

行业研究 | 专题报告 | 软件与服务

月之暗面：长文本能力点亮国产大模型发展之路

报告要点

月之暗面迅猛发展已成为国内大模型领域的领军企业之一。公司创始人杨植麟师从中美各路 AI 大拿，学术成果颇丰，在大模型长文本识别问题曾做出重要研究成果，并深度参与了多个国产大模型的研发工作，是中国大模型产业的业界领袖之一。他与清华校友共同打造了月之暗面，旗下核心产品，Kimi Chat 以对话交互方式为 C 端用户提供服务，长文本处理能力优秀。其特质体现了我国基础模型能力的跃升。

分析师及联系人



宗建树

SAC: S0490520030004



范超

SAC: S0490513080001

SFC: BQK473

月之暗面：长文本能力点亮国产大模型发展之路

月之暗面：国产 AI 独角兽

月之暗面迅猛发展已成为国内大模型领域的领军企业之一。公司创始人杨植麟师从中美各路 AI 大拿，学术成果颇丰，在大模型长文本识别问题曾做出重要研究成果，并深度参与了多个国产大模型的研发工作，是中国大模型产业的业界领袖之一。杨植麟在创立月之暗面之前，已积累了一定创业经验，后又联合其清华校友，共同打造国产“OpenAI”——月之暗面，目前被资本市场看好，已成为 AI 独角兽。

Kimi Chat：长文本技术世界领先

技术迭代助力性能跃升，Kimi Chat 长文本处理能力优秀。作为月之暗面旗下核心产品，Kimi Chat 以对话交互方式为 C 端用户提供服务，主要应用场景包括文本提炼、信息检索、数据处理、模拟交流等。Kimi Chat 以长文本处理技术著称，同时具备优秀的联网搜索与指令遵循能力。**1) 在长文本能力方面**，当前 Kimi Chat 长文本处理能力达 200 万字，性能较半年前提升 9 倍，对应约 133 万 Tokens，已超越 Gemini 1.5 Pro (100 万 Tokens)、Claude 3 (20 万 Tokens) 以及 GPT-4 Turbo (12.8 万 Tokens)，为全球领先水平。**2) 在模型性能方面**，Kimi chat 表现与 GPT4 接近，在学术论文总结、人物观点搜索、PDF 文件解析等任务环节，显示出较好的文本处理能力与交互效果。**3) 在定价方面**，上下文处理能力最强的 moonshot-v1-128k 每百万 tokens 的计价为 60 元，低于相同文本处理水平下的其他模型 (GLM4 与 GPT4 Turbo)。

如何看待 KIMI 的长文本能力跃升？

KIMI 长文本能力特质体现了我国基础模型能力的跃升。大模型的开发团队会在确保模型性能的前提下设定一个最大上下文窗口大小，窗口大小决定了模型在一次前向传播中能够处理的最大文本长度。本次 KIMI 实现了长文本能力的突破代表了公司在模型架构、内存管理、注意力机制优化等层面有深厚的技术积累。KIMI 模型的 lossless long-context 机制使得模型在推理阶段能够结合丰富的 Prompt 内容，避免“断章取义”，使得推理更加准确。此外，长文本处理能力使 KIMI 能够直接处理大型文件 Prompt，适用于复杂和需要大量文本分析的 B 端重应用场景，如办公、医疗、教育等垂类场景。相比 C 端轻应用场景，重应用场景用户留存率较高，付费意愿更为强烈。因此长文本能力或成为 KIMI 等大模型货币化之路的核心要点之一。KIMI 的 lossless long-context 机制或可以取代模型调参，催生个性化模型的诞生。KIMI 模型提升了推理阶段用户数据的权重，模型可以根据用户的行为进行反馈，因此建立了数据飞轮效应。从长期角度来看，lossless long-context 机制在未来有望取代模型的调参步骤，用户可以根据与模型长期互动建立个性化的模型。

风险提示

- 1、AI 模型技术推进不及预期。
- 2、AI 模型下游需求不及预期。

请阅读最后评级说明和重要声明

市场表现对比图(近 12 个月)



资料来源：Wind

相关研究

- 《新质生产力领航，低空经济将腾飞》2024-03-18
- 《Grok-1 专家模型开源，行业应用加速发展》2024-03-18
- 《关键词：NVIDIA 2024 GTC 大会 2024 年第 11 周计算机行业周报》2024-03-17



更多研报请访问
长江研究小程序

目录

月之暗面：国产 AI 独角兽	4
技术立身，公司创始人学术背景深厚	4
从学术到实业，明星团队打造月之暗面	6
Kimi Chat：长文本技术世界领先	8
如何看待 KIMI 的长文本能力跃升？	11
风险提示	14

图表目录

图 1：月之暗面发展历程	4
图 2：杨植麟与两位博士导师 Ruslan Salakhutdinov（最右）、William Cohen（最左）合影	4
图 3：杨植麟文章被大量引用	5
图 4：XLNet 架构	6
图 5：杨植麟个人经历	6
图 6：杨植麟在华为云 TechWave 全球技术峰会介绍盘古 NLP 大模型	7
图 7：北京月之暗面科技有限公司股权情况	8
图 8：Kimi chat VS GPT4：搜索总结	9
图 9：Kimi chat VS GPT4：论文总结	9
图 10：Kimi chat VS GPT4：财报解析	10
图 11：Kimi chat 上市至今已经历 4 轮迭代	10
图 12：按 200 万字测算，Kimi chat 长文本处理能力或领先全球	10
图 13：对比其他大模型，Kimi Chat 文本处理能力接近 GPT4	11
图 14：Kimi Chat 在长文本摘要、多选问答两个环节得分领先	11
图 15：长文本能力使得 KIMI 能依据更长的 Prompt 进行推理	12
图 16：KIMI-Chat 鼓励用户尝试超长文本的任务	13
图 17：KIMI 在长文本任务中有较好的表现	13
表 1：杨植麟参与的部分学术论文（截至 2024 年 3 月 21 日）	5
表 2：月之暗面创始股东学术背景	7
表 3：月之暗面融资情况	8
表 4：主要大模型定价及上下文窗口对比	11

月之暗面：国产 AI 独角兽

月之暗面迅猛发展已成为国内大模型领域的领军企业之一。月之暗面（Moonshot AI）成立于 2023 年 3 月，并于同年 10 月快速推出对话类产品 Kimi Chat，以 to C 为发展重点，寻求大模型的应用落地。2024 年 3 月 18 日，公司宣布 Kimi 智能助手已支持 200 万字超长无损上下文，在长文本这一能力取得突破，并在世界范围内领先。公司也借此成为我国 AI 独角兽之一，受到市场广泛关注。

图 1：月之暗面发展历程

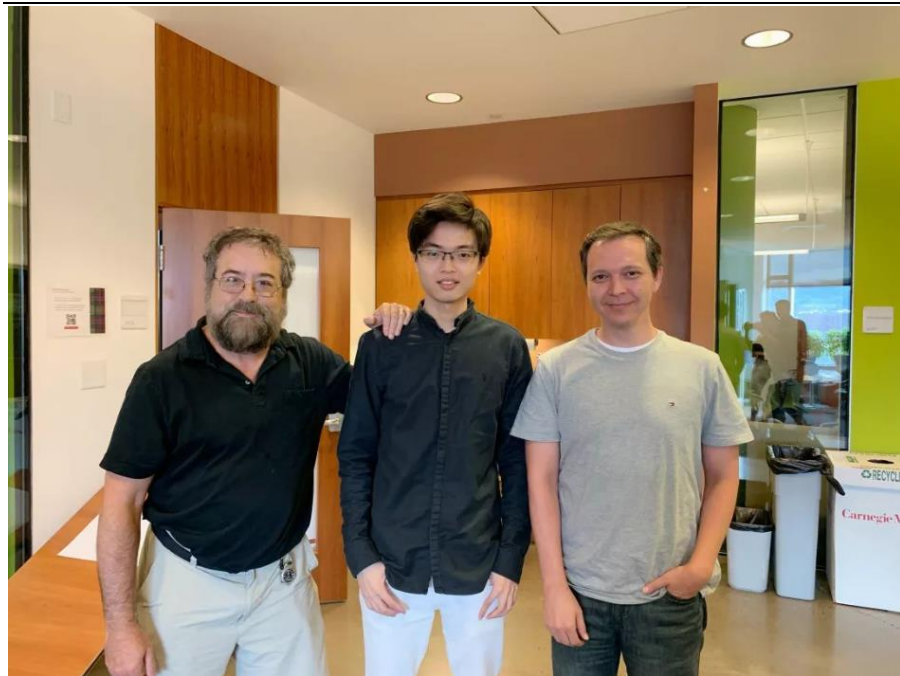


资料来源：澎湃新闻，IT 之家，长江证券研究所

技术立身，公司创始人学术背景深厚

公司创始人杨植麟师从中美各路 AI 大拿。杨植麟本科就读于清华大学计算机系，师从 IEEE Fellow 唐杰，2015 年以年级第一名的优异成绩毕业；随后，赴自然语言处理 (NLP) 研究全球排名第一的卡内基梅隆大学语言技术研究所 (LTI) 攻读博士，师从知名学者 Ruslan Salakhutdinov 与 William Cohen。

图 2：杨植麟与两位博士导师 Ruslan Salakhutdinov（最右）、William Cohen（最左）合影

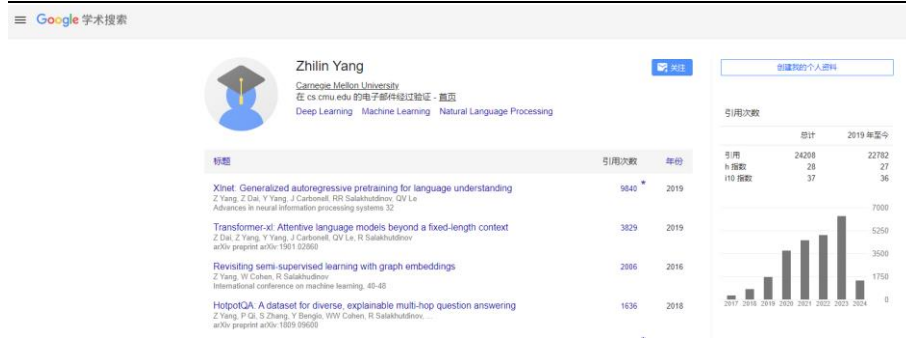


资料来源：腾讯云，长江证券研究所

杨植麟与多位图灵奖得主合作发表论文，学术成果颇丰。据不完全统计，杨植麟曾在 ICLR、NeurIPS、ICML、ACL、EMNLP 等计算机顶会发表论文 20 余篇。他曾与图灵

奖得主 Yoshua Bengio 合作发布「火锅问答」数据集 HotpotQA，并以一作身份发表《XLNet: Generalized autoregressive pretraining for language understanding》和《Transformer-XL: Attentive language models beyond a fixed-length context》，在 NLP 领域产生重要影响，成为 NeurIPS 2019 与 ACL 2019 的最高引论文之一，截至 2024 年 3 月 21 日，《XLNet》Google Scholar 引用次数高达 9840，《Transformer-XL》也超 3800 次，在业内具有强烈影响力。

图 3：杨植麟文章被大量引用



资料来源：Google Scholar，长江证券研究所

表 1：杨植麟参与的部分学术论文（截至 2024 年 3 月 21 日）

论文题目	发表时间	发布刊物	引用次数	第几作者
XLnet: Generalized autoregressive pretraining for language understanding	2019	NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems	9840	一作
Transformer-xl: Attentive language models beyond a fixed-length context	2019	Annual Meeting of the Association for Computational Linguistics	3829	共同一作
Revisiting semi-supervised learning with graph embeddings	2016	ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning	2006	一作
HotpotQA: A dataset for diverse, explainable multi-hop question answering	2018	Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing	1636	一作
GPT understands, too	2023	AI Open	986	通讯作者
Glm: General language model pretraining with autoregressive blank infilling	2021	Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics	638	通讯作者
Multi-task cross-lingual sequence tagging from scratch	2016	arXiv	626	一作
P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks	2021	Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics	617	通讯作者
Differentiable learning of logical rules for knowledge base reasoning	2017	NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems	615	二作
Good semi-supervised learning that requires a bad gan	2017	NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems	545	共同一作

资料来源：Google Scholar，ACM，Semantic Scholar，ACL Anthology，ScienceDirect，arXiv，长江证券研究所

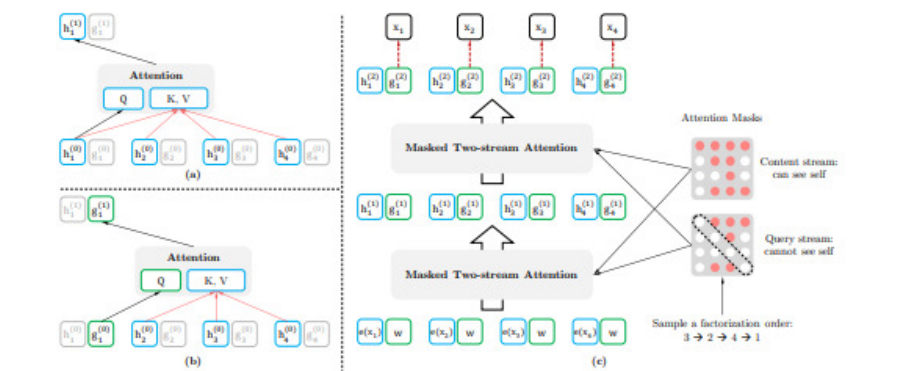
杨植麟曾在谷歌就大模型长文本识别问题做出重要研究成果。读博后期，他曾在 Facebook 人工智能研究院和 Google Brain 研究院工作，Google Brain 创始成员 Quoc Le 也同时参与《XLNet》和《Transformer-XL》两篇论文的研究。

- Transformer-XL 针对 Transformer 对长文本编码效果差的问题，围绕如何建模长距离依赖，提出 Transformer-XL，利用片段级递归机制解决编码长距离依赖和上下文碎片化问题，同时利用相对位置编码机制解决可能出现的时序混淆问题。

- XLNet 使用了《Transformer-XL》的优化方式，并提出了 Permutation Language Model (PLM)将 AR (Auto Regression, 自回归) 和 AE (Auto Encoder, 自编码器) 相结合以更好地捕捉上文信息并改善了模型对长距离依赖性的处理能力。

从这两篇论文不难看出，杨植麟在深度学习、机器学习、NLP 等方面有着深厚学术功底，这也为 Kimi 的长文本能力奠定了坚实的理论基础。

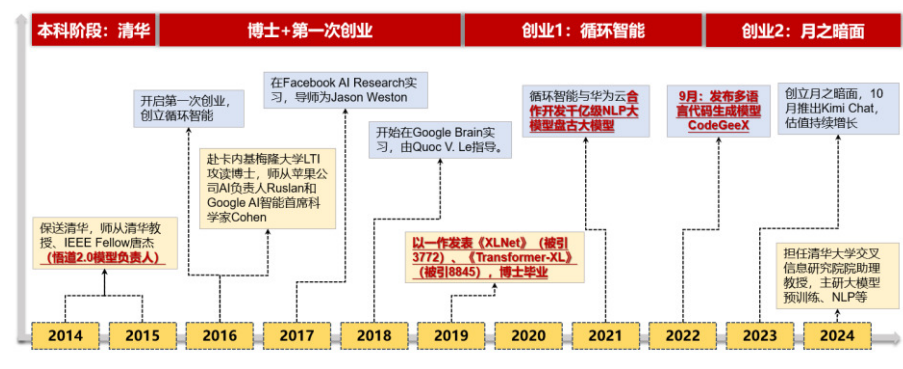
图 4: XLNet 架构



资料来源：《XLnet: Generalized autoregressive pretraining for language understanding》，杨植麟等人，长江证券研究所

杨植麟深度参与了多个国产大模型的研发工作，是中国大模型产业的业界领袖之一。悟道大模型是智源研究院副院长、清华大学教授唐杰带头研发的我国最早的大模型之一，杨植麟作为唐杰的得意门生、智源青年科学家，深度参与项目研发，是多篇关键论文的通讯作者。智谱 AI 所采用的 GLM 大模型也得益于杨植麟的深度参与。此外，他还率领团队联合华为云打造了华为云盘古 NLP 大模型。

图 5: 杨植麟个人经历



资料来源：Github，汕头发布，清华大学官网，AI 科技评论公众号，长江证券研究所

从学术到实业，明星团队打造月之暗面

杨植麟在创立月之暗面之前，已积累了一定创业经验。2016 年，杨植麟在博士二年级就读时就创立循环智能 (Recurrent.ai)，布局 AI 在销售科技领域的商业应用，并于 2021 年与华为云合作开发了千亿级 NLP 大模型盘古大模型，在部分实际场景中达到了优于 Bert 与 GPT 系列的效果。

图 6：杨植麟在华为云 TechWave 全球技术峰会介绍盘古 NLP 大模型



资料来源：华为云，长江证券研究所

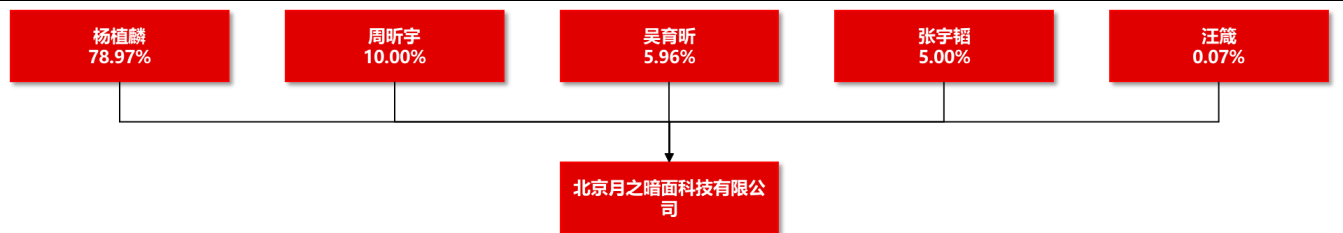
循环智能之后，杨植麟又联合其清华校友，共同打造国产“OpenAI”。除杨植麟外，公司创始成员张宇韬、周昕宇和吴育昕均出身清华大学计算机系，他们在知识图谱构建、计算机视觉、深度网络算法等各自研究方向均有优异的学术成果。创始团队核心成员还参与了 Google Gemini、Google Bard、盘古 NLP、悟道等多个大模型的研发，多项核心技术被 Google PaLM、Meta LLaMA、Stable Diffusion 等主流产品采用。

表 2：月之暗面创始股东学术背景

职位		学术背景
杨植麟	创始人，实控人	曾任循环智能联合创始人。CMU 计算机博士，师从苹果公司现任 AI 负责人、深度学习奠基人之一的 Ruslan Salakhutdinov；曾效力于全球顶级人工智能机构 Facebook AI Research, Google Brain；于 ICLR、NIPS、ICML、KDD、ACL 等顶级 AI 会议发表论文二十余篇；在所有六个主流语言建模数据集保持世界第一名 (State-of-the-art)。
张宇韬	联合创始人，CTO	清华大学计算机博士，师从清华大学计算机系副系主任、数据挖掘顶级专家唐杰教授；曾作为核心开发者研发了全球知名的科技大数据分析平台 AMiner，产品服务于 BATH 等科技巨头及国家科技部等政府科研管理机构。
周昕宇	联合创始人	清华大学 2011 级本科生，毕业后选择加入旷视，以共同一作的身份撰写《ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices》，中标 CVPR。
吴育昕	联合创始人	毕业于清华大学与卡耐基梅隆大学，曾获 2018 年欧洲计算机视觉会议 (ECCV) 最佳论文提名。他是 Meta (Facebook) 员工，系公司人工智能实验室 FAIR 团队的一员，曾和队员何恺明共同提出了组归一化 (Group Normalization, 简称 GN) 的方法。
汪箴	创始成员	毕业于复旦大学计算机科学系，后就读新加坡南洋理工大学计算机工程系，主要研究工作面向 AI 在鉴证科学上的应用。曾在 Google 担任高级软件工程师/技术主管，开发移动端网页搜索、社交平台等产品。后又担任新加坡 AI 技术驱动独角兽领创集团 (ADVANCE.AI) 技术委员会负责人/技术副总裁。

资料来源：36Kr，量子位，长江证券研究所

图 7：北京月之暗面科技有限公司股权情况



资料来源：Wind，长江证券研究所

月之暗面被资本市场看好，已成为 AI 独角兽。从 2023 年 3 月成立至今，月之暗面在一年时间内已完成 3 轮融资，收到资本市场认可。今年 2 月完成新一轮超 10 亿美金融资后，估值已达到 25 亿美元，投资方包括阿里、红杉中国、小红书、美团等。目前，公司已成为国内大模型领域的头部企业之一。

表 3：月之暗面融资情况

时间	投资方	金额	企业估值
2023 年 6 月	红杉中国、真格基金	超 2 亿美元	超 15 亿美元
2023 年 10 月	红杉资本、今日资本、砺思资本	近 20 亿人民币	
2024 年 2 月	阿里、红杉中国、小红书、美团等	超 10 亿美金	约 25 亿美金

资料来源：36Kr，长江证券研究所

Kimi Chat：长文本技术世界领先

Kimi Chat 是一款由月之暗面开发并推出的智能问答助手产品，以对话交互方式为 C 端用户提供服务。Kimi Chat 以长文本处理技术著称，同时具备优秀的联网搜索与指令遵循能力，能够准确识别用户问题并给出完善回答，其支持各类文档（PDF、Excel、CSV、PPT、TXT、图片等）解析，主要应用场景包括文本提炼、信息检索、数据处理、模拟交流等。对标当前主流大模型 GPT4，在学术论文总结、人物观点搜索、PDF 文件解析、等任务环节，Kimi chat 表现与 GPT4 接近，显示出较好的文本处理能力与交互效果。

图 8: Kimi chat VS GPT4: 搜索总结

Kimi Chat	GPT4
<p>瑞·达利欧是桥水基金的创始人，他在人生和投资方面提出了一系列的核心原则。以下是根据搜索结果总结的一些核心原则：</p> <p>人生原则：</p> <ol style="list-style-type: none"> 人生阶段划分： <ul style="list-style-type: none"> 学习阶段：依赖他人，像幼儿一样。 工作成长阶段：他人依赖你，希望你能成功。 成熟阶段：不再追求个人更大的成功，而是希望身边的人能成功。 三个人生原则： <ul style="list-style-type: none"> 保持天性：了解自己天生的驱动力和激励因素。 打造生活机器：确定方向后，通过五步流程（选择目标、识别问题、诊断问题根源、制定计划、执行计划）实现目标。 利用坦诚透明的力量：在沟通中保持真实，直接表达想法。 其他生活原则： <p>投资原则：</p> <ol style="list-style-type: none"> 五大交易原则： <ul style="list-style-type: none"> 分散化投资：成功的投资关键是要打造良好的投资组合，找到15个良好的、互不相关的回报流。 理论价值与实际价值：理解预期价值（未来现金流的现值）和实际价值（支出/销售量）。 最佳资产类别：没有固定的最佳资产类别，随市场局势变化。 系统化决策：设定投资方法与思路，遵循原则，通过投资博弈进行验证。 逆向思维：投资要逆千万人而为，逆自己的思维方式而行。 投资“圣杯”： 打造良好的投资组合，找到10-15个良好的、互不相关的投资或回报流。 其他投资理念： <ul style="list-style-type: none"> 痛苦+反省=进步。 当真相与愿望不符时，选择接受真相。 区分目标和欲望。 有原则的决策。 谦逊和自知之明。 	<p>聊天GPT</p> <p>瑞达利欧（Ray Dalio）是全球最大的对冲基金桥水基金（Bridgewater Associates）的创始人之一。他不仅重塑投资策略而闻名，也重塑关于工作和生活的核心原则而受到广泛关注瑞达利欧在他的著作《原则》（Principles）中详细阐述了这些观点，分为生活原则和工作原则两大部分，他认为这些原则帮助他在个人生活和职业生涯中取得成功以下是他的一些核心原则：</p> <p>人生原则</p> <ol style="list-style-type: none"> 拥抱现实并面对现实： 问题，诚实地面对自己和他人，接受困难的真相。 痛苦+遗憾=进步： 瑞达利欧认为痛苦是进步的催化剂，通过反思痛苦的原因，我们可以找到增长和改善的办法。 目标、障碍、解决方案： 明确你的目标，识别阻碍你实现目标的障碍，然后直接解决这些障碍。 相信进化过程： 生活和工作都是一个不断进化的过程，通过不断学习和适应，我们能够进化成更好的自己。 <p>投资原则</p> <ol style="list-style-type: none"> 分散投资组合： 不要把所有的鸡蛋放在一个篮子里，分散投资可以降低风险。 风险管理： 识别和管理风险是投资成功的关键，要努力做到风险最小化才能获得最大回报。 有效市场假说： 相信市场总体上是有效的，单一的投资者很难持续击败市场。 反周期投资： 鼓励勇气在市场恐慌时买入，在市场狂热时卖出，逆市场情绪而行。 <p>瑞达利欧认为，这些原则适用于投资，它们可以评估生活的方方面面。通过持续地反思和改善，个人可以在生活和职业中实现目标。这些原则的核心在于不断地学习、适应并优化决策流程，以实现持续进步和成功。</p>

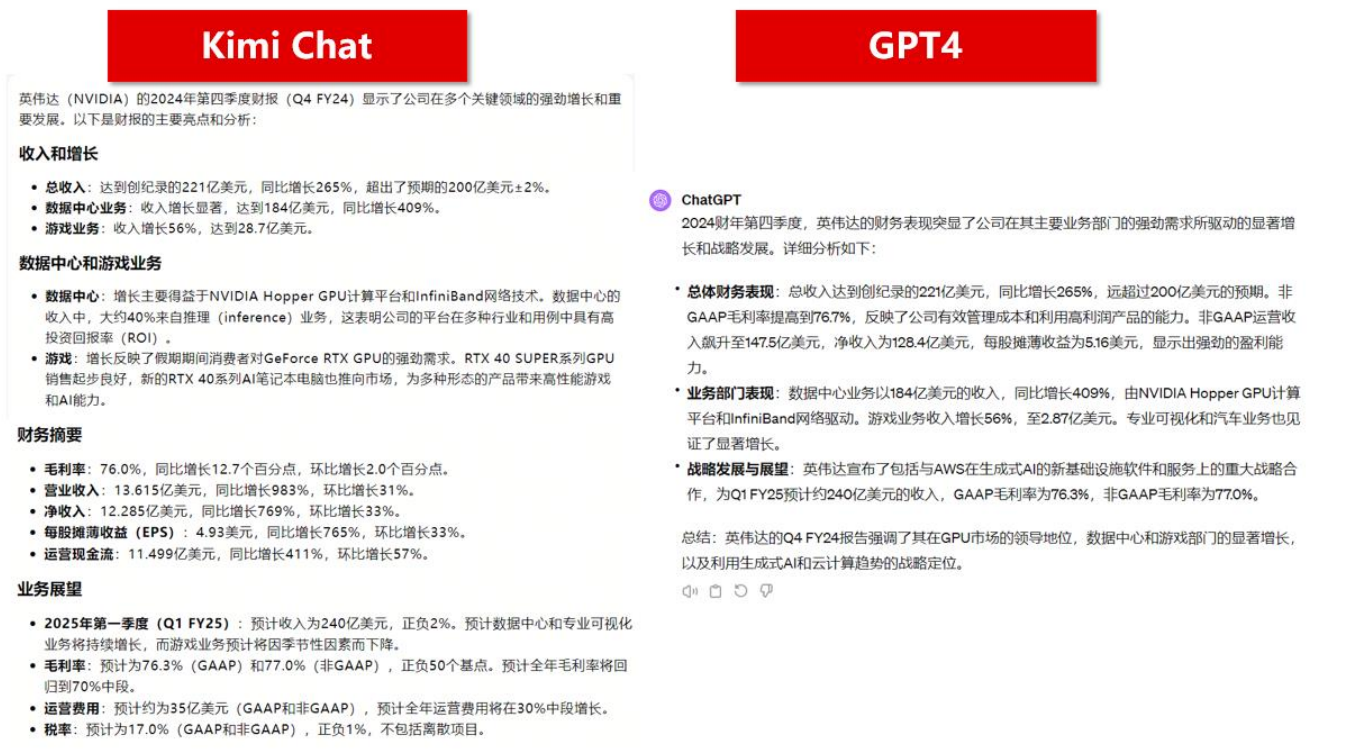
资料来源：Moonshot 官网，Open AI，长江证券研究所

图 9: Kimi chat VS GPT4: 论文总结

Kimi Chat	GPT4
<p>这篇论文介绍了一种名为AutoFT (Robust Fine-Tuning by Optimizing Hyperparameters on OOD Data) 的新方法。旨在通过优化超参数来提高模型在面对分布偏移 (out-of-distribution, OOD) 数据时的泛化能力。以下是从这篇论文中学到的关键新想法和新知识：</p> <ol style="list-style-type: none"> 数据驱动的微调方法： AutoFT提出了一种数据驱动的方法来指导基础模型 (foundation models) 的微调过程。这种方法通过优化超参数来最大化模型在一个小的OOD验证集上的性能，从而提高模型在未知OOD数据上的表现。 超参数空间的扩展： AutoFT扩展了传统的超参数空间，不仅包括学习率和权重衰减等常规超参数，还包括了多个不同损失函数的权重系数。这样的设计允许模型在微调过程中更细致地调整其学习目标。 OOD验证集的使用： 与传统方法不同，AutoFT使用与训练数据分布不同的小型OOD验证集来优化超参数。这使得模型能够在微调过程中学习到如何在保留基础模型先验知识的同时适应特定任务的数据。 实验评估： 作者在多个自然分布偏移的场景下评估了AutoFT，包括域偏移和子群体偏移。实验结果表明，AutoFT在这些场景下显著提高了模型对新OOD数据的泛化能力，并且在WILDs-iWildCam和WILDs-FMoW基准测试中取得了新的最佳性能。 计算效率： 尽管AutoFT引入了更复杂的超参数搜索空间，但其计算开销相对较小，仅比标准微调多出5%的总计算量。 鲁棒性与性能的提升： AutoFT在保持模型在原始分布 (in-distribution, ID) 上的性能的同时，显著提高了在OOD数据上的表现。这表明AutoFT能够在不牺牲ID性能的情况下，提高模型的OOD鲁棒性。 超参数的可转移性： 论文还探讨了通过AutoFT学习到的超参数在不同微调数据集和模型架构之间的可转移性，发现这些超参数在与原始微调分布相似的数据集上表现更好。 <p>这些新的想法和知识为机器学习领域，特别是在提高模型泛化能力和鲁棒性方面提供了有价值的见解，并可能激发未来在这一方向上的进一步研究。</p>	<p>You</p> <p>https://arxiv.org/html/2401.10220v1 请用中文回答：我们可以从这篇paper中学到哪些关键的新想法和新知识？</p> <p>ChatGPT</p> <p>这篇论文提出了AutoFT，一种优化超参数以在离群分布 (OOD) 数据上强化微调基础模型的方法。通过在小型OOD验证集上优化超参数，AutoFT显著提高了对新OOD数据的泛化能力。其关键创新在于通过搜索广泛的超参数空间（包括多种损失函数和正则化器）来精细控制微调过程，这种方法在诸如WILDs-iWildCam和WILDs-FMoW等基准测试上实现了最佳性能。详情请访问论文链接。</p>

资料来源：Moonshot 官网，Open AI，长江证券研究所

图 10: Kimi chat VS GPT4: 财报解析



资料来源: Moonshot 官网, Open AI, 长江证券研究所

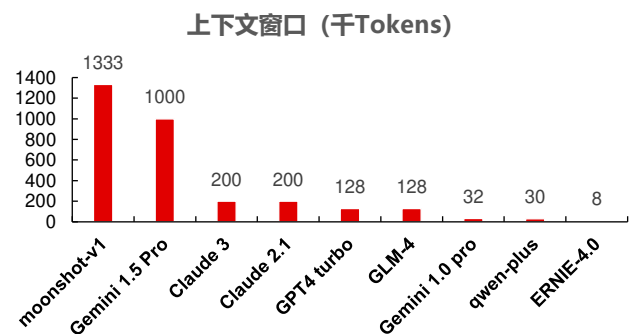
技术迭代助力性能跃升, Kimi Chat 长文本技术世界领先。月之暗面创始团队成立之初便专注于无损长上下文窗口技术, 凭借快速研发实现产品效果的领先。Kimi Chat 最早于2023年10月正式开放内测, 推出伊始具有长达20万字的上下文处理能力, 领先于当时其他可产品化使用的大模型, 至今产品已经历4轮迭代。2024年3月18日, Moonshot AI 官方公众号宣布启动200万字无损上下文内测, 是半年前长文本处理能力的10倍。根据月之暗面官方API文档, 一段通常的中文文本中Token与字数关系为1Token等于1.5-2个汉字, 按1.5汉字计算200万字对应约133万Tokens, 显示出Kimi Chat当前的长文本处理能力已超越Gemini 1.5 Pro (100万Tokens)、Claude 3 (20万Tokens) 以及GPT-4 Turbo (12.8万Tokens), 为全球领先水平。

图 11: Kimi chat 上市至今已经历4轮迭代



资料来源: Moonshot 官方公众号, 长江证券研究所

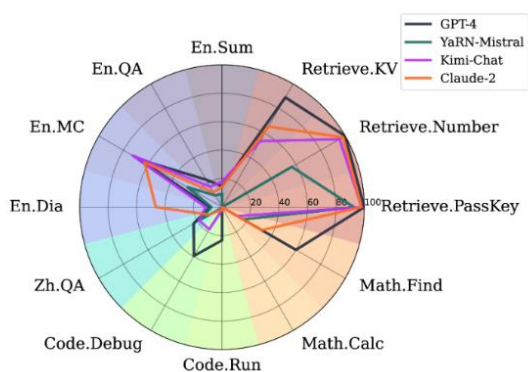
图 12: 按200万字测算, Kimi chat 长文本处理能力或领先全球



资料来源: 各公司官网, 长江证券研究所 (注: moonshot-v1 最大上下文窗口假设为1Token=1.5字)

对比其他国内外大模型，Kimi chat 兼具性能与性价比。性能方面，根据相关论文评估，Kimi chat 在生成长文本摘要、多选问答上要优于其他模型，而在代码与数学的能力上表现相对一般。定价方面，上下文处理能力最强的 moonshot-v1-128k 每百万 tokens 的计价 60 元，低于相同文本处理水平下的其他模型（GLM4 与 GPT4 Turbo）。

图 13: 对比其他大模型，Kimi Chat 文本处理能力接近 GPT4



资料来源:《∞ Bench: Extending Long Context Evaluation Beyond 100K Tokens》(Zhang et al, 2024), 长江证券研究所

图 14: Kimi Chat 在长文本摘要、多选问答两个环节得分领先

Task	GPT-4	YaRN-Mistral	Kimi-Chat	Claude 2
Retrieve.PassKey	100.00	92.71	98.14	97.80
Retrieve.Number	100.00	56.61	95.42	98.14
Retrieve.KV	89.00	0.00	53.60	65.40
En.Sum	14.73	9.09	17.93	14.45
En.QA	22.22	9.55	16.52	11.97
En.MC	67.25	27.95	72.49	62.88
En.Dia	8.50	7.50	11.50	46.50
Zh.QA	23.06	16.98	18.62	10.53
Code.Debug	39.59	0.76	18.02	2.28
Code.Run	23.25	1.25	2.00	2.50
Math.Calc	0.01	0.00	0.00	0.00
Math.Find	60.00	17.14	12.57	32.29
Average	45.63	19.96	34.73	37.06

资料来源:《∞ Bench: Extending Long Context Evaluation Beyond 100K Tokens》(Zhang et al, 2024), 长江证券研究所

表 4: 主要大模型定价及上下文窗口对比

模型	提供商	区域	输入价格 (元/M tokens)	输出价格 (元/M tokens)	上下文窗口
Claude 3 Opus	Anthropic	国外	107	533	200K
GPT4 turbo	OPENAI	国外	71	213	128K
Claude 3 Sonnet	Anthropic	国外	21	107	200K
GLM-4	智谱 AI	国内	100	100	128K
moonshot-v1-128k	Moonshot	国内	60	60	128K
qwen-plus	阿里巴巴	国内	20	20	30K
文心-4.0-8K	百度	国内	12	12	8K
混元-标准版	腾讯	国内	10	10	-
Claude 3 Haiku	Anthropic	国外	2	9	200K
Gemini 1.0 Pro	Google	国外	1	3	-

资料来源: Wind, Anthropic, Open AI, Moonshot, 阿里巴巴, 百度, 腾讯, Google, 智谱 AI, 长江证券研究所 (注: 美元与人民币换算为 7.10, 参考美元兑人民币中间价 2024 年至今均价)

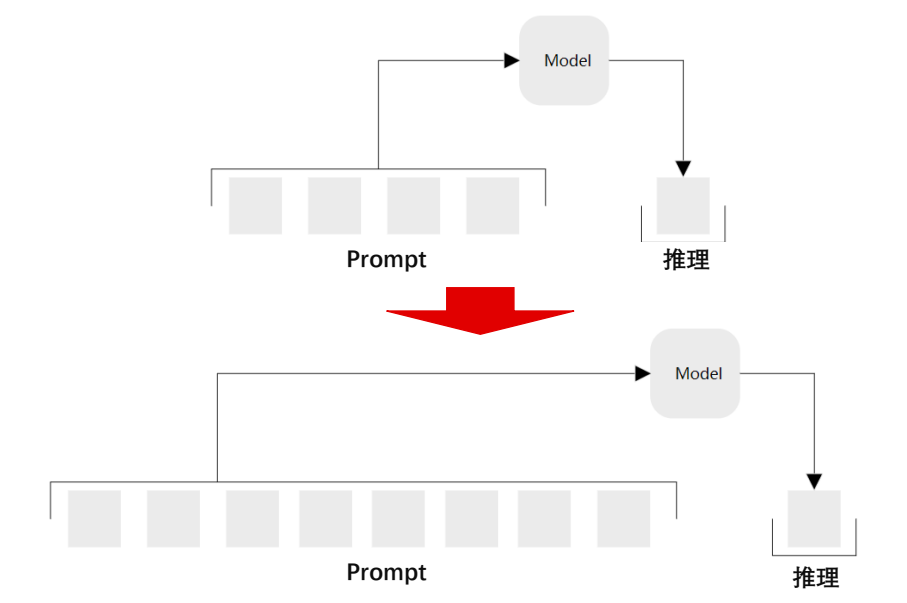
如何看待 KIMI 的长文本能力跃升?

KIMI 长文本能力特质体现了我国基础模型能力的跃升。大模型的上下文窗口通常是由模型开发团队根据模型的架构和预期的应用场景来设定的。在处理长文本时，由于计算资源和性能的限制，模型无法处理无限长的序列。因此，开发团队会在确保模型性能的前提下设定一个最大上下文窗口大小，窗口大小决定了模型在一次前向传播中能够处理

的最大文本长度。本次 KIMI 实现了长文本能力的突破代表了公司在模型架构、内存管理、注意力机制优化等层面有深厚的技术积累。

KIMI 的长文本能力使得模型能够更好的理解 Prompt 输入。KIMI 模型的长文本能力源自于公司的 lossless long-context 机制，lossless long-context 机制在训练阶段以无损压缩的形式实现模型的“scale up”，因此在推理阶段可以较好还原训练数据的细节。此外，长文本能力使得模型在推理阶段能够结合丰富的 Prompt 内容，避免“断章取义”，使得推理更加准确。后续随着模型持续迭代，模型长文本能力有望持续提升。从远期来看，模型调参步骤可能不再重要，用户或可以凭借长 Prompt 输入直接实现模型调参的过程。

图 15: 长文本能力使得 KIMI 能依据更长的 Prompt 进行推理



资料来源：长江证券研究所

KIMI 的 lossless long-context 机制或可以取代模型调参，催生个性化模型的诞生。之前的大语言模型数据主要来自于训练阶段，受推理阶段的用户数据的影响较小，因此无法建立起基于用户数据的持续进化。而 KIMI 模型提升了推理阶段用户数据的权重，模型可以根据用户的行为进行反馈，因此建立了数据飞轮效应。从长期角度来看，lossless long-context 机制在未来有望取代模型的调参步骤，用户可以根据与模型长期互动建立个性化的模型。

KIMI 模型长文本能力或适配 B 端重应用场景。在这一轮迭代后，KIMI 模型能够支持输入高达 200 万汉字的文本，在国内通用大模型领域处于技术领先地位。KIMI 模型的长文本处理能力使其能够直接处理大型文件 Prompt，适用于复杂和需要大量文本分析的 B 端重应用场景，如办公、医疗、教育等垂类场景。相比 C 端轻应用场景，重应用场景用户留存率较高，付费意愿更为强烈。因此长文本能力或成为 KIMI 等大模型货币化之路的核心要点之一。

图 16: KIMI-Chat 鼓励用户尝试超长文本的任务



资料来源: KIMI 官网, 长江证券研究所

图 17: KIMI 在长文本任务中有较好的表现



资料来源: KIMI 官网, 长江证券研究所

风险提示

- 1、AI 模型技术推进不及预期。由于模型技术当前还处于快速发展中，新技术的不成熟和不稳定性，可能会导致研发进展缓慢的推迟、性能表现不佳、安全风险等问题，这些问题可能会给企业带来不利影响。
- 2、AI 模型下游需求不及预期。目前 AI 模型处于技术迭代状态，模型性能在部分场景可用性不足。若模型能力在未来无法持续提升，下游需求或不及预期。

投资评级说明

行业评级 报告发布日后的 12 个月内行业股票指数的涨跌幅相对同期相关证券市场代表性指数的涨跌幅为基准，投资建议的评级标准为：

看 好： 相对表现优于同期相关证券市场代表性指数

中 性： 相对表现与同期相关证券市场代表性指数持平

看 淡： 相对表现弱于同期相关证券市场代表性指数

公司评级 报告发布日后的 12 个月内公司的涨跌幅相对同期相关证券市场代表性指数的涨跌幅为基准，投资建议的评级标准为：

买 入： 相对同期相关证券市场代表性指数涨幅大于 10%

增 持： 相对同期相关证券市场代表性指数涨幅在 5%~10%之间

中 性： 相对同期相关证券市场代表性指数涨幅在-5%~5%之间

减 持： 相对同期相关证券市场代表性指数涨幅小于-5%

无投资评级： 由于我们无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使我们无法给出明确的投资评级。

相关证券市场代表性指数说明：A 股市场以沪深 300 指数为基准；新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准。

办公地址

上海

Add /浦东新区世纪大道 1198 号世纪汇广场一座 29 层
P.C / (200122)

武汉

Add /武汉市江汉区淮海路 88 号长江证券大厦 37 楼
P.C / (430015)

北京

Add /西城区金融街 33 号通泰大厦 15 层
P.C / (100032)

深圳

Add /深圳市福田区中心四路 1 号嘉里建设广场 3 期 36 楼
P.C / (518048)

分析师声明

本报告署名分析师以勤勉的职业态度，独立、客观地出具本报告。分析逻辑基于作者的职业理解，本报告清晰地反映了作者的研究观点。作者所得报酬的任何部分不曾与，不与，也不将与本报告中的具体推荐意见或观点而有直接或间接联系，特此声明。

法律主体声明

本报告由长江证券股份有限公司及其附属机构（以下简称「长江证券」或「本公司」）制作，由长江证券股份有限公司在中华人民共和国大陆地区发行。长江证券股份有限公司具有中国证监会许可的投资咨询业务资格，经营证券业务许可证编号为：10060000。本报告署名分析师所持中国证券业协会授予的证券投资咨询执业资格证书编号已披露在报告首页的作者姓名旁。

在遵守适用的法律法规情况下，本报告亦可能由长江证券经纪（香港）有限公司在香港地区发行。长江证券经纪（香港）有限公司具有香港证券及期货事务监察委员会核准的“就证券提供意见”业务资格（第四类牌照的受监管活动），中央编号为：AXY608。本报告作者所持香港证监会牌照的中央编号已披露在报告首页的作者姓名旁。

其他声明

本报告并非针对或意图发送、发布给在当地法律或监管规则下不允许该报告发送、发布的人员。本公司不会因接收人收到本报告而视其为客户。本报告的信息均来源于公开资料，本公司对这些信息的准确性和完整性不作任何保证，也不保证所包含信息和建议不发生任何变更。本报告内容的全部或部分均不构成投资建议。本报告所包含的观点、建议并未考虑报告接收人在财务状况、投资目的、风险偏好等方面的具体情况，报告接收者应当独立评估本报告所含信息，基于自身投资目标、需求、市场机会、风险及其他因素自主做出决策并自行承担投资风险。本公司已力求报告内容的客观、公正，但文中的观点、结论和建议仅供参考，不包含作者对证券价格涨跌或市场走势的确定性判断。报告中的信息或意见并不构成所述证券的买卖出价或征价，投资者据此做出的任何投资决策与本公司和作者无关。本研究报告并不构成本公司对购入、购买或认购证券的邀请或要约。本公司有可能会与本报告涉及的公司进行投资银行业务或投资服务等其他业务（例如：配售代理、牵头经办人、保荐人、承销商或自营投资）。

本报告所包含的观点及建议不适用于所有投资者，且并未考虑个别客户的特殊情况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。投资者不应以本报告取代其独立判断或仅依据本报告做出决策，并在需要时咨询专业意见。

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、价值及投资收入可升可跌，过往表现不应作为日后的表现依据；在不同时期，本公司可以发出其他与本报告所载信息不一致及有不同结论的报告；本报告所反映研究人员的不同观点、见解及分析方法，并不代表本公司或其他附属机构的立场；本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。本公司及作者在自身所知情形范围内，与本报告中所评价或推荐的证券不存在法律法规要求披露或采取限制、静默措施的利益冲突。

本报告版权仅为本公司所有，本报告仅供意向收件人使用。未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布给其他机构及/或人士（无论整份和部分）。如引用须注明出处为本公司研究所，且不得对本报告进行有悖原意的引用、删节和修改。刊载或者转发本证券研究报告或者摘要的，应当注明本报告的发布人和发布日期，提示使用证券研究报告的风险。本公司不为转发人及/或其客户因使用本报告或报告载明的内容产生的直接或间接损失承担任何责任。未经授权刊载或者转发本报告的，本公司将保留向其追究法律责任的权利。

本公司保留一切权利。