

通过 Kimi，看长文本的实现

2024 年 3 月 26 日 | 中国内地

专题研究

计算机 增持 (维持)

Kimi 上下文长度 10 倍增长，引领国内大模型长上下文迭代新方向

大模型的长上下文支持能力已经成为重要的迭代方向。海外相对超前，Anthropic Claude 3 模型标配 200K 上下文，并可向特定客户提供 1M 长度；Google Gemini 1.5 Pro 标配支持 1M 上下文长度，内部已实现 10M。国产大模型初创公司中，月之暗面的 Kimi 智能助手在 23 年 10 月即实现了 20 万字上下文，并在 24 年 3 月进一步迭代成为 200 万字。同月，阿里通义千问宣布文档解析功能支持 1000 万字；百度文心一言将在 4 月的更新中支持 200 万字以上的长文本能力；360 智脑开始内测 500 万字长文本处理功能。长上下文已成为全球大模型迭代重要方向，关注其他国产模型厂商进展。

大模型长上下文，主要通过优化 Transformer 架构实现

目前，全球大模型仍然以 Transformer 解码器为主要架构基础。在此基础上，可以通过改进解码器架构来实现长上下文，主要改进方法包括：1) 高效的注意力机制：降低计算成本，在训练时实现更长的序列长度，相应的推理时序列长度也就更长；2) 实现长期记忆：设计显式记忆机制，以解决上下文记忆的局限性。3) 改进位置编码：对现有的位置编码进行改进，实现上下文外推。4) 对上下文进行处理：用额外的上下文预/后处理，确保每次调用中输入给 LLM 的输入始终满足最大长度要求。

国内大模型厂商可能采取了多种路线混合优化方法实现长上下文

长上下文作为核心技术，各厂商选择不公开。以月之暗面为例，其创始人杨植麟主要的学术论文 Transformer-XL 和 XL-Net，均探讨了长上下文的实现方法，且前者属于长期记忆力的优化，后者属于特殊目标函数的优化。百度的 ERNIE-Doc 则同时采用了长期记忆力和特殊目标函数的优化方法。阿里 Qwen-7B 则使用了优化的位置编码算法 extended RoPE。所以我们推测，国内模型厂商之所以能够在短期内实践出长上下文方法，或是在原有积累的基础上进行了算法迭代，采取多方法的混合优化，实现快速超车。

长上下文的通用性将解决多类场景需求，带来应用突破机会

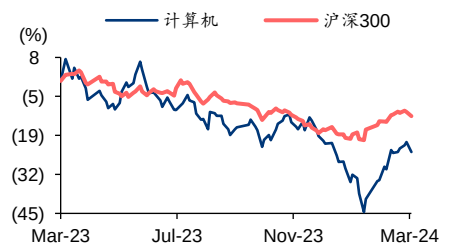
具有长上下文的大模型通用性更强，用户将特定领域的知识通过上下文的方式输入到模型中，模型即可以通过上下文学习掌握相应内容，一定程度上代替模型的微调。此外，长上下文模型能适应虚拟角色的个性化信息记忆、开发者的长 prompt 输入、AI Agent 的多轮调用需求，以及金融、法律等垂直客户长文档输入需求等多种场景，有望为 AI+ 应用带来新的突破机会。

关注大模型长文本潜在受益产业链

长文本应用场景：1) 文本工具：金山办公、福昕软件；2) 法律文案：华宇软件、通达海；3) 业务流程：泛微网络、致远互联；4) 其他文本：汉仪股份、汉王科技。专业领域+多任务+多模态场景：1) 金融领域：同花顺、恒生电子；2) 医疗领域：嘉和美康；3) 电商领域：光云科技。AI 算力：浪潮信息、神州数码、海光信息。

风险提示：宏观经济波动，技术进步不及预期。本研报中涉及到未上市公司或未覆盖个股内容，均系对其客观公开信息的整理，并不代表本研究团队对该公司、该股票的推荐或覆盖。

行业走势图



资料来源：Wind，华泰研究

重点推荐

股票名称	股票代码	目标价 (当地币种)	投资评级
金山办公	688111 CH	417.81	买入
福昕软件	688095 CH	86.72	买入
泛微网络	603039 CH	74.12	买入
汉王科技	002362 CH	25.74	买入
同花顺	300033 CH	169.06	买入
恒生电子	600570 CH	35.19	买入

资料来源：华泰研究预测

正文目录

- 长上下文已经成为当前阶段大模型的重要迭代趋势.....3
- Kimi 更新 10 倍长上下文，为国内大模型演进寻到差异化新方向.....4
- 实现 Transformer 长上下文，难点和限制在哪？6
 - 准备知识#1：Transformer 解码器架构仍是模型主流6
 - 准备知识#2：Transformer 架构的核心是注意力机制6
 - 准备知识#3：Transformer 架构需要 KV cache 来存储 KV 信息7
 - 准备知识#4：Transformer 架构需要位置编码7
 - 计算复杂度、内存和长度限制是长文本实现的核心矛盾点8
- 通过改进 Transformer 架构实现长上下文9
- 大模型产品长上下文实现方法预测12
 - 长上下文的通用性将解决多类场景需求13
 - 关注大模型长文本潜在受益产业链.....13
 - 风险提示.....14

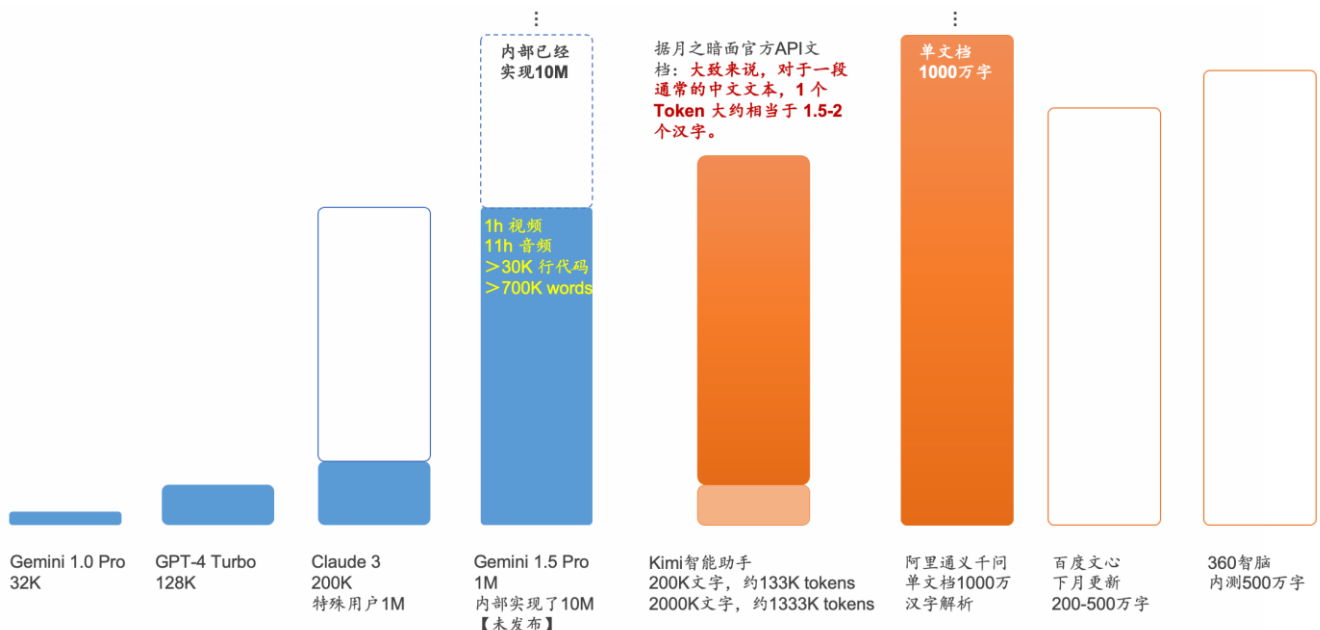
长上下文已经成为当前阶段大模型的重要迭代趋势

全球视角看，长上下文能力已经成为大模型重要的迭代趋势。我们认为，海外大模型发展相对超前，龙头公司在长上下文上布局略早于国内公司。Anthropic 旗下 Claude 一直以长文本能力著称。2023 年 11 月，Anthropic 发布 Claude 2.1 版本，将上下文支持能力从 100K 扩展到 200K tokens。24 年 3 月，Claude 3 发布，延续标配了 200K 上下文，并且可以向特定用户提供长达 1M token 的版本。Google Gemini 模型同样开始发力上下文，在 24 年 2 月发布 Gemini 1.5 Pro 时，将支持的上下文长度从 1.0 版本的 32K 大幅提升到 1M token，并宣称内部已经实现了 10M 的上下文，一举超越 Anthropic 成为闭源模型厂商中上下文长度最长的产品。

国内模型厂商迅速追赶，逐步补齐模型长文本能力。国内模型厂商中，较早实现优秀长文本效果的是杨植麟的初创公司月之暗面(Moonshot AI)，其 Kimi 智能助手(原名 Kimi Chat) 在 23 年 10 月发布时即支持 20 万汉字的长文本，长文本能力为当时国内模型 Top 1。24 年 3 月，Kimi 智能助手发布更新，将 20 万上下文扩展到 200 万上下文，并发布邀测。同月，阿里通义千问宣布推出文档解析功能，能够处理超万页的极长资料，换算成中文篇幅约 1000 万字。随后，百度文心一言也宣布将在 4 月的更新中支持 200 万字以上的长文本能力；360 官方也宣布 360 智脑开始内测 500 万字长文本处理功能，即将入驻 360AI 浏览器。此外，大模型初创公司阶跃星辰也发布 Step-1 和 1V 模型，支持 200K 上下文，且万亿参数 MoE 模型 Step 2 也已加入预览版申请。

我们认为，之所以长上下文会在当下成为趋势，主要原因包括，1) 阶段性需求：ChatGPT 和 GPT-4 问世已经超过 1 年，在基于 Transformer 解码器架构没有重大革新的情况下，模型的推理能力（GPT-4 能力）、成本控制（GPT-4 Turbo 的降价）、多模态能力（GPT-4V 等）、智能体能力（GPTs 等）已经取得阶段性成果，而上下文支持能力尚未被显著开发。2) 场景需求：尤其是对于虚拟陪伴类 AI 产品（如 Character.ai），用户希望在交互过程中，模型能够记忆长期的用户信息，需要依赖模型的长下文能力。以及对于逐渐丰富的大模型垂类场景，如金融分析、法律辅助、个性化教育等，需要模型分析较长的文档。3) AGI 的需求：更远期的看，长下文能够很好的解决模型在执行下游任务时需要做 fine-tune（微调）的问题。只需要将知识通过上下文输入，即可实现上下文学习，这是更加通用的方法，更符合 AGI 的定义。

图表1：全球主流模型厂商的长上下文布局（线框代表暂未落地，实框代表已经落地）

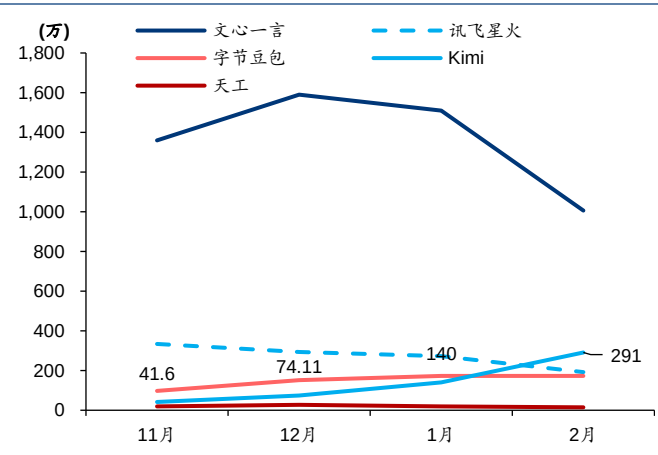


资料来源：各公司官网、华泰研究

Kimi 更新 10 倍长上下文，为国内大模型演进寻到差异化新方向

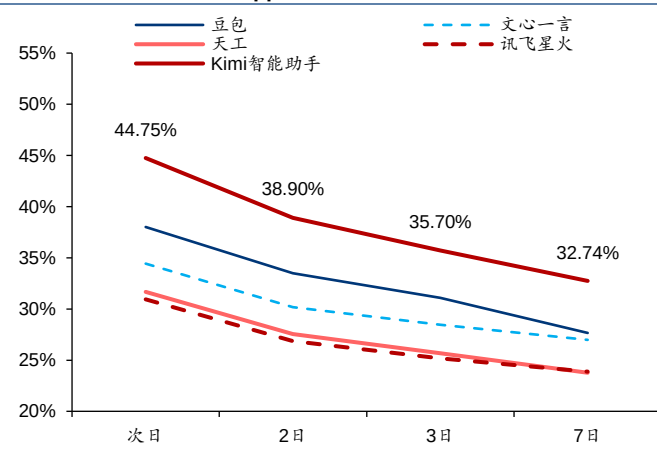
Kimi 智能助手访问量显著增长，App 留存率高于同类大模型产品。Kimi 智能助手是国内大模型初创公司月之暗面旗下主要产品。据 SimilarWeb 数据，Kimi 智能助手网页版的访问量从去年 11 月有统计数据开始（Kimi 于 23 年 10 月推出），即呈现一路上升的趋势。2 月环比提升超 100%。而同期同类国内主要大模型产品访问情况呈持平或下降状态。移动端看，据 QuestMobile 数据，Kimi 智能助手的 7 日用户留存率为 32.74%，高于同类国内大模型产品。体现了用户侧对 Kimi 产品的认可。

图表2：国内主流大模型访问量情况



资料来源：SimilarWeb、华泰研究

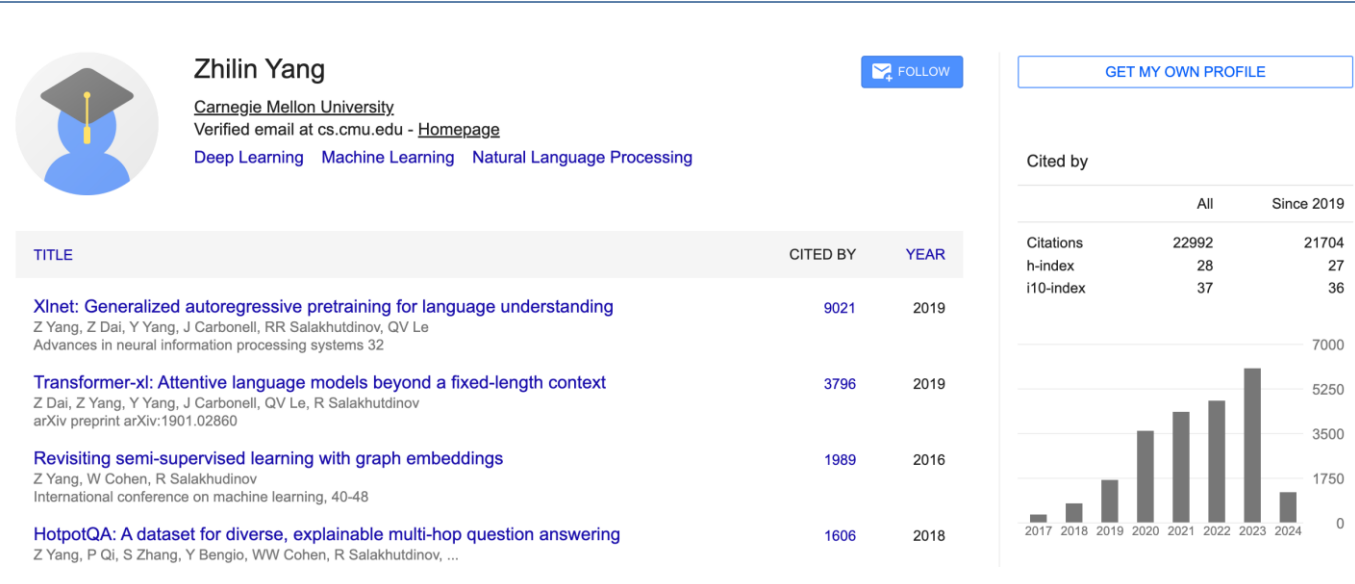
图表3：国内主流大模型 App 留存率情况



资料来源：QuestMobile、华泰研究

月之暗面创始人杨植麟博士深耕大模型领域，长上下文实现是其研究重点。杨植麟在国内师从清华教授、IEEE Fellow 唐杰教授（智谱创始人之一）。后前往卡内基梅隆大学（CMU）语言技术研究所（LTI），跟随苹果公司 AI 负责人 Ruslan Salakhutdinov 和 Google AI 智能首席科学家 William W. Cohen 攻读博士学位。毕业后曾在 Facebook AI Research 和 Google Brain 工作，发明的算法曾在 30 多项 AI 标准任务取得 SOTA。其代表性学术著作作为 Transformer-XL 与 XLNet 均探讨了 Transformer 长上下文的实现方法，Google Scholar 引用高达近 3800 和 9000 次，论文总引用数超 20000 次。学术论文在华人学者引用排名中位居前 10，在 40 岁以下排名第一，学术功底深厚。公司创始团队核心成员参与了 Google Gemini、Google Bard、盘古 NLP、悟道等多个大模型的研发，多项核心技术被 Google PaLM、Meta LLaMa、Stable Diffusion 等主流产品采用。

图表4：杨植麟博士学术代表著作及引用情况



资料来源：Google Scholar、华泰研究

闭源+2C 是月之暗面的核心标签，长上下文是产品核心技术，多模态是下一步迭代重点。据月之暗面信息，公司坚定 2C 方向，加速模型迭代，打造 super-app；同时产品以闭源为主，通过闭源形成产品差异化，从而形成数据的虹吸效应，进一步提高迭代效率、优化模型技术。23 年 10 月 9 日推出全球首个支持输入 20 万汉字的智能助手产品 Kimi Chat，底座模型为千亿参数级大模型 moonshot。当时支持的上下文长度是海外 Claude 2-100k（约 8 万字）的 2.5 倍，GPT-4-32k（约 2.5 万字）的 8 倍。杨植麟认为，上下文长度就是大模型的“内存”，它是决定大模型应用最关键的两个因素（参数量和上下文）之一。多模态产品也在跟进中，预计 24 年推出。

坚持无损长程注意力机制，Kimi 上下文长度扩展 10 倍，引领国内大模型下一步迭代方向。3 月 18 日，月之暗面官方宣布 Kimi 智能助手在长上下文窗口技术上再次取得突破，无损上下文长度提升了一个数量级到 200 万字，官网可约内测。官方指出，在技术路径上没有采用常规的渐进式提升路线，技术难度指数级增加。从模型预训练到对齐、推理环节均进行了原生的重新设计和开发，不走“滑动窗口”“降采样”等技术捷径，坚持无损的长程注意力机制，攻克了网络结构、算力、存储、带宽等很多底层技术难点。更长的上下文进一步拓宽模型使用场景，例如完整代码库的分析理解、自主帮人类完成多步骤复杂任务的智能体 Agent、不会遗忘关键信息的终身助理、真正统一架构的多模态模型等。此外，更新后的 Kimi 智能助手支持上传的文档多达 500 个，相同成本、相同设备情况下，模型响应速度提升了 3 倍左右。Kimi 后，阿里、百度、360 均宣布了自家大模型对长上下文的支持。

图表5：Kimi 智能助手内测支持 200 万汉字上下文



资料来源：月之暗面、华泰研究

图表6：Kimi 智能助手支持一次性上传数百个文档



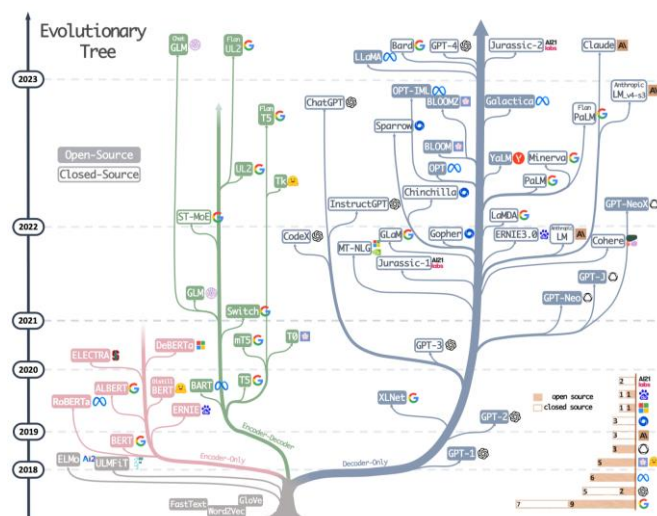
资料来源：月之暗面、华泰研究

实现 Transformer 长上下文，难点和限制在哪？

准备知识#1：Transformer 解码器架构仍是模型主流

Transformer 架构主流地位未被撼动。截至 24 年 3 月，大语言模型（LLM）绝大部分仍然以 Transformer 为基础架构，包括当前最先进的 GPT-4 系列，以及 Google 在 24 年 2 月更新的 Gemini 1.5 Pro，均是基于 Transformer 的解码器架构。虽然有研究者提出了 Mamba 等基于状态空间模型（SSM）的新模型架构，实现了：1）推理时的吞吐量为 Transformer 的 5 倍；2）序列长度可以线性扩展到百万级别；3）支持多模态；4）测试集结果优于同等参数规模的 Transformer 模型。但从工程实现来看，暂时未得到大范围的使用。我们认为，短期内替换掉 Transformer 可能性不大。

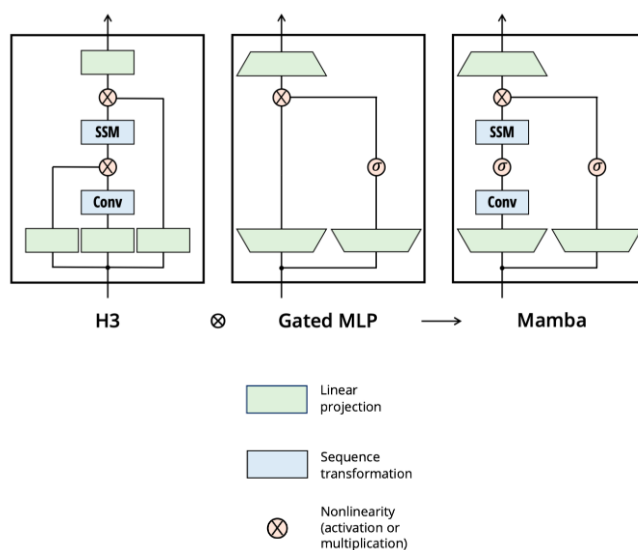
图表7：Transformer 架构仍是 LLM 主流架构



注：灰色以外的分支均为 Transformer 架构，粉色为编码器架构，绿色为编解码器架构，蓝色为解码器架构

资料来源：《Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond》，Jingfeng Yang (2023.4)、华泰研究

图表8：Mamba 架构

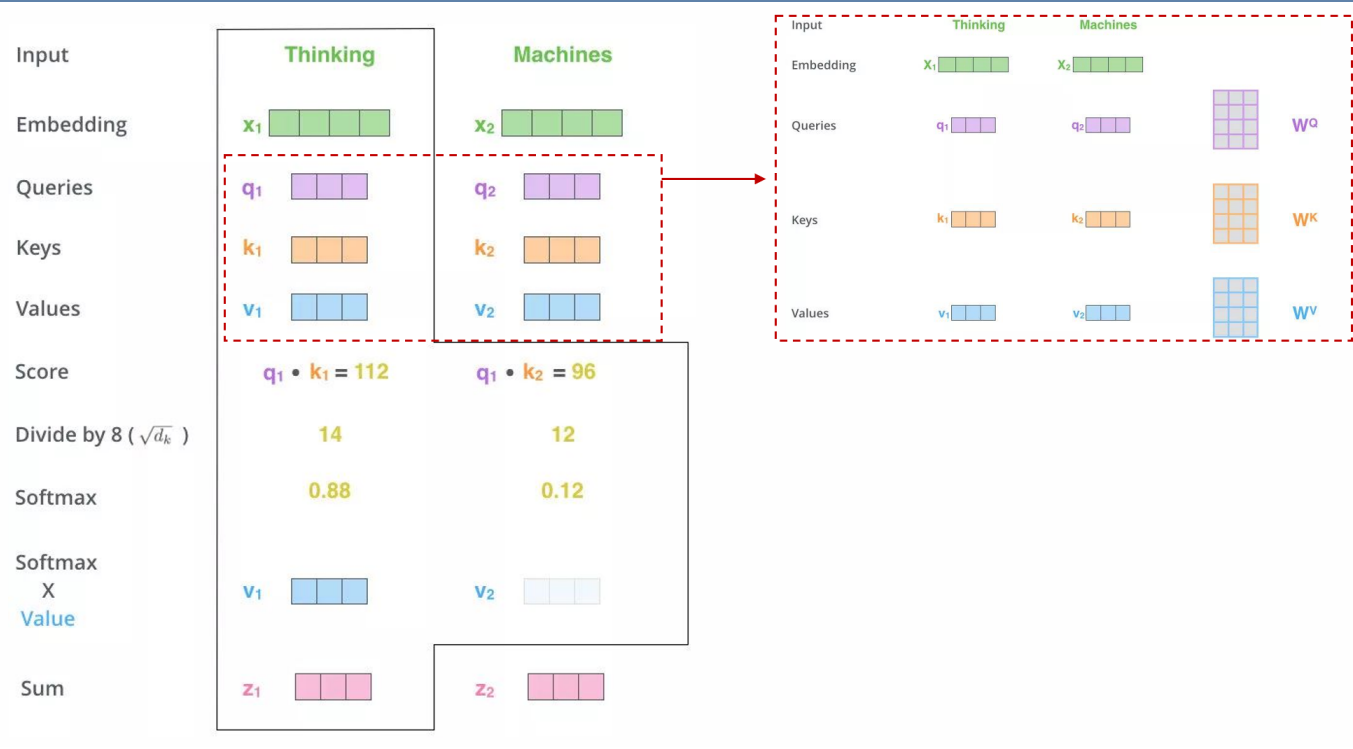


资料来源：《Mamba: Linear-Time Sequence Modeling with Selective State Spaces》，Albert Gu (2023.12)、华泰研究

准备知识#2：Transformer 架构的核心是注意力机制

注意力模型可以用 Query、Key 和 Value 模型进行描述。本质上注意力（Attention）机制是针对数据源的 Query 与 Key，计算出每个 Key 对应 Value 的权重系数（相似性或者相关性），得到最具有吸引力的部分，然后对 Value 进行加权求和，即得到最终的 Attention Value 数值。例如，输入单词“Thinking”，经过 embedding 后用 x_1 表示。 x_1 会分别与 W_Q 、 W_K 、 W_V 矩阵相乘（这三个矩阵为模型要学习的参数），得到“Thinking 单词”（也即 x_1 ）对应的 q_1 、 k_1 、 v_1 。其他所有的输入单词，如“machines”均同理计算。得到每个单词的 Q、K、V 后， q_1 分别与其他单词的 k 进行相似度计算，得到相关性（如图中的 112 和 96），并经过必要的缩放、Softmax 等操作，得到权重。权重与所有的 v 值加权求和后即得到 q_1 做完注意力后的结果 z_1 。 z_2 即其他 token 的计算同理。

图表9：注意力机制的计算方法



资料来源：GitHub、华泰研究

准备知识#3: Transformer 架构需要 KV cache 来存储 KV 信息

KV cache 类似大语言模型（LLM）的内存存储，用来存储每个 token 的 KV 信息。在一次查询 Q 中，LLM 可能会输入或生成很多 token，这些 token 都有对应的 Key-Value 信息。而 KV cache 本质是一个张量列表，将所有先前 token 的 K、V 嵌入存储在每个块的注意力层中，在因果 LLM 的自回归生成过程中使用和更新。**KV cache 的使用机制：** 1) 在有 token 对应的 KV 值进入前，cache 初始化为空；2) 第一个生成的 token 对应的 KV 值会进入 KV cache 存起来；3) 下一个新的 token 对应 KV 值，一方面和 KV cache 中存储的旧 KV 值做注意力机制，另一方面也会将自己存到 KV cache 中；4) 因此，KV cache 随着 token 数的增加，而线性增加。

准备知识#4: Transformer 架构需要位置编码

所有元素两两之间都并行计算注意力机制，保证了模型能够处理复杂的全局关系。即使在长文本时，位于句首的单词（token）也会与末尾单词（token）计算注意力，因此模型能够捕获整个输入中每对 token 的全局依赖关系，使该模型能够处理具有复杂关系的序列。

但由于是并行计算，因此 token 之间没有位置信息，需要额外的位置信息。所有 token 之间的注意力机制都是并行计算的，因此并没有位置信息，需要人为添加。典型的位置编码，包括 Transformer 开篇之作《Attention is all you need》中使用的正弦位置嵌入（position embedding, PE），以及在 Llama 和智谱 GLM 模型中使用的旋转 PE（RoPE）。

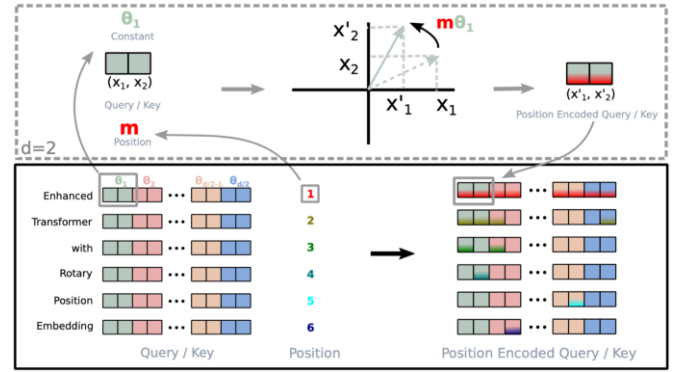
图表10：正弦 PE 和 RoPE 的数学表述

$$\text{SinPE}(n) := \begin{bmatrix} \sin(n\theta^0) \\ \cos(n\theta^0) \\ \sin(n\theta^1) \\ \cos(n\theta^1) \\ \vdots \\ \sin(n\theta^{\frac{d}{2}-1}) \\ \cos(n\theta^{\frac{d}{2}-1}) \end{bmatrix}, \text{ where } \theta = \text{base}^{-\frac{2}{d}}, n \in \{0, 1, \dots, L-1\}$$

$$\text{RoPE}(n) := \begin{bmatrix} R_n^{(0)} & & & \\ & R_n^{(1)} & & \\ & & \ddots & \\ & & & R_n^{(\frac{d}{2}-1)} \end{bmatrix}, \text{ where } R_n^{(i)} := \begin{bmatrix} \cos(n\theta^i) & -\sin(n\theta^i) \\ \sin(n\theta^i) & \cos(n\theta^i) \end{bmatrix}$$

资料来源：《Attention is all you need》，Ashish（2017）、《Roformer: Enhanced transformer with rotary position embedding》，Su（2024）、华泰研究

图表11：RoPE 的可视化表示



资料来源：《The What, Why, and How of Context Length Extension Techniques in Large Language Models – A Detailed Survey》，Saurav（2024）、华泰研究

计算复杂度、内存和长度限制是长文本实现的核心矛盾点

实现 Transformer 的长上下文，难点在于计算复杂度、上下文内存和最大长度限制。

1) 计算复杂度：由于 Transformer 的注意力机制需要 token 间两两计算注意力，因此其时间复杂度 $O(L^2d)$ 和空间复杂度 $O(L^2)$ 与序列长度 (L) 的平方成正比。随着序列长度的提高，对计算资源的需求增长巨大，导致模型时间和算力成本上升。因此不做任何优化，直接扩充上下文长度，盲目的去“堆算力”并非很好的解决方法。

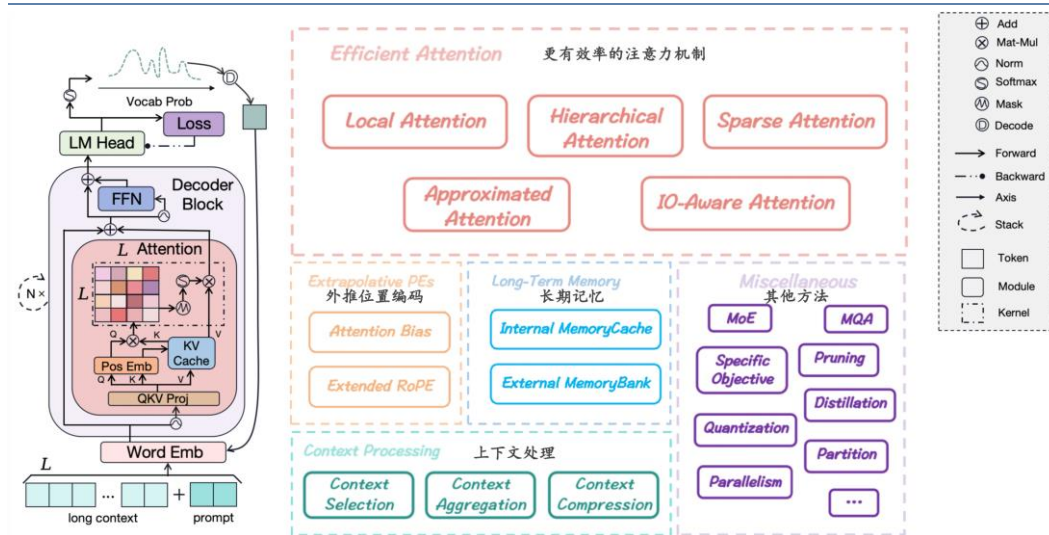
2) 上下文内存：LLM 缺乏显式内存机制，仅依靠 KV cache 在列表中存储所有先前 token 的表示。一旦查询 (Q) 在一次调用中完成，Transformer 不会在后续调用中保留或调用任何先前的状态或序列，除非整个历史记录逐个 token 重新加载到 KV cache 中。因此，每次调用期间模型只拥有较短上下文工作记忆，而不是长短期记忆等固有记忆机制。在聊天机器人应用程序等对长期记忆保留的任务中不适用。

3) 最大长度限制：在训练阶段，工程师通常需要确定一个关键的超参数最大长度，代表了批处理中任何训练样本的序列长度上限 $L_{\max}$ 。通常根据可用的计算资源设置为 1K、2K、4K 或更多，以避免 GPU 上的内存不足的错误。虽然 Transformer 架构理论上可以处理任何长度序列，并没有限制，但是当前的语言模型在处理超过最大值的输入序列时表现出明显的性能下降。所以在推理时，各模型厂商只能限制用户 prompt 上下文，或采取截断操作，使得与训练时序列长度一致，来保证效果。

通过改进 Transformer 架构实现长上下文

拆解 Transformer 解码器，可以通过改进架构中的各个模块来实现上下文长度的拓展。具体优化方法涉及注意力机制、存储、上下文、位置编码等。

图表12：从 Transformer 架构来拆解长上下文的改进方法

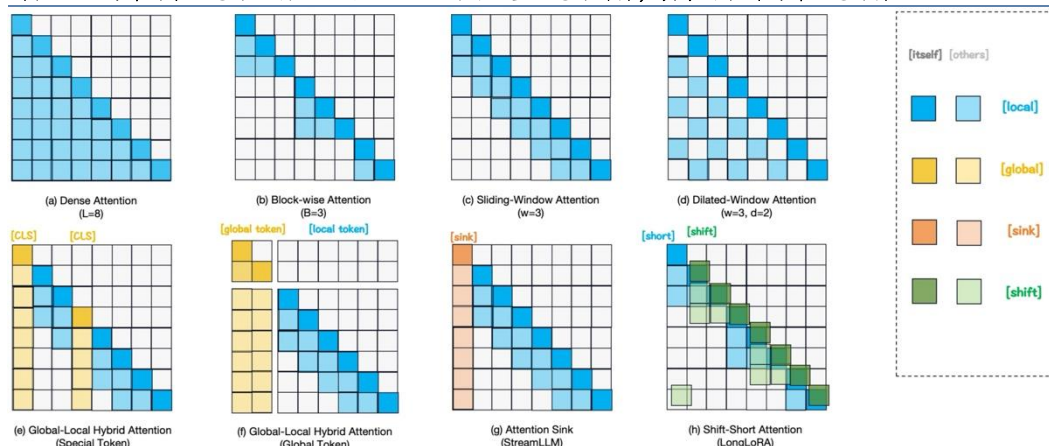


资料来源：《Advancing Transformer Architecture in Long-Context Large Language Models: A Comprehensive Survey》，Huang (2024)、华泰研究

1) 高效注意力机制: 实现更高效的注意力机制，降低计算成本，甚至实现线性时间复杂度。这样在训练时就可以实现更长的序列长度，相应的推理时序列长度也就更长。简言之，一个新的 token 生成后，需要和前面所有已经生成的 token 做注意力机制，但是可以通过不同的方法，舍弃掉一部分注意力的计算，来提高效率。例如局部注意力，将每个 token 的注意力限制在其相邻 token 上；分层注意力，利用更高层级的全局信息和更低层次的局部注意力来实现多尺度的上下文接受域；稀疏注意力，引入稀疏的注意力掩码，既提供了计算效率，也提供了捕获全域上下文信息的能力；近似注意力，以精度为代价，用稀疏或低秩等线性复杂度，来近似完整注意力机制。

另有一种 IO 感知注意力比较特殊，该方法涉及硬件，会考虑内存瓶颈，同时保持注意力核计算的准确性。典型的方法是 Flash Attention，在 GPU HBM 和片上 SRAM 之间进行 IO 感知，减少时间和内存消耗，保持计算精度。Flash Attention 的高效性，已经被直接纳入机器学习框架 Pytorch v2.0 中，被广泛使用。

图表13：局部因果注意力机制的可视化（左上角为全量注意力计算，其余为各种局部注意力算法）



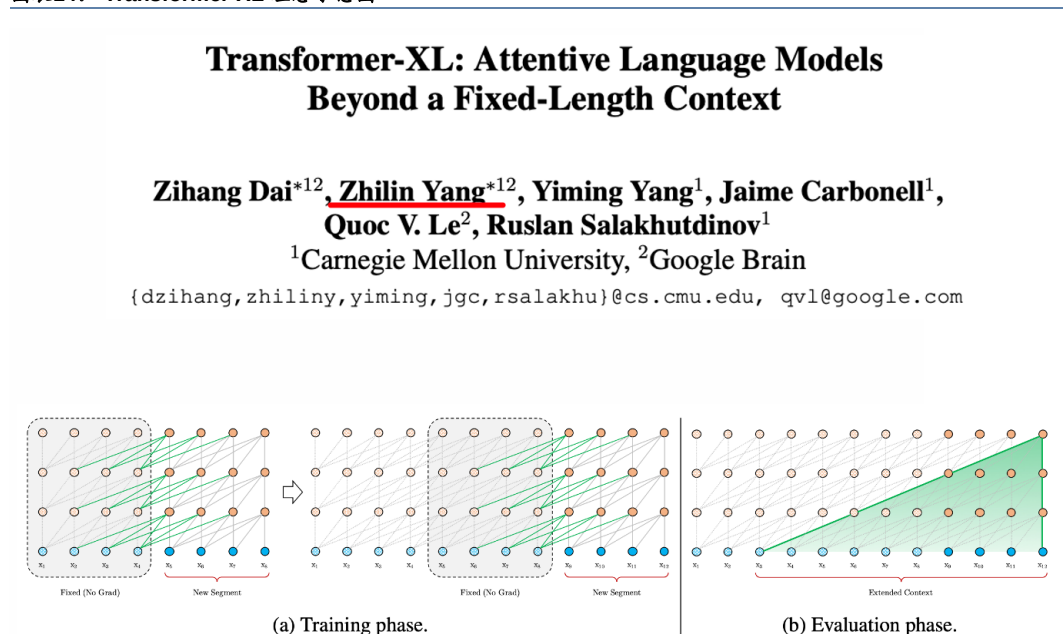
资料来源：《Advancing Transformer Architecture in Long-Context Large Language Models: A Comprehensive Survey》，Huang (2024)、华泰研究

2) 实现长期记忆：设计显式记忆机制，以解决上下文记忆的局限性。由于 LLM 内部的存储支持有限，因此可以通过提供额外的内部或外部存储支持来提高其长期记忆能力，进而提升上下文支持程度。

内部来看，可以基于递归、循环机制，将长文本划分为固定长度的片段流，当前查询的上下文将包含之前片段的缓存或蒸馏信息，相当于之前的信息被保留了下来，变相增加了上下文的支持能力。典型的算法如月之暗面创始人杨植麟的 Transformer-XL。但是递归方法可能面临内存机制小改变就需要模型重训练，或者内存可能会失效等问题。

外部来看，可以为 LLM 提供外部内存，最典型的算法就是 RAG（检索增强生成）。RAG 将模型与长期内存解耦，使用上下文编码器，将长序列作为分段嵌入向量，存储到外部内存库（如向量数据库）。查询期间，模型根据特定标准从该内存库中检索信息，并动态连接它以形成上下文工作内存，从而实现对长上下文的支持。典型代表是 OpenAI 提出的智能体 GPTs，能够在其中上传用户的知识作为 RAG 信息源。

图表14：Transformer-XL 理念示意图



资料来源：《Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context》，Dai（2019）、华泰研究

3) 改进位置编码 PE：对现有的位置编码 PE 进行改进，实现上下文外推。第一类是注意力偏差，给算好的非归一化注意力权重矩阵 P 加上一个偏差值 B ，或者像论文 ALiBi 中引入负因果注意力偏差，实现了在推理比训练时最大序列长度长 16 倍的 token 时，仍然能够保持很低的困惑度（即正确率高）。第二类是扩展 RoPE（旋转位置 embedding）。RoPE 本身性能优异，研究人员用各种策略扩展 RoPE 的上下文长度外推能力，包括 scaling 策略，阶段策略、重排序策略。Scaling 策略如 XPOS、PI、NTK-RoPE、Dynamic-NTK 等。

图表15：外推位置 embedding 的不同方法

注意力偏差

$$\tilde{P} := P + B, \quad B \in \mathbb{R}^{L \times L}, \quad \text{where } B_{ij} := \mathcal{B}(i, j), \quad \forall i, j \in \{0, 1, \dots, L-1\}$$

负因果注意力偏差

$$\mathcal{B}_{ALiBi}^{(h)}(i, j) := -\lambda^{(h)} \cdot |i - j|, \quad \lambda^{(h)} := \frac{1}{2^h} \text{ or } \frac{1}{2^{h/2}}$$

扩展RoPE等方法

$$\text{XPOS: } P_{i,j} := \langle \tilde{\mathbf{q}}_i, \tilde{\mathbf{k}}_j \rangle = \gamma^{i-j} (\mathbf{q}^T R_{j-i} \mathbf{k}), \quad \text{where } \tilde{\mathbf{q}}_i := \gamma^i (R_i \mathbf{q}), \quad \tilde{\mathbf{k}}_j := \gamma^{-j} (R_j \mathbf{k}), \quad i \geq j$$

$$\text{PI: } \tilde{P}_{i,j} := \langle R_{i/\kappa} \mathbf{q}, R_{j/\kappa} \mathbf{k} \rangle = \mathbf{q}^T R_{\frac{j-i}{\kappa}} \mathbf{k}$$

$$\text{NTK: } \tilde{\beta} := c_{\kappa} \cdot \beta, \quad \text{s.t.} \quad \frac{n}{\beta^{d/2-1}} = \frac{n/\kappa}{\beta^{d/2-1}} \Rightarrow c_{\kappa} = \kappa^{2/(d-2)}$$

$$\text{Power Scaling: } \tilde{\beta}^i := \beta^i / (1 - 2i/d)^{\kappa}$$

资料来源：《Advancing Transformer Architecture in Long-Context Large Language Models: A Comprehensive Survey》，Huang（2024）、华泰研究

4) 对上下文进行处理：用额外的上下文预/后处理，在已有的 LLM（视为黑盒）上改进，确保每次调用中给 LLM 的输入始终满足最大长度要求。这种方法“治标不治本”，模型本身的能力并没有提高，只是在模型的输出和输入文本上做处理。第一类是上下文选择，将长文本划分为短的小段，并根据预定义的标准选择特定段处理。例如 LangChain 会在长文本划分为小段后，分别查询 LLM 以独立输出每个段的答案和置信度分数，并选择置信得分最高的答案作为最终输出。第二类与第一类相似，同样会分段查询，但是在最终给结果时会从每个片段中单独提取相关信息，然后采用各种融合策略来聚合检索到的信息，得出最终答案。第三类是上下文压缩，模型提供商会将输入的长上下文，压缩到支持的最大长度。例如学习一个更短的表述代替用户的长文本表述（软压缩），或根据预训练模型计算各种分值来去除用户上下文中的冗余信息（硬压缩）。

5) 其他方法：以更广泛的视角来增强 LLM 的有效上下文窗口或优化使用现成 LLM 时的效率。包括 **MoE**：MoE 混合专家架构增强了通用性，减少了计算需求，并提高了建模大规模上下文的效率和有效性。**特定的优化目标函数**：适应特定任务的预训练，旨在提高 LLM 在较长文本中捕获复杂的长期依赖性和话语结构的有效性。如杨植麟的 XL-Net，百度的 ERNIE-Doc。**并行策略**：利用节点内部和跨节点的聚合 GPU 内存，引入各种并行策略来扩展模型大小和扩展序列长度。如数据并行、张量并行、pipeline 并行、序列并行、专家并行（将不同的专家放在不同的 GPU 上，并行执行）。**权重压缩**：例如剪枝、因子分解、量化、分区和蒸馏等，其中量化策略在大规模 LLM 的实际部署中作用关键，通过降低参数精度，从而缓解记忆需求并加速训练和推理（如 B200 配套推出新的 FP4/FP6 等精度）；或者多查询注意力（MQA）和分组查询注意力（GQA）能够缓解较大的 KV cache 压力。

大模型产品长上下文实现方法预测

大模型的长上下文实现方法是闭源模型厂商的核心 Know-How，并未公开。以 Google Gemini 1.5 Pro 为例，Google 已经在内部实现了 10M tokens 的上下文长度，几乎是全球大模型支持上下文最长的。但其技术报告中仅描述为“Gemini 1.5 Pro 在一系列重要架构上改进，才将上下文扩展到 10M tokens”。

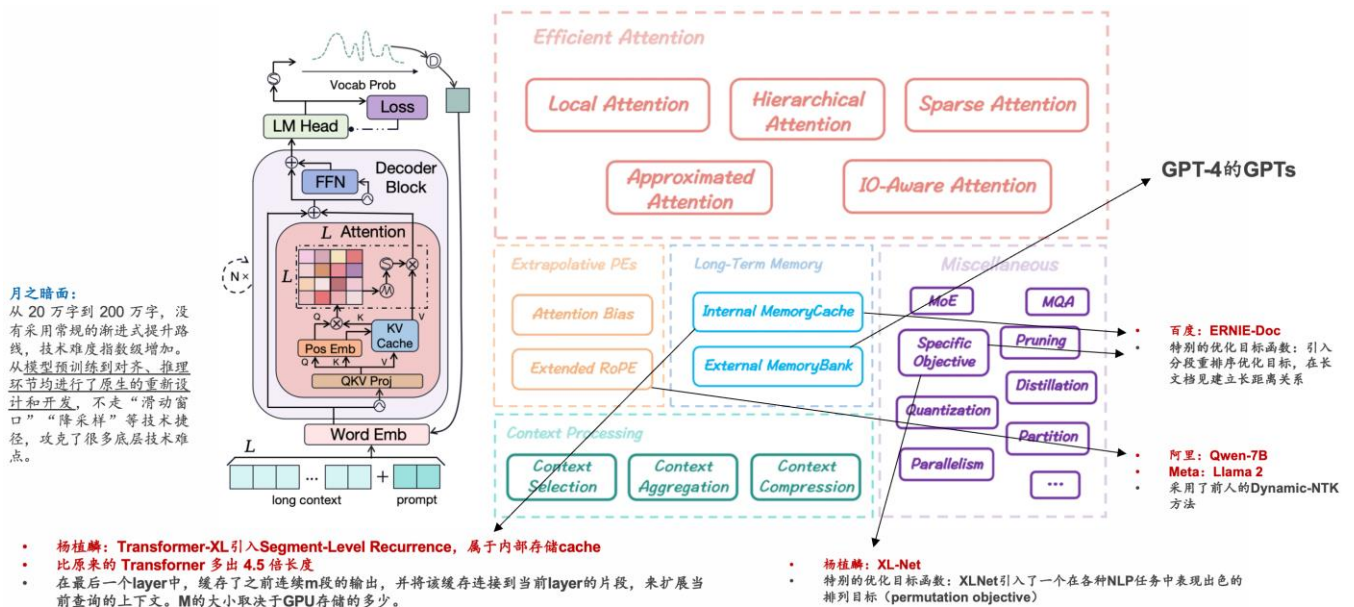
图表16： Gemini 1.5 Pro 技术报告关于长上下文实现方法的描述

Gemini 1.5 Pro also incorporates a series of significant architecture changes that enable long-context understanding of inputs up to 10 million tokens without degrading performance. Translated into

资料来源： Gemini 1.5 Pro 技术报告、华泰研究

国内大模型厂商实现长上下文，可能采取了多种路线的混合优化方法。以月之暗面为例，其创始人杨植麟主要的学术论文 Transformer-XL 和 XL-Net，均探讨了长上下文的实现方法，且前者属于长期记忆力的优化，后者属于特殊目标函数的优化。另外，月之暗面官方表述，“从 20 万字到 200 万字，没有采用常规的渐进式提升路线，技术难度指数级增加。从模型预训练到对齐、推理环节均进行了原生的重新设计和开发，不走滑动窗口、降采样等技术捷径，攻克了很多底层技术难点。”百度的 ERNIE-Doc 则同时采用了长期记忆力和特殊目标函数的优化方法。阿里 Qwen-7B 则使用了优化的位置编码算法，即 extended RoPE。所以我们推测，国内模型厂商之所以能够在短期内实践出长上下文方法，或是在原有积累的基础上进行了改进，采取多方法的混合优化。

图表17： 长上下文实现方法的预测



资料来源：《Advancing Transformer Architecture in Long-Context Large Language Models: A Comprehensive Survey》，Huang（2024）、华泰研究

长上下文的通用性将解决多类场景需求

长上下文的通用性，将一定程度上代替模型的微调。通常在大模型预训练完毕之后，对于特定的场景和任务，往往需要利用该场景下的数据进行进一步的微调训练。虽然微调的算力需求远低于预训练，但是仍需要准备数据集并进行模型训练，步骤上仍较为繁琐。但是模型的长上下文可以很好的解决这一问题，用户只需要将特定领域的知识通过上下文的方式输入到模型中，模型即可以通过上下文学习掌握相应内容，避免了微调的繁琐步骤，实现了模型在各种场景下的通用性。

长上下文模型能适应虚拟角色、长 prompt、AI Agent、垂直客户需求等多种场景。1) 虚拟角色 Chatbot: 长文本能力能够使得虚拟角色长期记住用户的个性化信息，提升产品体验。2) 开发者需求: 对于大模型开发者来说，输入 prompt 长度的限制约束了大模型应用的场景和能力的发挥，比如基于大模型开发剧本杀类游戏时，往往需要将数万字甚至超过十万字的剧情设定以及游戏规则作为 prompt，只有长文本能力才能满足开发需求。3) AI Agent: 由于 Agent 运行需要自动进行多轮规划和决策，且每次行动都需要参考历史记忆信息才能完成，因此需要更长的上下文来全面准确的基于历史信息进行新的规划和决策，提高成功率。4) 垂直场景客户需求: 在使用大模型作为法律、金融等工作助理完成任务的过程中，需要将较长的资料作为上下文输入模型，只有长文本支持能力才能实现长文档的推理分析。

关注大模型长文本潜在受益产业链

大模型的长文本能力，有望赋能现实中需要长文本输入的多种应用场景。此外，根据 OpenAI 发表的 Scaling Law 论文，推理所需算力近似为 $2^{\text{模型参数} \times \text{数据 token}}$ 。虽然各厂商会对长文本进行算力效率优化，但是一般来说长文本意味着更长的 token，因此长文本推理的算力需求同样增长。相关产业链公司有：

长文本应用场景：1) 文本工具：金山办公、福昕软件；2) 法律文案：华宇软件、通达海；3) 业务流程：泛微网络、致远互联；4) 其他文本：汉仪股份、汉王科技。

专业领域+多任务+多模态场景：1) 金融领域：同花顺、恒生电子；2) 医疗领域：嘉和美康；3) 电商领域：光云科技。

AI 算力：浪潮信息、神州数码、海光信息。

图表18：提及公司列表

公司代码	公司简称
MSFT US	微软
GOOGL US	谷歌
Meta US	Meta
BIDU US	百度
BABA US	阿里
601360 CH	360
未上市	Anthropic
未上市	OpenAI
未上市	月之暗面
未上市	阶跃星辰

资料来源：Bloomberg、华泰研究

风险提示

宏观经济波动。若宏观经济波动，产业变革及新技术的落地节奏或将受到影响，宏观经济波动还可能对 AI 投入产生负面影响，从而导致整体行业增长不及预期。

技术进步不及预期。若 AI 技术和大模型技术进步不及预期，或将对相关的行业落地情况产生不利影响。

本研报中涉及到未上市公司或未覆盖个股内容，均系对其客观公开信息的整理，并不代表本研究团队对该公司、该股票的推荐或覆盖。

图表19：重点推荐公司一览表

股票名称	股票代码	投资评级	最新收盘价	目标价	市值 (百万)	EPS (元)				PE (倍)			
			(当地币种)	(当地币种)	(当地币种)	2022	2023E	2024E	2025E	2022	2023E	2024E	2025E
金山办公	688111 CH	买入	306.40	417.81	141,501	2.42	2.85	3.63	4.85	126.61	107.51	84.41	63.18
福昕软件	688095 CH	买入	68.60	86.72	6,276	-0.02	-1.02	-0.03	0.32	-3,430.00	-67.25	-2,286.67	214.38
泛微网络	603039 CH	买入	39.15	74.12	10,203	0.86	0.98	1.57	2.18	45.52	39.95	24.94	17.96
汉王科技	002362 CH	买入	20.96	25.74	5,124	-0.55	0.13	0.45	0.53	-38.11	161.23	46.58	39.55
同花顺	300033 CH	买入	138.00	169.06	74,189	3.15	2.61	3.19	3.92	43.81	52.87	43.26	35.20
恒生电子	600570 CH	买入	23.43	35.19	44,517	0.57	0.75	0.93	1.17	41.11	31.24	25.19	20.03

资料来源：Bloomberg，华泰研究预测

图表20：重点推荐公司最新观点

股票名称	最新观点
金山办公 (688111 CH)	AI 产品化顺利推进，订阅收入有望加速增长 金山办公发布年报，2023 年实现营收 45.56 亿元（yoy+17.27%），归母净利 13.18 亿元（yoy+17.92%），扣非净利 12.62 亿元（yoy+34.45%）。其中 Q4 实现营收 12.86 亿元（yoy+17.99%，qoq+17.06%），归母净利 4.24 亿元（yoy+39.44%，qoq+44.40%）。2023 年公司价格体系调整完成，推出 WPS AI 实现 AI 产品化，未来随着 AI 逐步改善产品使用体验。2B 业务逐步推进，订阅收入或加速增长。我们预计 24-26 年 EPS 为 3.63/4.85/6.44 元，收入 58.47/77.36/102.64 亿元，可比公司平均 24E 20.4xPS(Wind)，考虑公司或率先实现 AI 商业化，给予 24 年 33xPS，目标价 417.81 元，维持“买入”。 风险提示：付费用户增长不及预期、AI 商业化不及预期。 报告发布日期：2024 年 03 月 21 日 点击下载全文：金山办公(688111 CH,买入)：订阅驱动增长，AI+成效或逐步显现
福昕软件 (688095 CH)	大模型能力持续提升，关注 AI 应用产品化落地 2024 年 3 月 4 日 Anthropic 发布第三代 Claude 模型，在长文本及多模态等方面能力均有所提升，长文本方面该模型的三种版本都能接受超过 100 万个 token 的输入；多模态方面可以处理照片、表格、图形、科研图表等内容。我们认为海外大模型能力的持续提升或将推动 AI 应用端产品的不断迭代。公司具备良好的产品基础、海外业务基础，随着转型进入新阶段，收入增速有望逐步加速。预计 2023-2025 年 EPS 为 -1.02/-0.03/0.32 元。可比公司 24 年 ifind 一致预期 PS 均值为 10.6 倍，给予 2024 年 10.6 倍 PS，目标价 86.72 元，维持“买入”评级。 风险提示：新应用场景拓展不及预期；市场竞争风险；联营企业业绩波动风险。 报告发布日期：2024 年 03 月 07 日 点击下载全文：福昕软件(688095 CH,买入)：模型能力持续提升，关注 AI 应用落地
泛微网络 (603039 CH)	费用投入加大业绩承压，维持“买入”评级 公司 23Q1-Q3 收入 13.98 亿元，同增 4.0%；归母净利 0.11 亿元，同减 90.8%；扣非归母净利-0.13 亿元，同减 127.3%；23Q3 单季收入 4.92 亿元，同减 11.3%；归母净利-0.27 亿元，同减 151.5%；扣非归母净利-0.35 亿元，同减 755.9%。业绩承压的主要原因为收入增速放缓，销售、研发费用投入力度加大。我们认为公司持续加大投入，产品竞争力有望持续增强，随着下游行业景气度恢复，收入增速有望回升。预计 2023-2025 年公司 EPS 为 0.98/1.57/2.18 元，参考可比公司 2023 年平均估值 75.4x PE（Wind），给予 2023E 75.4x PE，对应目标价 74.12 元，维持“买入”评级。 风险提示：产品拓展不及预期，下游需求不及预期，市场竞争加剧。 报告发布日期：2023 年 10 月 28 日 点击下载全文：泛微网络(603039 CH,买入)：LLM 产品持续升级，成长空间打开
汉王科技 (002362 CH)	收入端逐步恢复，利润端短期承压 汉王科技发布三季报，2023 年 Q1-Q3 实现营收 9.73 亿元（yoy+5.04%），归母净利-8985 万元（yoy-98.32%），扣非净利-8919 万元（yoy-86.04%）。其中 Q3 实现营收 3.52 亿元（yoy+5.64%，qoq+12.44%），归母净利-3799 万元（yoy-13259.32%，qoq-73.11%）。利润端短期承压，系公司创新技术、产品、AI 等研发投入，以及新产品上市市场投入增加。我们预计公司 2023-2025 年 EPS 分别为 0.13、0.45、0.53 元。可比公司 24 年 Wind 一致预期 PE 均值为 57.2 倍，给予公司 24 年 57.2 倍 PE，目标价 25.74 元（前值 34.13 元），维持“买入”评级。 风险提示：下游需求不及预期，产品化进展低于预期。 报告发布日期：2023 年 10 月 28 日 点击下载全文：汉王科技(002362 CH,买入)：AI+产品矩阵完善，利润端短期承压
同花顺 (300033 CH)	2023 年投入加大利润承压，有望受益于 AI 机遇 同花顺发布年报，2023 年实现营收 35.64 亿元（yoy+0.14%），归母净利 14.02 亿元（yoy-17.07%），扣非净利 13.69 亿元（yoy-14.80%）。其中 23Q4 实现营收 11.92 亿元（yoy-7.14%，qoq+32.15%），归母净利 6.32 亿元（yoy-21.86%，qoq+102.96%）。23 年净利下滑原因：1）资本市场波动影响收入增速；2）公司把握 AI 机遇，加大销售研发投入力度；考虑公司加大研发销售投入，预计公司 2024-2026 年 EPS 为 3.19/3.92/4.86 元（前值 24-25 年 4.01/4.76 元）。可比公司 24 年平均 53xPE(Wind)，给予公司 24 年 53xPE，目标价 169.06 元（前值 204.04 元），维持“买入”评级。 风险提示：A 股交易量波动；技术落地效果不及预期。 报告发布日期：2024 年 02 月 28 日 点击下载全文：同花顺(300033 CH,买入)：投入加大利润承压，受益于 AI 机遇
恒生电子 (600570 CH)	恒生 2023 年营收 yoy+11.98%、归母净利润 yoy+30.50%，高于预告 预计 2023 年实现营收 72.81 亿（yoy+11.98%），略高于预告；归母净利润 14.24 亿（yoy+30.50%），扣非净利润 14.48 亿（yoy+26.51%），高于预告（业绩预告：营收 72.92 亿，归母净利润 13.45 亿，扣非净利润 13.78 亿）；鉴于公司成本费用管控力度超我们预期，因此上调公司 24-25 年 EPS 分别至 0.93/1.17 元（前值 0.89/1.11 元），新增 2026 年 EPS 为 1.33 元。可比公司 2024PE 均值 38 倍（Wind），我们仍看好恒生核心竞争力及未来发展前景，给予公司 24 年目标 PE 38 倍，给予目标 35.19 元（前值 31.08 元），维持“买入”评级。 风险提示：下游景气度不及预期，行业竞争加剧，AI 产品推进不及预期。 报告发布日期：2024 年 03 月 25 日 点击下载全文：恒生电子(600570 CH,买入)：营收稳健增长，迈入精细化管理阶段

资料来源：Bloomberg，华泰研究预测