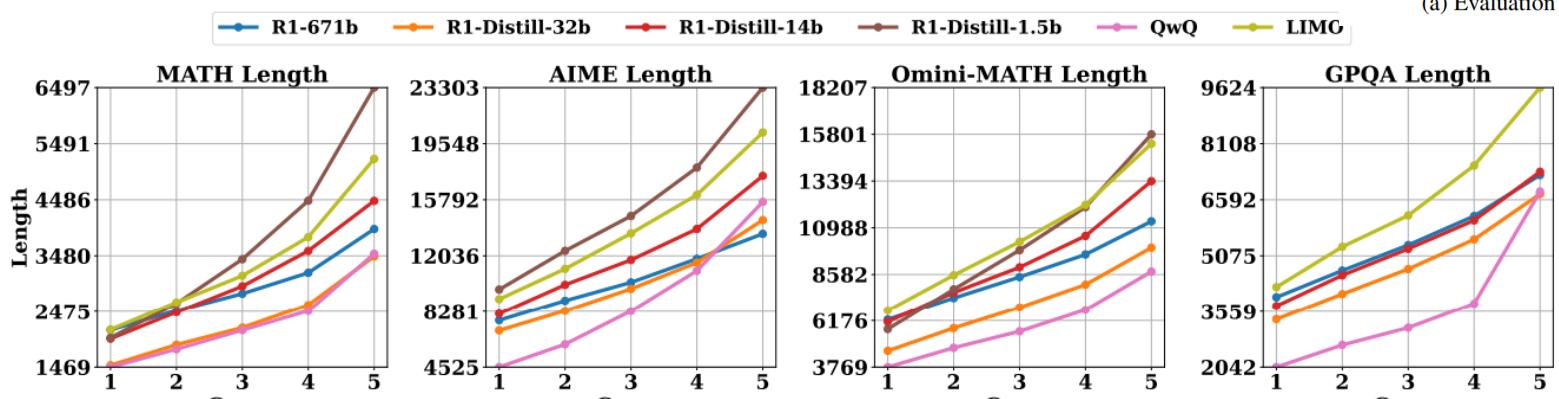
Huang, Liu and Zhou, et al. LongSafety: Enhance Safety for Long-Context LLMs, ACL 2025 under review

训练部分： Revisit TTS：长推理提升大模型能力

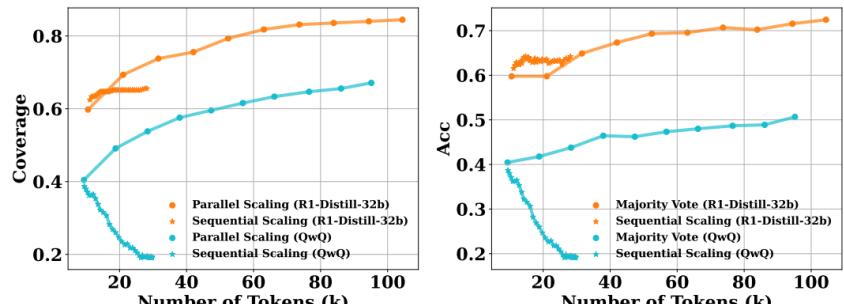
发现当前推理模型扩展性的不足，提出并行推理扩展修正

发现R1、QwQ、LIMO等模型的推理效果
和推理长度和修正次数呈现负相关

通过并行推理取代顺序推理修正scaling曲线

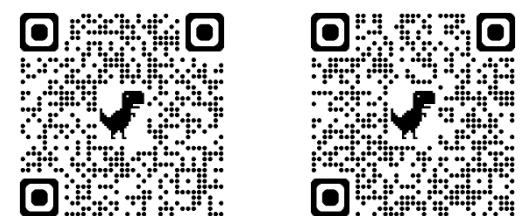


(a) Evaluation for Solution length.



(a) Evaluation on Coverage.

(b) Evaluation on Accuracy



arXiv

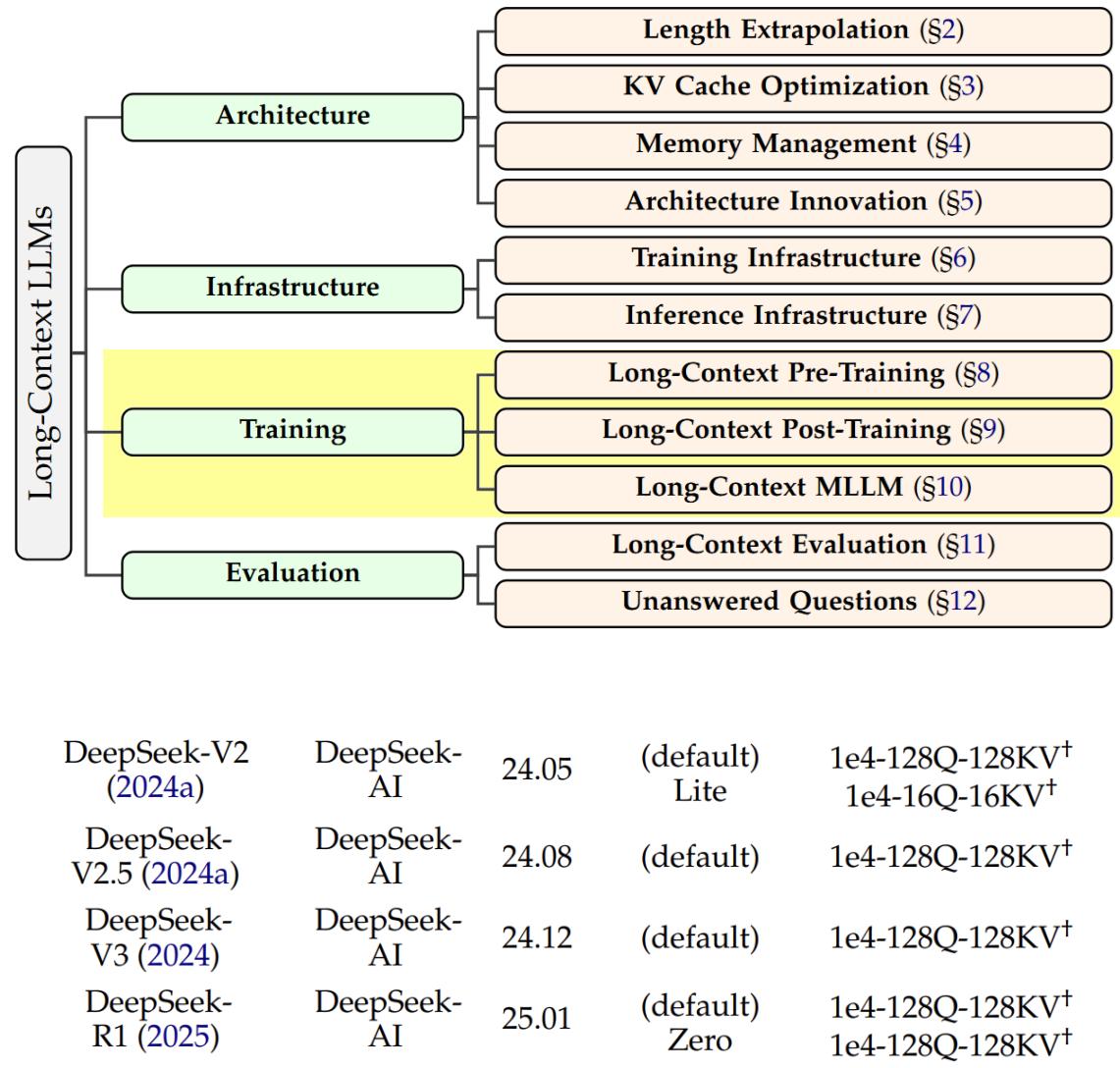
Github

Zeng, et al. Revisiting the Test-Time Scaling of o1-like Models: Do they Truly Possess Test-Time Scaling Capabilities? arXiv 2025

训练部分：纯文本训练

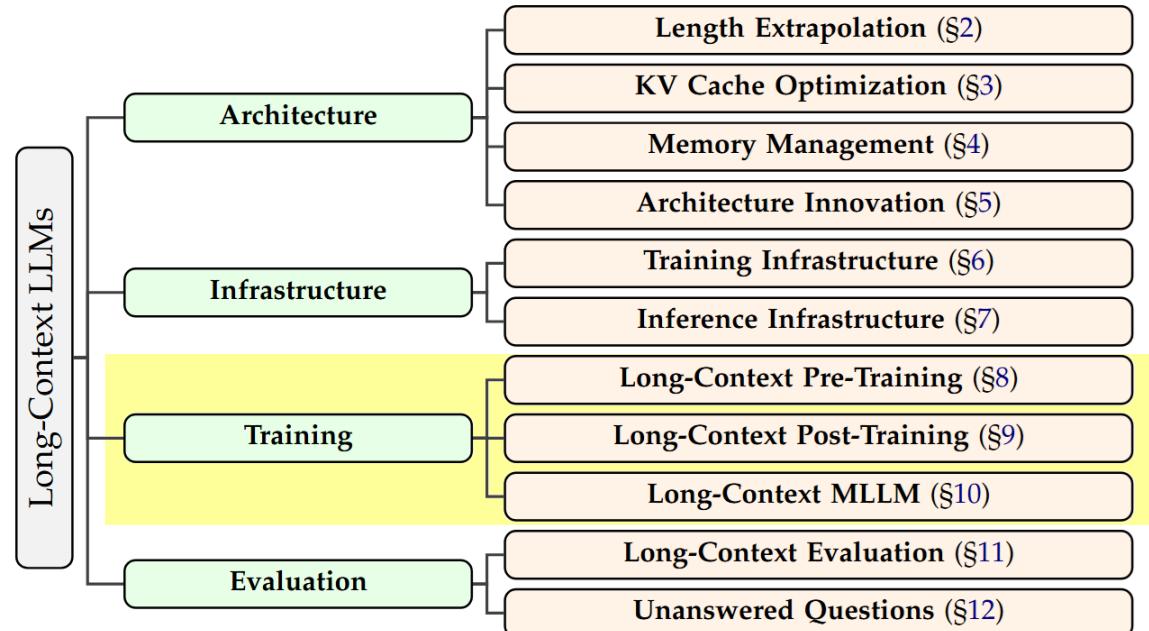
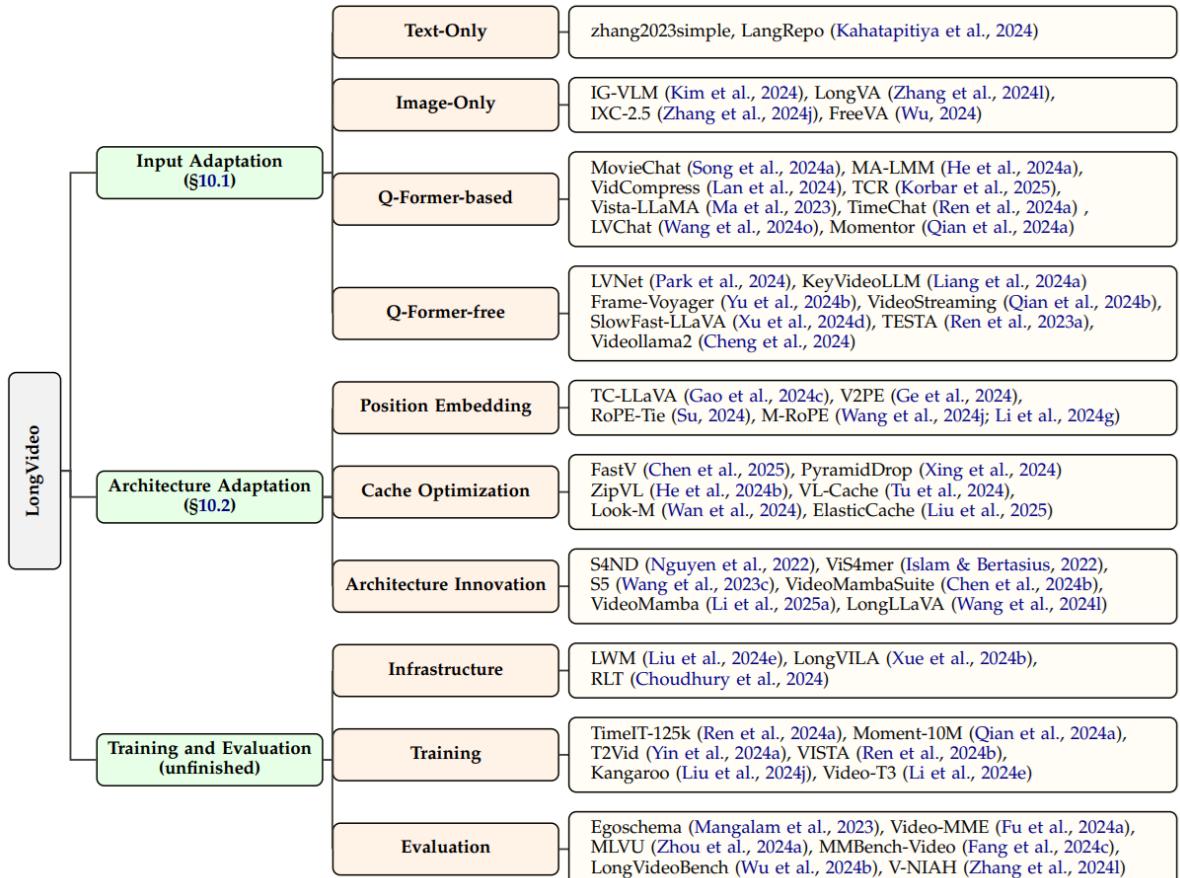
- ▶ 对无限长上下文的第二波冲刺！
- ▶ 主流大模型长文表现与架构细节的整理

Model	Organization	Time	Version	Architecture Detail (Base-Q-KV)	Context Length
LLaMA (2023a)	Meta	23.03	7B	1e4-32Q-32KV	2k
			13B	1e4-40Q-40KV	
			33B	1e4-52Q-52KV	
			65B	1e4-64Q-64KV	
			7B	1e4-32Q-32KV	
LLaMA2 (2023b)	Meta	23.07	13B	1e4-40Q-40KV	4k
			70B	1e4-64Q-8KV	
			8B	5e5-32Q-32KV	
LLaMA3 (2024a)	Meta	24.04	70B	5e5-64Q-8KV	8k
			8B	5e5-32Q-8KV	
			freq 1,4; factor 8	freq 1,4; factor 8	
LLaMA3.1° (2024)	Meta	24.07	70B	5e5-64Q-8KV	128k
			405B	freq 1,4; factor 8	
			1B	5e5-128Q-8KV	
			freq 1,4; factor 8	5e5-32Q-8KV	
			freq 1,4; factor 32	5e5-32Q-8KV	
LLaMA3.2° (2024b)	Meta	24.09	3B	freq 1,4; factor 32	128k
			11B	5e5-32Q-8KV	
			freq 1,4; factor 8	5e5-64Q-8KV	
LLaMA3.3° (2024)	Meta	24.12	70B	freq 1,4; factor 8	128k



训练部分：多模态训练

- ▶ 对无限长上下文的第二波冲刺！
- ▶ 相比于文本，**多模态更需要长上下文**



- ▶ 多模态长文：长视频大模型
- ▶ 已有研究的重点更多在于输入的适配：纯文本、纯图像、QFormer-based/free
- ▶ 基础架构的设计仍然缺少统一的共识

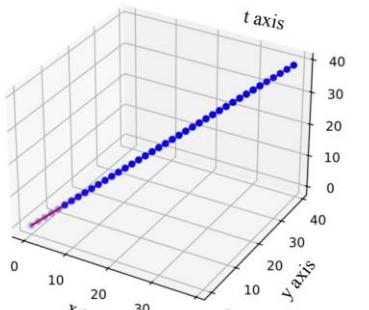
训练部分： VideoRoPE：长视频位置编码设计

系统分析视频位置编码设计要素，显著提升长视频建模效果

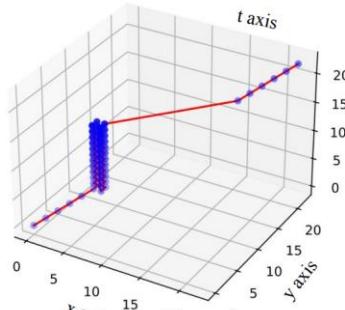
分析视频位置编码设计的四个要素：

3D表征、频率分配、空间对称、时序对齐

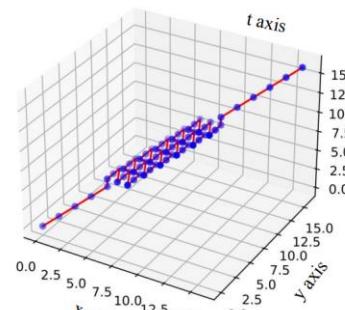
低频 时序长程依赖，高频 空间局部语义
在视频理解、检索、幻觉等任务上提升



(a) 3D visualization for Vanilla RoPE.



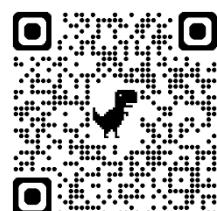
(b) 3D visualization for M-RoPE.



(c) 3D visualization for VideoRoPE.



arXiv



Github

	2D/3D Structure	Frequency Allocation	Spatial Symmetry	Temporal Index Scaling
Vanilla RoPE (Su et al., 2024)	✗	✗	✗	✗
TAD-RoPE (Gao et al., 2024)	✗	✗	✗	✓
RoPE-Tie (Su, 2024a)	✓	✗	✓	✗
M-RoPE (Wang et al., 2024b)	✓	✗	✗	✗
VideoRoPE (Ours)	✓	✓	✓	✓

Table 1. Comparison between different RoPE variants for Video Large Language Models (Video LLMs).

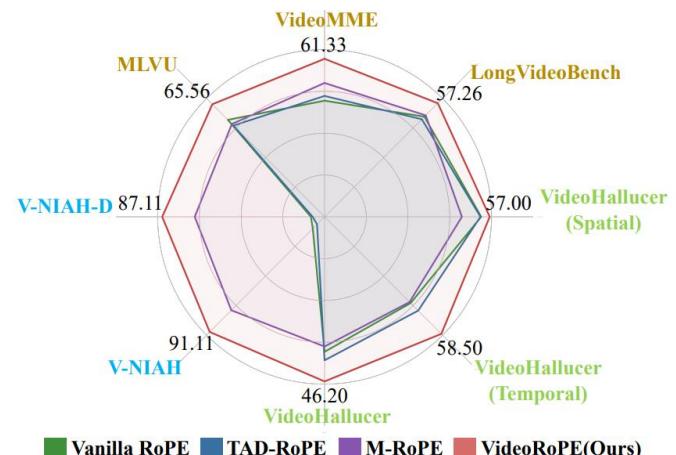
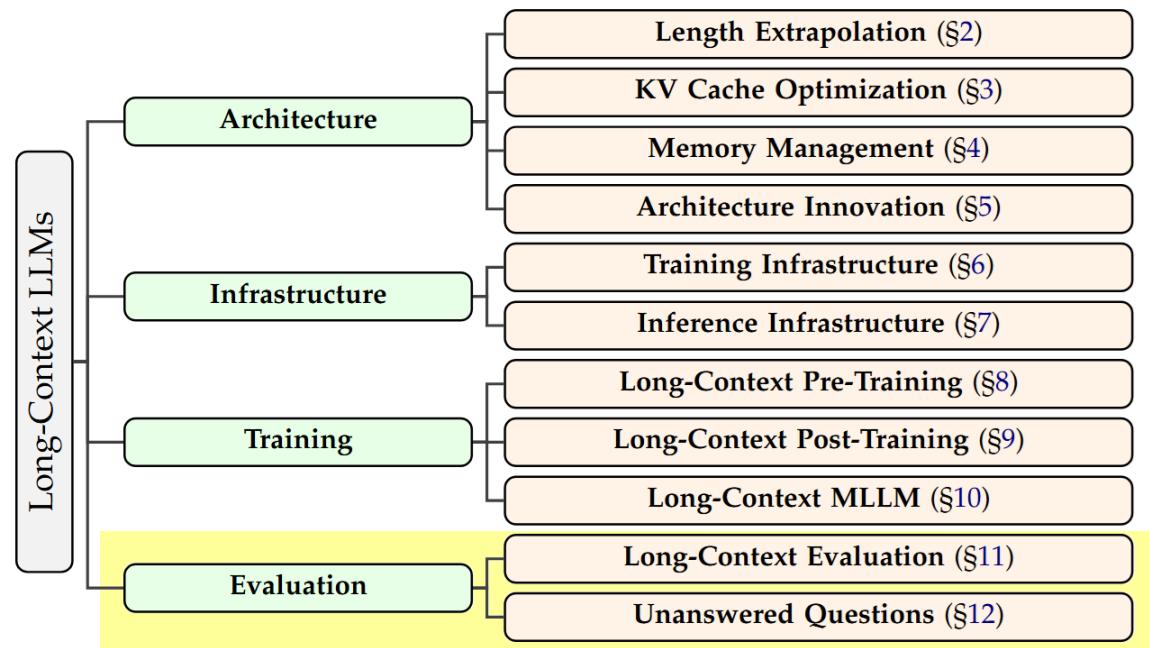


Figure 1. VideoRoPE outperforms RoPE variants on benchmarks.

Wei & Liu, et al. VideoRoPE: What Makes for Good Video Rotary Position Embedding? ICML 2025 under review

评测部分：长文评测

Name	Time	Benchmark Feature						
		Len.	Lang.	Flexible	Stable	D.C.	Align.	L.O.
Scroll (Shaham et al., 2022)	22.01	~8k	En	✗	✗	✗	✗	✗
ZeroScrolls (Shaham et al., 2023)	23.05	~8k	En	✗	✗	✗	✗	✗
LEval (An et al., 2023)	23.07	4k-60k	En	✗	✓	✗	✗	✗
LongBench (Bai et al., 2023b)	23.08	~10k	En, Zh	✗	✗	✗	✗	✗
BAMBOO (Dong et al., 2024f)	23.09	4k-16k	En	○	✓	✓	✓	✗
M4LE (Kwan et al., 2023)	23.10	1k-128k	En, Zh	○	✗	✗	✗	✗
LoOGLE (Li et al., 2023b)	23.11	~20k	En	○	✗	✓	✗	✗
Marathon (Zhang et al., 2023e)	23.12	~80k	En	✗	✓	✗	✗	✗
Needle-In-A-Haystack (Kamradt, 2023)	23.11	1k-128k	En	✓	✓	✗	✗	✗
InfiniteBench (Zhang et al., 2024q)	24.02	~200k	En, Zh	✗	✗	✗	✗	✓
LV-Eval (Yuan et al., 2024b)	24.02	16k-56k	En	✓	✓	✓	✗	✗
Multi-NIHA (Reid et al., 2024)	24.03	1k-1M	En	✓	✓	✗	✓	✗
CLongEval (Qiu et al., 2024)	24.03	1k-100k	Zh	○	✗	✗	✗	✗
LongICLBench (Li et al., 2024j)	24.04	2k-50k	En	✓	✓	✗	✗	✗
XL2Bench (Ni et al., 2024)	24.04	~200k	En, Zh	✗	✗	✓	✗	✗
RULER (Hsieh et al., 2024a)	24.04	4k-1M	En	✓	✓	✗	✗	✗
Ada-LEval (Wang et al., 2024a)	24.04	2k-128k	En	○	✓	✗	✗	✗
LoFT (Lee et al., 2024a)	24.06	32k-1M	En, Es, Fr, Hi, Zh	○	✓	✗	✗	✗
Loong (Wang et al., 2024i)	24.06	10k-250k	En, Zh	○	✓	✓	✗	✗
BABILong (Kuratov et al., 2024)	24.06	4k~10M	En	✓	✓	✓	✗	✗
LongIns (Gavin et al., 2024)	24.06	256-16k	En	✓	✓	✗	✓	✗
NeedleBench (Li et al., 2024f)	24.07	20k-1M	En, Zh	✓	✓	✗	✓	✗
HelloBench (Que et al., 2024)	24.09	~2k	En	✗	✓	✗	✓	✓
LongGenBench ₁ (Wu et al., 2024k)	24.09	~20k	En	○	✓	✗	✓	✓
LongGenBench ₂ (Liu et al., 2024n)	24.10	4k-128k	En	✓	✓	✗	✓	✓
HELMET (Yen et al., 2024b)	24.10	8k-128k	En	○	✓	✗	✗	✗
LongSafetyBench (Huang et al., 2024a)	24.11	~40k	En	✗	✓	✗	✓	✗
LIFBench (Wu et al., 2024i)	24.11	4k-128k	En	✓	✓	✗	✓	✗
LongBench v2 (Bai et al., 2024b)	24.12	32k-128k	En, Zh	○	✓	✗	✗	✗
LongProc (Ye et al., 2025a)	25.01	500 8k	En	○	✓	✗	✓	✓



- ▶ 长文评测的发展历程就像一首盛大的舞曲
- ▶ 长文模型推动着长文评测迈进
- ▶ 早期（22.12-23.11）以QA和摘要为主
- ▶ 任务比较真实，原生长文；长度不够灵活，评测指标不够稳定

评测部分：L-Eval：高质量长文评测基准

高质量的数据 + 稳定的指标 = 高质量的评测基准

长度外推后，最早的质量长文评测基准

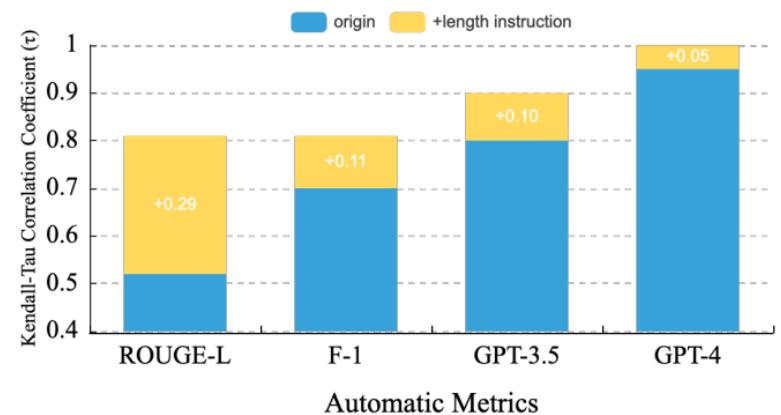
Dataset	Question-style	Domain	Avg len	Max len	#Instr	#Doc
<i>Closed - Ended Tasks</i>						
TOEFL	Multiple choice	English test	3,907	4,171	269	15
GSM(16-shot) [†]	Solving math problems	In-context examples	5,557	5,638	100	100
QuALITY [†]	Multiple choice	Gutenberg	7,169	8,560	202	15
Coursera*	Multiple choice	Advanced courses	9,075	17,185	172	15
TopicRet [†]	Retrieving topics	Conversation	12,506	15,916	150	50
SFiction*	True or False Questions	Scientific fictions	16,381	26,918	64	7
CodeU*	Deducing program outputs	Python Codebase	31,575	36,509	90	90
<i>Open - Ended Tasks</i>						
MultiDoc2Dial	Goal-oriented dialogues	Grounded documents	3,905	7888	136	20
Qasper	QA on papers	NLP papers	5,019	6,547	160	20
LongFQA*	QA on earning call	Finance	6,032	7824	52	6
NQ	QA from Google Search	Wikipedia	23,698	47,726	104	20
CUAD	Extracting key information	Law	30,966	68,625	130	20
NarrativeQA	QA on narratives	Gutenberg	62,335	210,541	182	20
Multi-News	Multi-doc Summarization	Multiple News articles	7,320	19,278	11	11
GovReport	Single-doc Summarization	Government reports	7,495	27,128	13	13
BigPatent	Single-doc Summarization	Lengthy patents	7,718	12,867	13	13
SummScreen	Transcripts Summarization	TV series transcripts	10,688	14,544	13	13
Openreview [†]	Paper writing & reviewing	Papers from Openreview	11,170	33,303	60	20
QMSum	Query-based summarization	Meeting transcripts	16,692	33,310	156	20
SPACE [†]	Aspect-based summarization	Reviews on Hotels	19,978	22,158	120	20

模型+人工校对并富集

高质量长文评测数据

基于**长度控制**的更反映

人类偏好的模型打分



arXiv



Github

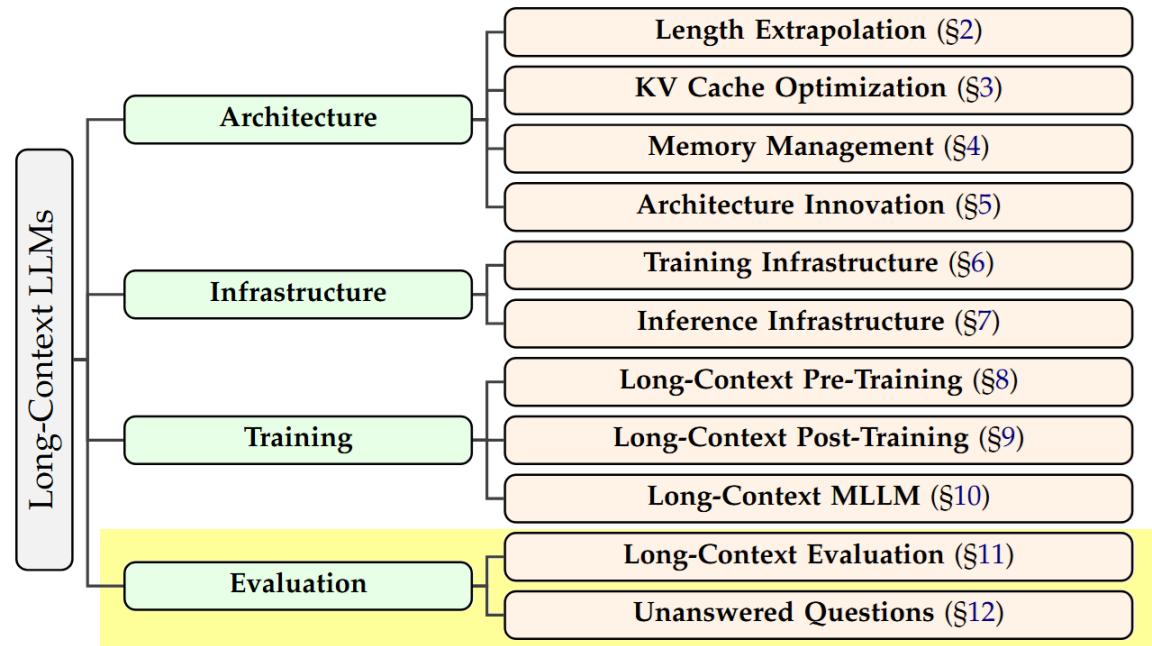


huggingface

An, et al. L-Eval: Instituting Standardized Evaluation for Long Context Language Models, **ACL 2024 Outstanding paper**

评测部分：长文评测

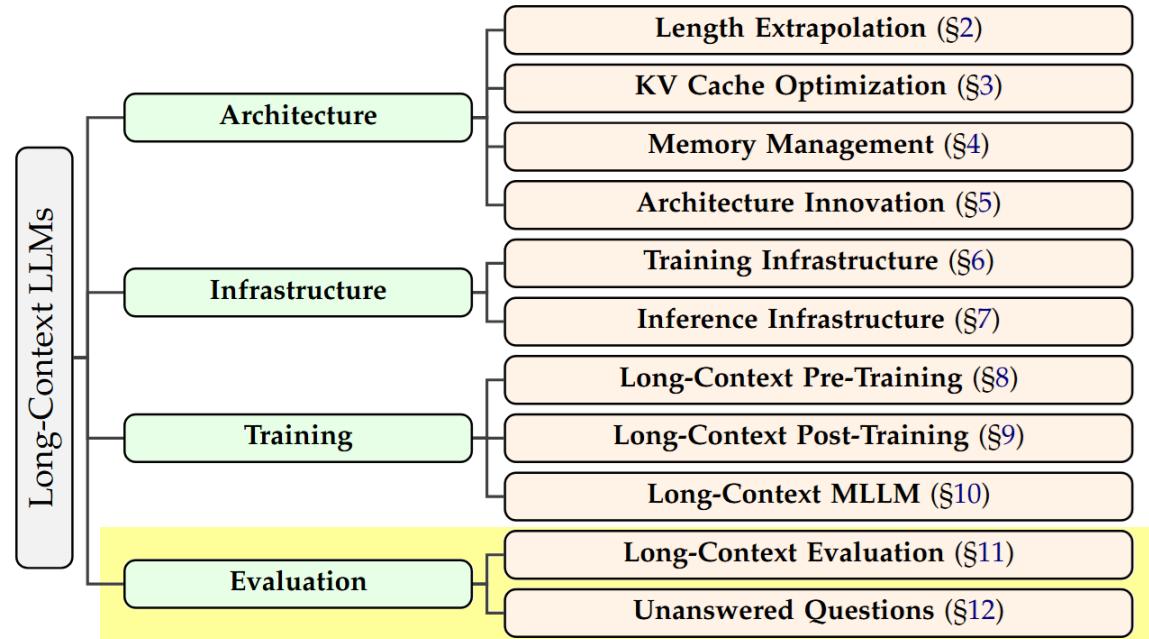
Name	Time	Benchmark Feature						
		Len.	Lang.	Flexible	Stable	D.C.	Align.	L.O.
Scroll (Shaham et al., 2022)	22.01	~8k	En	✗	✗	✗	✗	✗
ZeroScrolls (Shaham et al., 2023)	23.05	~8k	En	✗	✗	✗	✗	✗
LEval (An et al., 2023)	23.07	4k-60k	En	✗	✓	✗	✗	✗
LongBench (Bai et al., 2023b)	23.08	~10k	En, Zh	✗	✗	✗	✗	✗
BAMBOO (Dong et al., 2024f)	23.09	4k-16k	En	○	✓	✓	✓	✗
M4LE (Kwan et al., 2023)	23.10	1k-128k	En, Zh	○	✗	✗	✗	✗
LoOGLE (Li et al., 2023b)	23.11	~20k	En	○	✗	✓	✗	✗
Marathon (Zhang et al., 2023e)	23.12	~80k	En	✗	✓	✗	✗	✗
Needle-In-A-Haystack (Kamradt, 2023)	23.11	1k-128k	En	✓	✓	✗	✗	✗
InfiniteBench (Zhang et al., 2024q)	24.02	~200k	En, Zh	✗	✗	✗	✗	✓
LV-Eval (Yuan et al., 2024b)	24.02	16k-56k	En	✓	✓	✓	✗	✗
Multi-NIHA (Reid et al., 2024)	24.03	1k-1M	En	✓	✓	✗	✓	✗
CLongEval (Qiu et al., 2024)	24.03	1k-100k	Zh	○	✗	✗	✗	✗
LongICLBench (Li et al., 2024j)	24.04	2k-50k	En	✓	✓	✗	✗	✗
XL2Bench (Ni et al., 2024)	24.04	~200k	En, Zh	✗	✗	✓	✗	✗
RULER (Hsieh et al., 2024a)	24.04	4k-1M	En	✓	✓	✗	✗	✗
Ada-LEval (Wang et al., 2024a)	24.04	2k-128k	En	○	✓	✗	✗	✗
LoFT (Lee et al., 2024a)	24.06	32k-1M	En, Es, Fr, Hi, Zh	○	✓	✗	✗	✗
Loong (Wang et al., 2024i)	24.06	10k-250k	En, Zh	○	✓	✓	✗	✗
BABILong (Kuratov et al., 2024)	24.06	4k~10M	En	✓	✓	✓	✗	✗
LongIns (Gavin et al., 2024)	24.06	256-16k	En	✓	✓	✗	✓	✗
NeedleBench (Li et al., 2024f)	24.07	20k-1M	En, Zh	✓	✓	✗	✓	✗
HelloBench (Que et al., 2024)	24.09	~2k	En	✗	✓	✗	✓	✓
LongGenBench ₁ (Wu et al., 2024k)	24.09	~20k	En	○	✓	✗	✓	✓
LongGenBench ₂ (Liu et al., 2024n)	24.10	4k-128k	En	✓	✓	✗	✓	✓
HELMET (Yen et al., 2024b)	24.10	8k-128k	En	○	✓	✗	✗	✗
LongSafetyBench (Huang et al., 2024a)	24.11	~40k	En	✗	✓	✗	✓	✗
LIFBench (Wu et al., 2024i)	24.11	4k-128k	En	✓	✓	✗	✓	✗
LongBench v2 (Bai et al., 2024b)	24.12	32k-128k	En, Zh	○	✓	✗	✗	✗
LongProc (Ye et al., 2025a)	25.01	500 8k	En	○	✓	✗	✓	✓



► 23.11：评测的长度追不上模型的长度

评测部分：长文评测

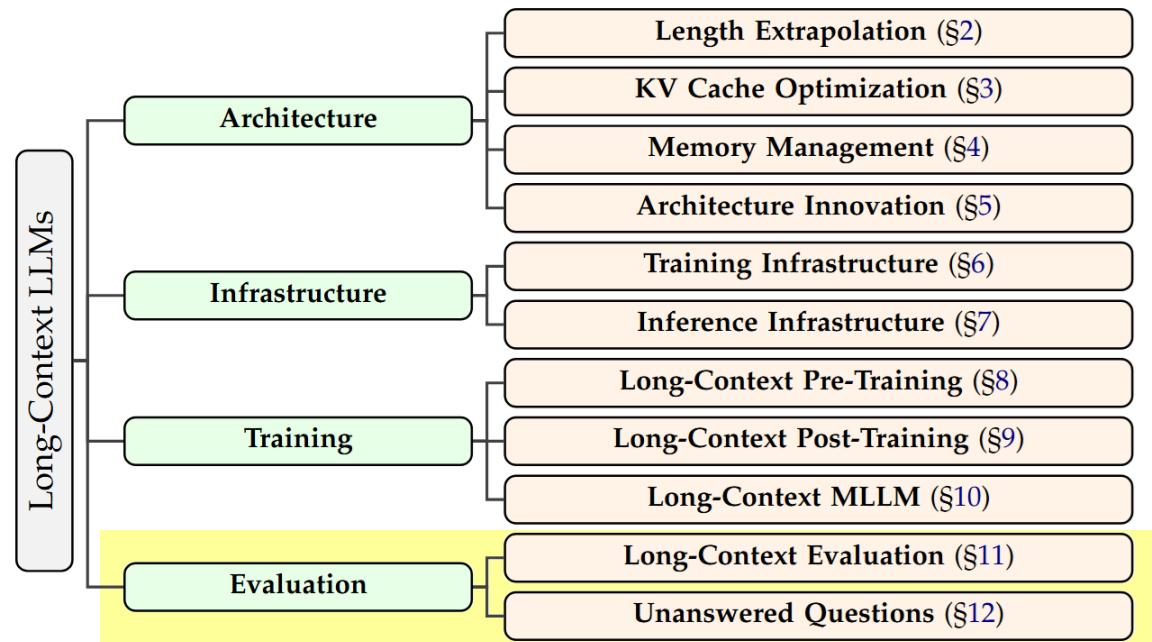
Name	Time	Benchmark Feature						
		Len.	Lang.	Flexible	Stable	D.C.	Align.	L.O.
Scroll (Shaham et al., 2022)	22.01	~8k	En	✗	✗	✗	✗	✗
ZeroScrolls (Shaham et al., 2023)	23.05	~8k	En	✗	✗	✗	✗	✗
LEval (An et al., 2023)	23.07	4k-60k	En	✗	✓	✗	✗	✗
LongBench (Bai et al., 2023b)	23.08	~10k	En, Zh	✗	✗	✗	✗	✗
BAMBOO (Dong et al., 2024f)	23.09	4k-16k	En	○	✓	✓	✓	✗
M4LE (Kwan et al., 2023)	23.10	1k-128k	En, Zh	○	✗	✗	✗	✗
LoogLE (Li et al., 2023b)	23.11	~20k	En	○	✗	✓	✗	✗
Marathon (Zhang et al., 2023e)	23.12	~80k	En	✗	✓	✗	✗	✗
Needle-In-A-Haystack (Kamradt, 2023)	23.11	1k-128k	En	✓	✓	✗	✗	✗
InfiniteBench (Zhang et al., 2024q)	24.02	~200k	En, Zh	✗	✗	✗	✗	✓
LV-Eval (Yuan et al., 2024b)	24.02	16k-56k	En	✓	✓	✓	✗	✗
Multi-NIHA (Reid et al., 2024)	24.03	1k-1M	En	✓	✓	✗	✓	✗
CLongEval (Qiu et al., 2024)	24.03	1k-100k	Zh	○	✗	✗	✗	✗
LongICLBench (Li et al., 2024j)	24.04	2k-50k	En	✓	✓	✗	✗	✗
XL2Bench (Ni et al., 2024)	24.04	~200k	En, Zh	✗	✗	✓	✗	✗
RULER (Hsieh et al., 2024a)	24.04	4k-1M	En	✓	✓	✗	✗	✗
Ada-LEval (Wang et al., 2024a)	24.04	2k-128k	En	○	✓	✗	✗	✗
LoFT (Lee et al., 2024a)	24.06	32k-1M	En, Es, Fr, Hi, Zh	○	✓	✗	✗	✗
Loong (Wang et al., 2024i)	24.06	10k-250k	En, Zh	○	✓	✓	✗	✗
BABILong (Kuratov et al., 2024)	24.06	4k~10M	En	✓	✓	✓	✗	✗
LongIns (Gavin et al., 2024)	24.06	256-16k	En	✓	✓	✗	✓	✗
NeedleBench (Li et al., 2024f)	24.07	20k-1M	En, Zh	✓	✓	✗	✓	✗
HelloBench (Que et al., 2024)	24.09	~2k	En	✗	✓	✗	✓	✓
LongGenBench ₁ (Wu et al., 2024k)	24.09	~20k	En	○	✓	✗	✓	✓
LongGenBench ₂ (Liu et al., 2024n)	24.10	4k-128k	En	✓	✓	✗	✓	✓
HELMET (Yen et al., 2024b)	24.10	8k-128k	En	○	✓	✗	✗	✗
LongSafetyBench (Huang et al., 2024a)	24.11	~40k	En	✗	✓	✗	✓	✗
LIFBench (Wu et al., 2024i)	24.11	4k-128k	En	✓	✓	✗	✓	✗
LongBench v2 (Bai et al., 2024b)	24.12	32k-128k	En, Zh	○	✓	✗	✗	✗
LongProc (Ye et al., 2025a)	25.01	500 8k	En	○	✓	✗	✓	✓



- 23.11: 评测的长度追不上模型的长度
- 长文评测推动着长文模型迈进
- 后续 (23.11-24.09) 以检索和聚合为主
- 长度灵活，可以无限长，评测更加稳定；合成任务不够真实，典型代表：NIAH

评测部分：长文评测

Name	Time	Benchmark Feature						
		Len.	Lang.	Flexible	Stable	D.C.	Align.	L.O.
Scroll (Shaham et al., 2022)	22.01	~8k	En	✗	✗	✗	✗	✗
ZeroScrolls (Shaham et al., 2023)	23.05	~8k	En	✗	✗	✗	✗	✗
LEval (An et al., 2023)	23.07	4k-60k	En	✗	✓	✗	✗	✗
LongBench (Bai et al., 2023b)	23.08	~10k	En, Zh	✗	✗	✗	✗	✗
BAMBOO (Dong et al., 2024f)	23.09	4k-16k	En	○	✓	✓	✓	✗
M4LE (Kwan et al., 2023)	23.10	1k-128k	En, Zh	○	✗	✗	✗	✗
LoogLE (Li et al., 2023b)	23.11	~20k	En	○	✗	✓	✗	✗
Marathon (Zhang et al., 2023e)	23.12	~80k	En	✗	✓	✗	✗	✗
Needle-In-A-Haystack (Kamradt, 2023)	23.11	1k-128k	En	✓	✓	✗	✗	✗
InfiniteBench (Zhang et al., 2024q)	24.02	~200k	En, Zh	✗	✗	✗	✗	✓
LV-Eval (Yuan et al., 2024b)	24.02	16k-56k	En	✓	✓	✓	✗	✗
Multi-NIHA (Reid et al., 2024)	24.03	1k-1M	En	✓	✓	✗	✓	✗
CLongEval (Qiu et al., 2024)	24.03	1k-100k	Zh	○	✗	✗	✗	✗
LongICLBench (Li et al., 2024j)	24.04	2k-50k	En	✓	✓	✗	✗	✗
XL2Bench (Ni et al., 2024)	24.04	~200k	En, Zh	✗	✗	✓	✗	✗
RULER (Hsieh et al., 2024a)	24.04	4k-1M	En	✓	✓	✗	✗	✗
Ada-LEval (Wang et al., 2024a)	24.04	2k-128k	En	○	✓	✗	✗	✗
LoFT (Lee et al., 2024a)	24.06	32k-1M	En, Es, Fr, Hi, Zh	○	✓	✗	✗	✗
Loong (Wang et al., 2024i)	24.06	10k-250k	En, Zh	○	✓	✓	✗	✗
BABILong (Kuratov et al., 2024)	24.06	4k~10M	En	✓	✓	✓	✗	✗
LongIns (Gavin et al., 2024)	24.06	256-16k	En	✓	✓	✗	✓	✗
NeedleBench (Li et al., 2024f)	24.07	20k-1M	En, Zh	✓	✓	✗	✓	✗
HelloBench (Que et al., 2024)	24.09	~2k	En	✗	✓	✗	✓	✓
LongGenBench ₁ (Wu et al., 2024k)	24.09	~20k	En	○	✓	✗	✓	✓
LongGenBench ₂ (Liu et al., 2024n)	24.10	4k-128k	En	✓	✓	✗	✓	✓
HELMET (Yen et al., 2024b)	24.10	8k-128k	En	○	✓	✗	✗	✗
LongSafetyBench (Huang et al., 2024a)	24.11	~40k	En	✗	✓	✗	✓	✗
LIFBench (Wu et al., 2024i)	24.11	4k-128k	En	✓	✓	✗	✓	✗
LongBench v2 (Bai et al., 2024b)	24.12	32k-128k	En, Zh	○	✓	✗	✗	✗
LongProc (Ye et al., 2025a)	25.01	500 8k	En	○	✓	✗	✓	✓



- ▶ 24.09: OpenAI-o1: 从长上文迈向长下文
- ▶ 长文评测和长文模型共同进步
- ▶ 近期 (24.09-25.03) 以推理和输出为主
- ▶ 在任务真实实用的前提下，维持指标可靠、长度可伸缩、多任务、多domain 等

评测部分：DetectiveQA：侦探小说长文推理评测

侦探推理小说 更加真实的任务，推理过程评价 更有挑战的指标

Question:
Which of the following is the reason for the disappearance of Sainsbury Seale?

Options:
A. left voluntarily. B. met an untimely end
C. eloped with someone. D. Sudden memory loss.

Answer:
B

Reasoning Steps:
Step #1: Ms Sainsbury-Seal did not take her luggage with her when she disappeared.
Step #2: This does not appear to be a voluntary departure.
Step #3: Ms Seale had a dinner date with a friend to play solitaire.
Step #4: Normally at the appointed time she would have been back at the hotel.
Step #5: Therefore, based on the above evidences, it is surmised that it was Sainsbury Seale who met an untimely end.

Evidence Position:
740, -1, 734, -1, -1

Answer Position:
1202



利用侦探小说的**原生长本**和
复杂推理链条构造真实任务

通过**推理线索覆盖率**
提供**长文推理的过程**

**评价，更符合
偏好，能反映
长文模型不足**

Models	Question+Context		
	Answer	Reasoning	G.M.
GPT-4-1106-preview-128k	73.99	27.43	45.05
OpenAI-O1-mini-128k	60.83	23.80	38.05
KimiChat-200k	64.13	27.79	42.21
Claude3-Opus-200k	81.95	37.33	55.30
LLaMA3.1-8B-Instruct-128k	28.17	21.15	24.41
GLM3-6B-128k	40.58	22.08	33.63
GLM4-9B-chat-1M	59.00	24.07	37.68
Qwen2.5-7B-Instruct-128k	61.75	21.16	36.15
InternLM2-7B-chat-200k	57.95	23.94	37.24
InternLM2.5-7B-chat-1M	60.92	22.45	36.98



arXiv

Xu, Ye, Liu and Liu, et al. DetectiveQA: Evaluating Long-Context Reasoning on Detective Novels, ICLR 2025 workshop

评测部分：未做回答的问题

- ▶ 在评测中发现激动人心的应用
 - ▶ 早期的长文模型：阅读长文本
 - ▶ Gemini：长上下文学习翻译小语种
 - ▶ **范式的转变：长上文到长下文**
 - ▶ DeepSeek：长推理的aha moment
 - ▶ RWKV：长推理解数独，性能与效率

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both ...

$$(\sqrt{a - \sqrt{a+x}})^2 = x^2 \implies a - \sqrt{a+x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

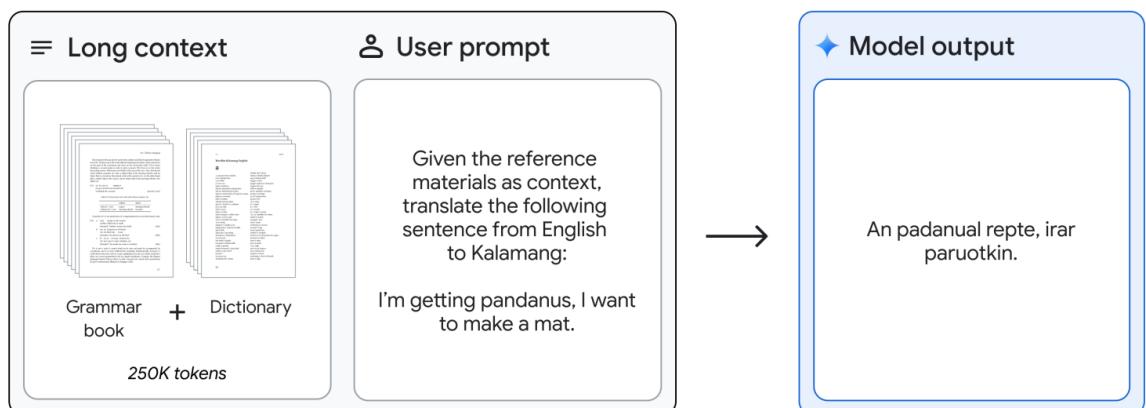
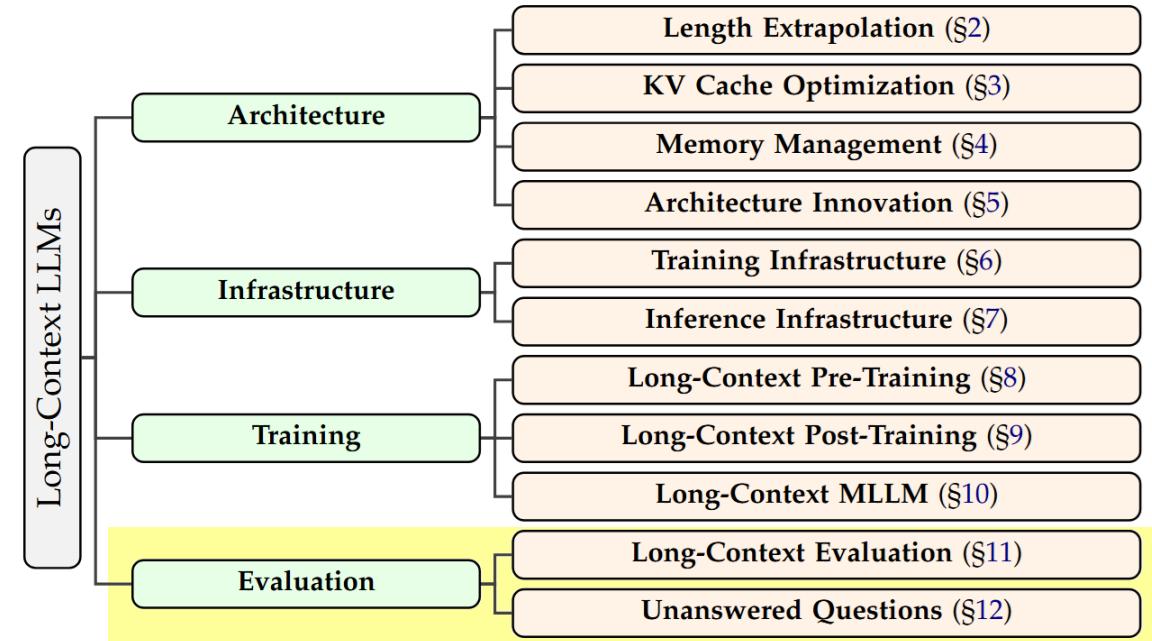
$$\sqrt{a - \sqrt{a+x}} = x$$

First, let's square both sides:

$$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

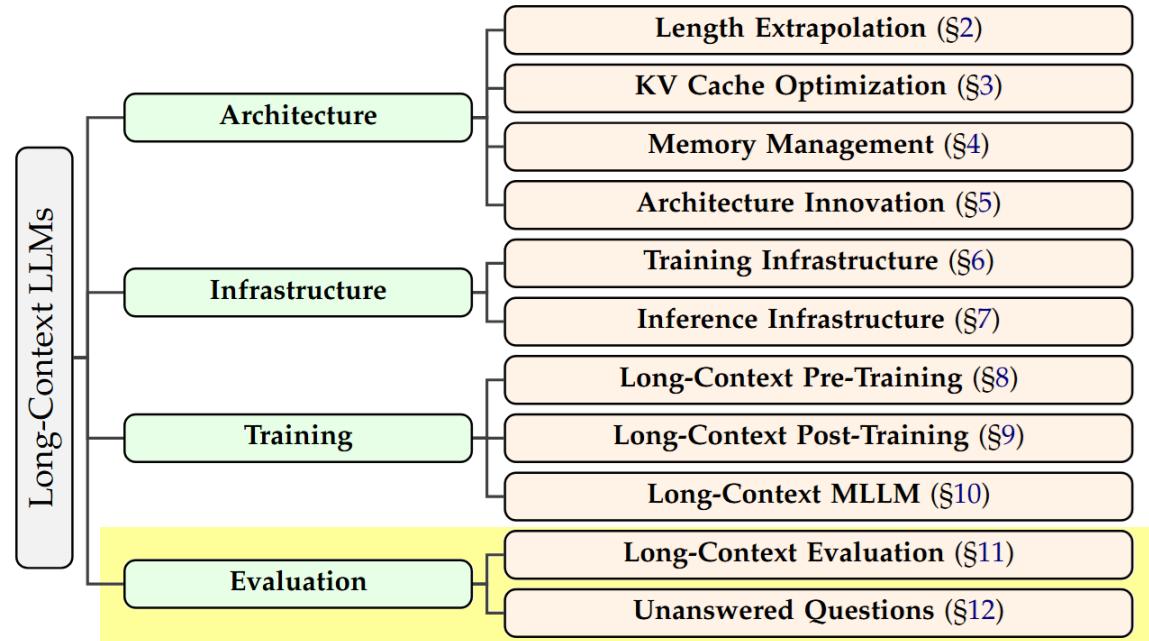
...



评测部分：未做回答的问题

▶ 狂欢终了，又是新一轮的沉思与尝试

- ▶ 位置偏置
- ▶ RoPE设计
- ▶ 长文PPL问题
- ▶ 长文 v.s. RAG
- ▶ 长文新架构
- ▶ 端侧长文
- ▶ 长短文混训
- ▶ 长文语料
- ▶ 长文推理 TTS
- ▶ 终身学习 TTT: **长文的星辰大海等待我们**
- ▶ **对无限长文的渴望与接受长文有限的需求**



▶ 终极的拷问

- ▶ 既然无限长上下文还无法达到，那百万 token上下文、长文研究的意义何在
- ▶ 正是这段研究让我们对LLM有了更深入的认知，让我们面对长文更加智慧

总结

长上下文大语言模型如是说

- ▶ 标题: *Thus Spake Long-Context Large Language Model*
- ▶ 灵感: *Thus Spake Zarathustra* by Richard Strauss
- ▶ 模型扩长上下文的过程就是人突破自身局限的过程!

SII-OpenMOSS & InternLM

Thus Spake Long-Context Large Language Model

Xiaoran Liu^{1,2,4*}, Ruixiao Li^{2,4*}, Mianqiu Huang^{2,*}, Zhigeng Liu^{2,*}, Yuerong Song^{2,4*}, Qipeng Guo^{1,4}, Siyang He², Qiqi Wang², Linlin Li³, Qun Liu³, Yaqian Zhou², Xuanjing Huang², Xipeng Qiu^{1,2,4†}

¹Shanghai AI Lab, ²School of Computer Science Fudan University,

³Huawei Noah's Ark Lab, ⁴Shanghai Innovation Institute

xrliu24@m.fudan.edu.cn, guoqipeng@pjlab.org.cn, xpqiu@fudan.edu.cn

- ▶ 当文字不够表达我的热情，我想用音乐来传递它



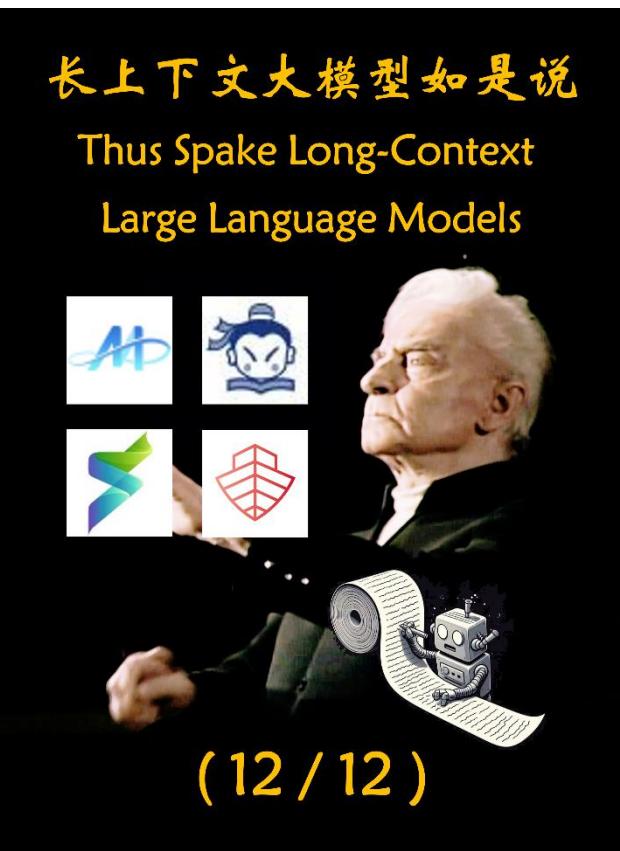
小红书



BiliBili



arXiv



Github



huggingface

同期工作

▶ 长上下文语言模型的全面综述

- ▶ 人类处理长文本的历史就是一部效率进化史
- ▶ 问题一：获得高效有效的模型
 - ▶ 训练数据、模型架构、工作流
- ▶ 问题二：实现高效的训练与部署
 - ▶ 训练框架、推理框架
- ▶ 问题三：全面的评测与分析
 - ▶ 长文评测、问题分析、应用、发展讨论



arXiv



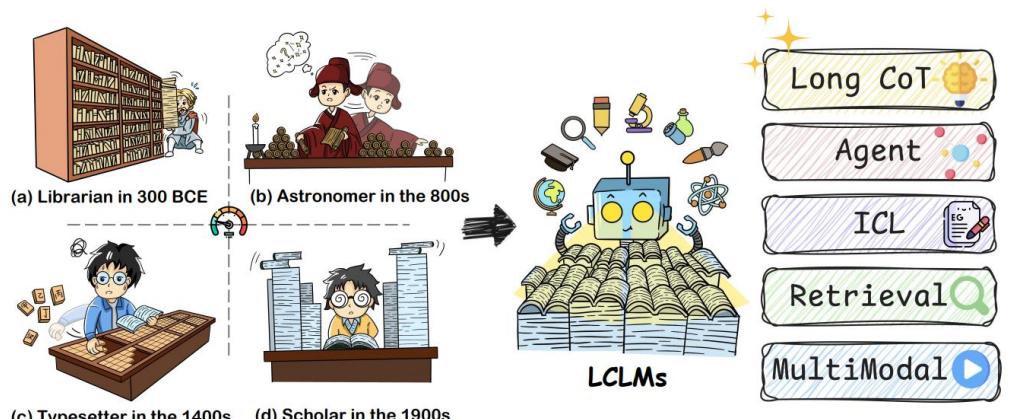
Github



A Comprehensive Survey on Long Context Language Modeling

Jiaheng Liu^{*,†}, Dawei Zhu^{*,†}, Zhiqi Bai^{*}, Yancheng He^{*}, Huanxuan Liao^{*}, Haoran Que^{*}, Zekun Wang^{*}, Chenchen Zhang^{*}, Ge Zhang^{*}, Jiebin Zhang^{*}, Yuanxing Zhang^{*}, Zhuo Chen, Hangyu Guo, Shilong Li, Ziqiang Liu, Yong Shan, Yifan Song, Jiayi Tian, Wenhao Wu, Zhejian Zhou, Ruijie Zhu, Junlan Feng, Yang Gao, Shizhu He, Zhoujun Li, Tianyu Liu, Fanyu Meng, Wenbo Su, Yingshui Tan, Zili Wang, Jian Yang, Wei Ye, Bo Zheng, Wangchunshu Zhou, Wenhao Huang[†], Sujian Li[†], Zhaoxiang Zhang[†]

NJU, PKU, CASIA, Alibaba, ByteDance, Tencent, Kuaishou, M-A-P



News

- [2025.03.13] 🎉 🎉 🎉 We have a good communication with the authors of [concurrent work](#), and will promote work of both parties in the future.

总结

▶ 架构部分

- ▶ 位置编码与外推；高效缓存；记忆管理；高效注意力、非注意力
- ▶ 我们的工作：ScalingRoPE ICLR24; ReAttention ICLR25; EM2 EMNLP24

▶ 框架部分

- ▶ 三个角度：计算优化、存储优化、分布式；两个场景：训练、推理
- ▶ 我们的工作：IMDC EMNLP24; ParallelTokenizer; CoLLiE EMNLP23

▶ 训练部分

- ▶ 长文预训练、长文后训练；长文提升/激发大模型能力；多模态长文（架构 训练 评测）
- ▶ 我们的工作：LongWanjuan EMNLP24; LongSafety ACL25'; Revisit TTS; VideoRoPE ICML25,

▶ 评测部分

- ▶ 长文评测：问答、检索、推理；我们的工作：L-Eval ACL24 Outstanding; DetectiveQA ICLR25
- ▶ 未做回答的问题，总结；同期工作



arXiv



Github



BiliBili



小红书



谢谢大家!