谷歌再更新 Gemini 大模型，立足 MoE 架构性能更加卓越

——计算机行业跟踪报告

强于大市(维持)

2024 年 02 月 21 日

行业核心观点：

谷歌推出 Gemini 1.5 Pro 版本，性能水平与 Gemini 1.0 Ultra 类似。2024 年 2 月 15 日，谷歌再次更新其 Gemini 大模型至 Gemini 1.5 代，并推出 Gemini 1.5 Pro 版本。Gemini 1.5 Pro 的性能水平与谷歌至今为止最大的模型 Gemini 1.0 Ultra 类似。与 Gemini 1.0 代对比，Gemini 1.5 Pro 的性能大大超过了 Gemini 1.0 Pro，在绝大多数（27/31）的基准测试（benchmarks）中表现更好；而在与 Gemini 1.0 Ultra 的对比中，Gemini 1.5 Pro 在超过一半的基准测试上表现更好，尤其是在多数文本基准测试（10/13）和部分视觉基准测试（6/13）中都表现优于 Gemini 1.0 Ultra。

投资要点：

建立在 MoE 架构上，能更高效的训练和服务。Gemini 1.5 大模型建立在对稀疏(sparse)混合专家(mixture-of-expert,MoE)架构及 Transformer 架构领先的研究上，其训练和服务更为高效。传统的 Transformer 是一个大型神经网络，而 MoE 模型则被划分为更小的“专家”神经网络。混合专家模型（MoE）主要由两个关键部分组成：1）稀疏 MoE 层：这些层代替了传统 Transformer 模型中的前馈网络(FFN)层。MoE 层包含若干“专家”，每个“专家”本身是一个独立的神经网络；2）门控网络或路由：这个部分用于决定每个 token 被发送到哪个“专家”。这种“术业有专攻”的架构，能够极大的提高模型的效率，让 MoE 能够在远少于稠密模型所需的计算资源下进行有效的预训练，因此基于 MoE 架构的 Gemini 1.5 在训练和服务上也更为高效。

具备超大容量的上下文窗口，可对大量信息进行复杂推理。Gemini 1.5 Pro 是一种中等规模（mid-size）的多模态模型，引入了在上下文理解方面的突破性实验特征。Gemini 1.5 Pro 除了配有标准的 128,000token 的上下文窗口，少数开发人员和企业客户还可以通过 AI Studio 和 Vertex AI 的私人预览版在最多 1,000,000 个 token 的上下文窗口中进行尝试和体验。100 万个 token 的上下文窗口容量相当于 Gemini 1.5 Pro 可以一次性处理 1 小时视频/11 小时音频/超过 30,000 行代码/超过 700,000 个单词（word）的信息库，能够对大量的信息进行复杂推理。

投资建议：Gemini 1.5 Pro 的超大容量上下文窗口有助于其应用在更多的领域。同时，MoE 架构能让模型更高效的训练和服务，也有助于多模型大模型在应用端的加速落地。我们认为 MoE 架构有望成为多模态大模型的主流应用架构之一，建议关注超大容量上下文长度以及 MoE 架构助力多模态大模型在应用端加速落地带来的投资机遇，同时继续关注多模态大模型对算力的持续需求。

风险因素：人工智能产业发展不及预期，应用落地不及预期，AI 带来的隐私、版权与技术风险。

行业相对沪深 300 指数表现

数据来源：聚源，万联证券研究所

相关研究

利润端整体承压，过半标的呈现向好趋势

Gemini1.5 和 Sora 相继发布，关注 AIGC 应用落地及对算力的需求提振

OpenAI 推出首个文生视频大模型 Sora，引领 AI 文生视频行业跨越式发展

分析师： 夏清莹
执业证书编号：S0270520050001
电话：075583223620
邮箱：xiaqy1@wlzq.com.cn

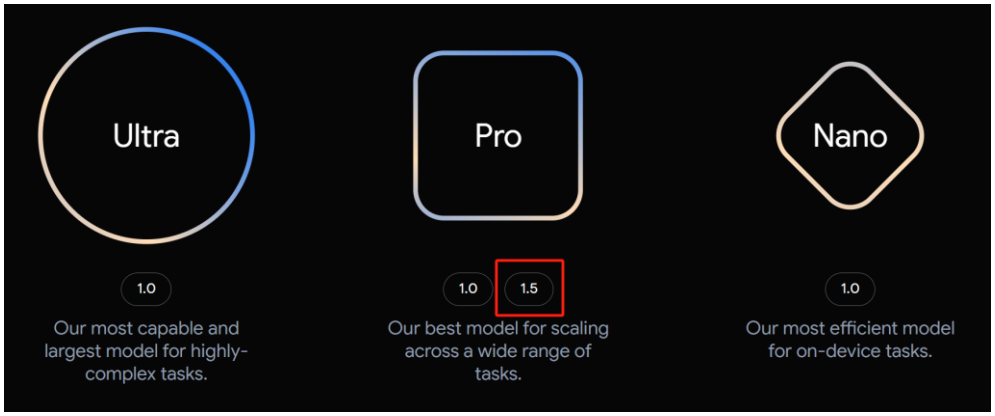
正文目录

1 Gemini 1.5 Pro 发布，立足 MoE 架构，性能更加优越

1.1 谷歌 Gemini 系列再更新，Gemini 1.5 Pro 性能可媲美 Gemini 1.0 Ultra

谷歌多模态大模型再更新，推出Gemini 1.5 Pro版本。Gemini系列大模型是谷歌的多模态（multimodality）大模型，能够处理跨越文本、图片、音频、视频、代码等多模态信息。此前，谷歌推出的Gemini 1.0总共有Nano、Pro、Ultra三个版本。2024年2月15日，谷歌再次更新其Gemini大模型至Gemini 1.5代，并推出Gemini 1.5 Pro版本。

图表1：谷歌 Gemini 系列产品一览



资料来源：Google DeepMind官网，万联证券研究所

Gemini 1.5 Pro的性能水平与谷歌至今为止最大的模型Gemini 1.0 Ultra类似。与Gemini 1.0代对比，Gemini 1.5 Pro的性能大大超过了Gemini 1.0 Pro，在绝大多数（27/31）的基准测试（benchmarks）中表现更好；而在与Gemini 1.0 Ultra的对比中，Gemini 1.5 Pro在超过一半的基准测试上表现更好，尤其是在多数文本基准测试（10/13）和部分视觉基准测试（6/13）中都表现优于Gemini 1.0 Ultra。

图表2：Gemini 1.5 Pro 和 Gemini 1.0 Pro 及 Gemini 1.0 Ultra 的对比

Gemini 1.5 Pro	Relative to 1.0 Pro	Relative to 1.0 Ultra
Long-Context Text, Video & Audio	from 32k up to 10M tokens	from 32k up to 10M tokens
Core Capabilities	Win-rate: 87.1% (27/31 benchmarks)	Win-rate: 54.8% (17/31 benchmarks)
Text	Win-rate: 100% (13/13 benchmarks)	Win-rate: 77% (10/13 benchmarks)
Vision	Win-rate: 77% (10/13 benchmarks)	Win-rate: 46% (6/13 benchmarks)
Audio	Win-rate: 60% (3/5 benchmarks)	Win-rate: 20% (1/5 benchmarks)

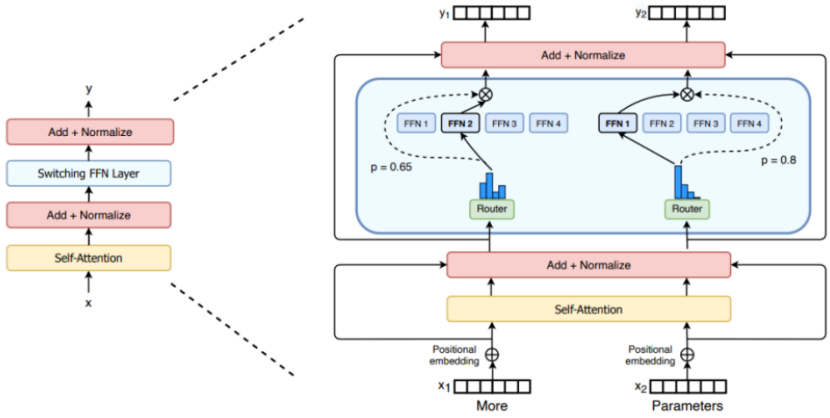
资料来源：Google技术报告，万联证券研究所

1.2 建立在 MoE 架构上，能更高效的训练和服务

Gemini 1.5大模型建立在对稀疏()混合专家（mixture-of-expert，MoE）架构及Transformer架构领先的研究上，其训练和服务更为高效。传统的Transformer是一个大型神经网络，而MoE模型则被划分为更小的“专家”神经网络。混合专家模型（MoE）主要由两个关键部分组成：1）稀疏MoE层：这些层代替了传统Transformer模型

中的前馈网络(FFN)层。MoE层包含若干“专家”，每个“专家”本身是一个独立的神经网络;2）门控网络或路由:这个部分用于决定每个token被发送到哪个“专家”。例如，在下图中，“More”这个token被发送到第二个专家，而“Parameters”这个token被发送到第一个专家。这种“术业有专攻”的架构，能够极大的提高模型的效率，让MoE能够在远少于稠密模型所需的计算资源下进行有效的预训练,基于MoE架构的Gemini 1.5在训练和服务上也更为高效。

图表3：MoE 架构原理示意图

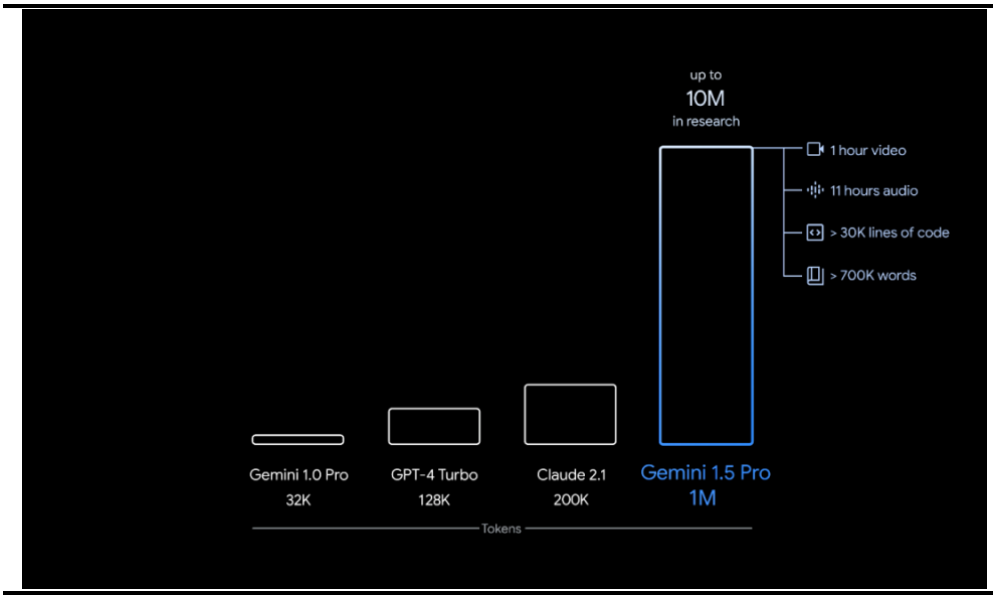


资料来源：Hugging Face、Switch Transformers paper，万联证券研究所

1.3 具备超大容量的上下文窗口，可对大量信息进行复杂推理

Gemini 1.5 Pro是中等规模的多模态模型，具有超大容量的上下文窗口。谷歌现在推出的Gemini 1.5 Pro是一种中等规模（mid-size）的多模态模型，引入了在上下文理解方面的突破性实验特征。Gemini 1.5 Pro除了配有标准的128,000token的上下文窗口，少数开发人员和企业客户还可以通过AI Studio和Vertex AI的私人预览版在最多1,000,000个token的上下文窗口中进行尝试和体验。**100万个token的上下文窗口容量相当于Gemini 1.5 Pro可以一次性处理1小时视频/11小时音频/超过30,000行代码/超过700,000个单词（word）的信息库。**

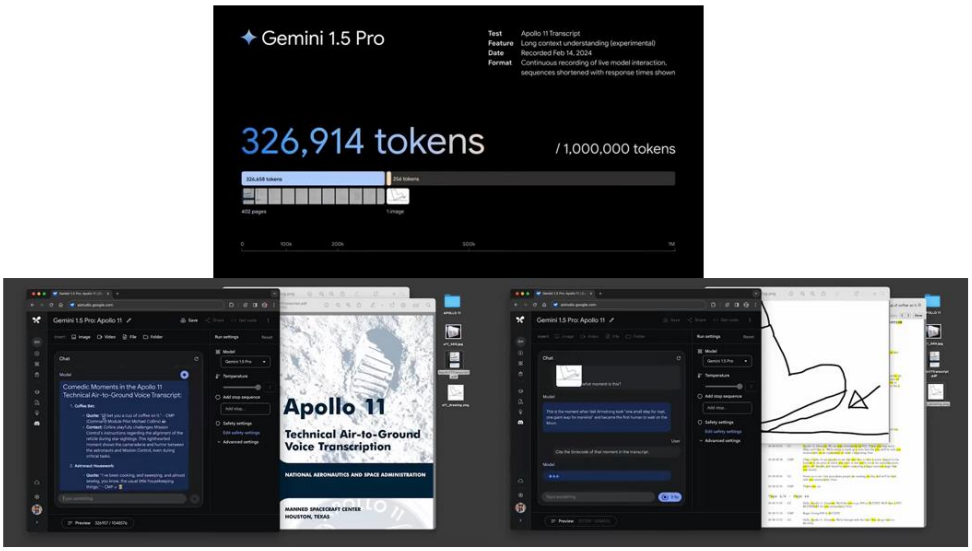
图表4：Gemini 1.5 具有超大容量的上下文窗口



资料来源：机器之心、腾讯网，万联证券研究所

Gemini 1.5 Pro能够对大量的信息进行复杂推理，可以在给定提示内无缝分析、分类和总结大量内容。例如，当给出阿波罗11号登月任务的402页记录时，Gemini 1.5 Pro可以推理文档中的对话、事件和细节。

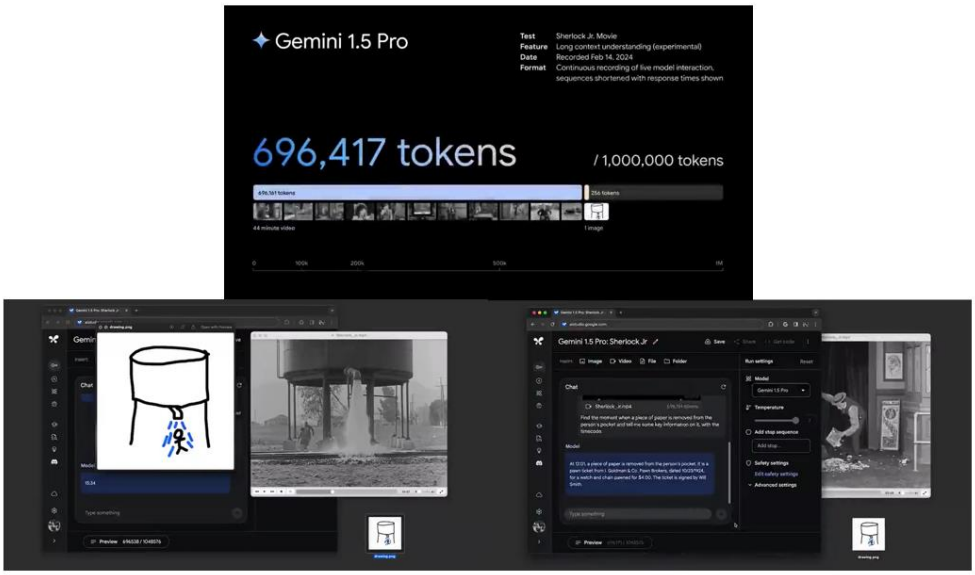
图表5：Gemini 1.5 Pro 可以处理阿波罗 11 号登月任务 402 页的记录



资料来源：机器之心、腾讯网，万联证券研究所

Gemini 1.5 Pro能够更好地理解和推理跨模态，可以针对包括视频在内的不同模式执行高度复杂的理解和推理任务。例如，当给定一部44分钟的巴斯特·基顿无声电影时，该模型可以准确分析各种情节点和事件，甚至推理出电影中容易被忽略的小细节。当给出简单的线条图作为现实生活中物体的参考材料时，Gemini 1.5 Pro可以识别44分钟的巴斯特基顿无声电影中的场景。

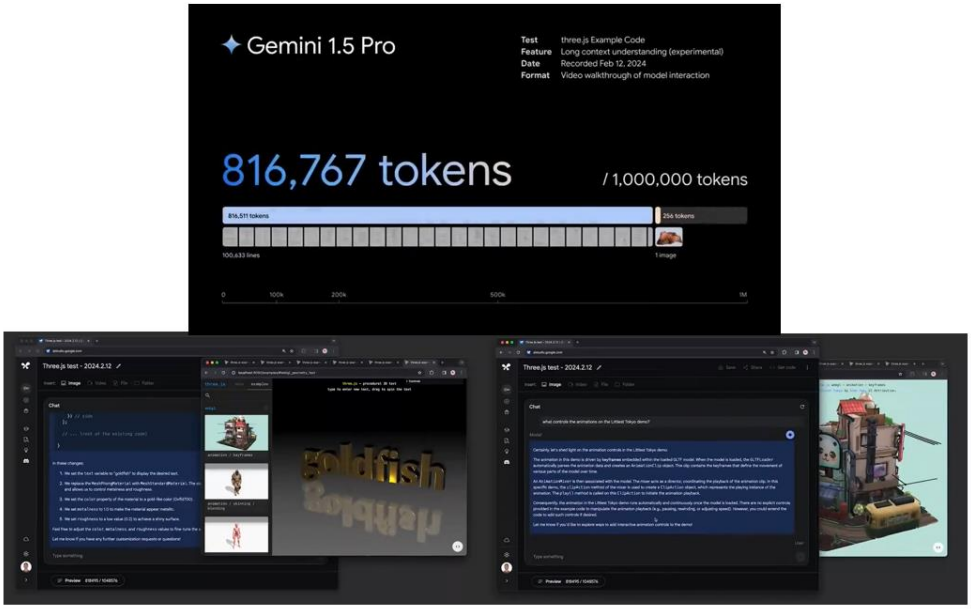
图表6：Gemini 1.5 Pro 可以识别一部 44 分钟无声电影中的场景



资料来源：机器之心、腾讯网，万联证券研究所

Gemini 1.5 Pro能够使用较长的代码块解决相关问题。例如，当给出超过100,000行代码的提示时，它可以更好地推理示例、建议有用的修改并解释代码不同部分的工作原理。

图表7：Gemini 1.5 Pro 可以推理超过 100,000 行代码



资料来源：机器之心、腾讯网，万联证券研究所

2 投资建议

Gemini 1.5 Pro的超大容量上下文窗口有助于其应用在更多的领域。同时，MoE架构能让模型更高效的训练和服务，也有助于多模型大模型在应用端的加速落地。我们认为MoE架构有望成为多模态大模型的主流应用架构之一，建议关注超大容量上下文长度以及MoE架构助力多模态大模型在应用端加速落地带来的投资机遇，同时继续关注多模态大模型对算力的持续需求。

3 风险提示

人工智能产业发展不及预期，应用落地不及预期，AI带来的隐私、版权与技术风险。

行业投资评级

强于大市：未来6个月内行业指数相对大盘涨幅10％以上；
同步大市：未来6个月内行业指数相对大盘涨幅10％至-10％之间；
弱于大市：未来6个月内行业指数相对大盘跌幅10％以上。

公司投资评级

买入：未来6个月内公司相对大盘涨幅15％以上；
增持：未来6个月内公司相对大盘涨幅5％至15％；
观望：未来6个月内公司相对大盘涨幅-5％至5％；
卖出：未来6个月内公司相对大盘跌幅5％以上。
基准指数 ： 沪深300指数

风险提示

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议；投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。

证券分析师承诺

本人具有中国证券业协会授予的证券投资咨询执业资格并登记为证券分析师，以勤勉的执业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

免责条款

万联证券股份有限公司 研究所
上海浦东新区世纪大道 1528 号陆家嘴基金大厦
北京西城区平安里西大街 28 号中海国际中心
深圳福田区深南大道 2007 号金地中心
广州天河区珠江东路 11 号高德置地广场