证券研究报告

事件点评

上海证券 SHANGHAI SECURITIES

谷歌发布开源模型 Gemma，端侧生成式 AI 或现增量需求

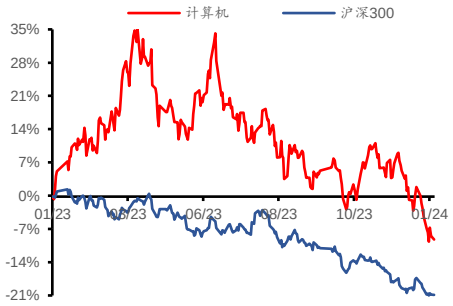
——人工智能行业跟踪报告

增持（维持）

行业： 计算机
日期： 2024年03月04日

分析师： 刘京昭
SAC 编号： S0870523040005

最近一年行业指数与沪深 300 比较



■ 主要观点

2024年2月21日，谷歌发布生成式AI模型Gemma，并在全球范围内开放使用。本次发布的Gemma 2B和Gemma 7B模型，参数量分别为20亿和70亿，每个模型都有对应的预训练和指令微调版本。其中，Gemma 7B模型主要用于在GPU和TPU上进行部署，Gemma 2B模型主要用于在CPU上进行部署，特别是在端侧设备上实现应用。

相比于Meta、Mistral AI等厂商推出的LLaMA-2、Mistral等相似参数量的开源模型，Gemma 7B在多个英文数据集上的表现相对较优。同时，Gemma也有望进一步缩小开源模型与闭源模型间的性能差距。

我们认为：Gemma模型的推出是对端侧生成式AI模型的进一步丰富，Gemma模型能够有效满足个人电脑等端侧设备的生成式AI模型部署需求，进一步挖掘生成式AI在端侧的应用场景。在新应用场景的支持下，模型训练相关的算力基础设施需求有望得到提振，为以光模块为代表的算力供应链带来新的增量空间。

■ 投资建议

建议关注：

中际旭创：中高端数通市场龙头，2022年与II-VI并列光模块业务营收全球第一。根据iFinD机构一致预期，截至2024年2月23日，公司2024年的预测PE为29倍，位于近五年的85%分位。

天孚通信：光器件整体解决方案提供商。根据iFinD机构一致预期，截至2024年2月23日，公司2024年的预测PE为43倍，位于近五年的98%分位。

新易盛：光模块领域龙头，成本管控优秀，具备切入增量云计算/AI客户的能力。根据iFinD机构一致预期，截至2024年2月23日，公司2024年的预测PE为33倍，位于近五年的90%分位。

■ 风险提示

下游需求不及预期；人工智能技术落地和商业化不及预期；产业政策转变；宏观经济不及预期等。

目 录

图

表

1 Gemma 模型支持本地部署，性能优于相似参数规模开源模型

2024 年 2 月 21 日，谷歌发布生成式 AI 模型 Gemma，并在全球范围内开放使用。本次发布的 Gemma 2B 和 Gemma 7B 模型，参数量分别为 20 亿和 70 亿，每个模型都有对应的预训练和指令微调版本。其中，Gemma 7B 模型主要用于在 GPU 和 TPU 上进行部署，Gemma 2B 模型主要用于在 CPU 上进行部署，特别是在端侧设备上实现应用。

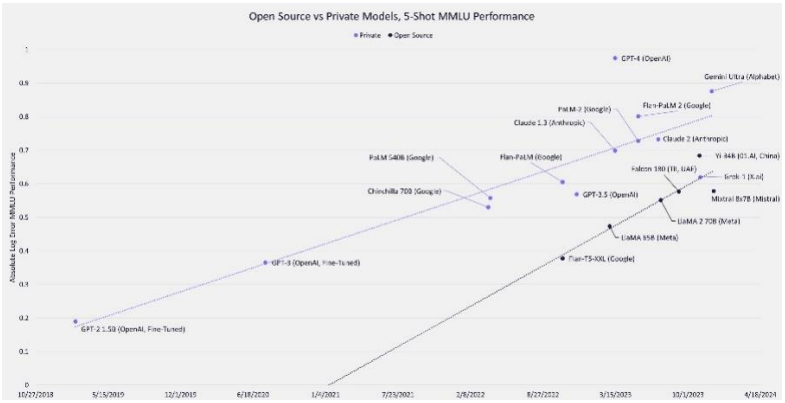
图 1：Gemma 7B 的测试表现优于相似参数量的其他开源模型

		LLaMA-2		Mistral	Gemma	
Benchmark	metric	7B	13B	7B	2B	7B
MMLU	5-shot, top-1	45.3	54.8	62.5	42.3	**64.3**
HellaSwag	0-shot	77.2	80.7	81.0	71.4	**81.2**
PIQA	0-shot	78.8	80.5	**82.2**	77.3	81.2
SIQA	0-shot	48.3	50.3	47.0*	49.7	**51.8**
Boolq	0-shot	77.4	81.7	**83.2***	69.4	**83.2**
Winogrande	partial scoring	69.2	72.8	**74.2**	65.4	72.3
CQA	7-shot	57.8	67.3	66.3*	65.3	**71.3**
OBQA		**58.6**	57.0	52.2	47.8	52.8
ARC-e		75.2	77.3	80.5	73.2	**81.5**
ARC-c		45.9	49.4	**54.9**	42.1	53.2
TriviaQA	5-shot	72.1	**79.6**	62.5	53.2	63.4
NQ	5-shot	25.7	**31.2**	23.2	12.5	23.0
HumanEval	pass@1	12.8	18.3	26.2	22.0	**32.3**
MBPP†	3-shot	20.8	30.6	40.2*	29.2	**44.4**
GSM8K	maj@1	14.6	28.7	35.4*	17.7	**46.4**
MATH	4-shot	2.5	3.9	12.7	11.8	**24.3**
AGIEval		29.3	39.1	41.2*	24.2	**41.7**
BBH		32.6	39.4	**56.1***	35.2	55.1
Average		47.0	52.2	54.0	44.9	**56.4**

资料来源：Google，上海证券研究所

相比于 Meta、Mistral AI 等厂商推出的 LLaMA-2、Mistral 等相似参数量的开源模型，Gemma 7B 在多个英文数据集上的表现相对较优。同时，Gemma 也有望进一步缩小开源模型与闭源模型间的性能差距。

图 2：Gemma 有望进一步缩小开源模型与闭源模型间的性能差距



资料来源：ARK Invest，36 氪，上海证券研究所

值得关注的是，本次 Gemma 模型训练使用谷歌于 2023 年 8 月推出的 TPU v5e 芯片进行训练。TPU v5e 集群由 Pod 构成，其中，256 块 TPU v5e 芯片组成一个 Pod。谷歌分别使用 2 个和 16 个 Pod 对 Gemma 2B 和 Gemma 7B 模型进行预训练。

图 3：Gemma 能够在本地通过 API 调用

```
from transformers import AutoTokenizer, pipeline
import torch

model = "google/gemma-7b-it"

tokenizer = AutoTokenizer.from_pretrained(model)
pipeline = pipeline(
    "text-generation",
    model=model,
    model_kwargs={"torch_dtype": torch.bfloat16},
    device="cuda",
)

messages = [
        {"role": "user", "content": "Who are you? Please, answer in pirate-speak."},
]
prompt = pipeline.tokenizer.apply_chat_template(messages, tokenize=False, add_generation_prompt=True)
outputs = pipeline(
    prompt,
    max_new_tokens=256,
    do_sample=True,
    temperature=0.7,
    top_k=50,
    top_p=0.95
)
print(outputs[0]["generated_text"][len(prompt):])
```

资料来源：GitHub，Hugging Face，上海证券研究所

Gemma 模型目前在开源社区 Hugging Face 平台上进行开放。经过指令微调后，Gemma 7B 模型的 GGUF 文件大小为 5.88GB，Gemma 2B 的 GGUF 文件大小为 1.36GB，完全满足本地部署和运行的需要。

我们认为：Gemma 模型的推出是对端侧生成式 AI 模型的进一步丰富，Gemma 模型能够有效满足个人电脑等端侧设备的生成式 AI 模型部署需求，进一步挖掘生成式 AI 在端侧的应用场景。在新应用场景的支持下，模型训练相关的算力基础设施需求有望得到提振，为以光模块为代表的算力供应链带来新的增量空间。

表 1：人工智能领域相关公司对比

所属板块	股票代码	股票简称	22 营业收入	22 归母净利润	24E 营业收入	24E 归母净利润	24E 估值	近五年 PE 分位数（%）
算力	688041.SH	海光信息	51.25	8.04	84.37	16.70	112	81
	688256.SH	寒武纪	7.29	-12.57	15.38	-5.98	--	--
	300474.SZ	景嘉微	11.54	2.89	17.39	3.60	79	85
	688521.SH	芯原股份	26.79	0.74	31.51	0.66	194	--
	603019.SH	中科曙光	130.08	15.44	174.04	24.91	23	25
PCB	002463.SZ	沪电股份	83.36	13.62	110.82	19.66	24	85
光模块/光器件	300308.SZ	中际旭创	96.42	12.24	229.31	40.33	29	85
	300502.SZ	新易盛	33.11	9.04	53.38	12.73	33	90
	300394.SZ	天孚通信	11.96	4.03	30.99	11.52	43	98
光芯片	688498.SH	源杰科技	2.83	1.00	3.36	1.06	105	61
液冷	872808.BJ	曙光数创	5.18	1.17	8.57	2.11	43	90
	002837.SZ	英维克	29.23	2.80	53.88	5.28	27	31
服务器/交换机	301165.SZ	锐捷网络	113.26	5.50	164.23	7.95	24	64
	603118.SH	共进股份	109.74	2.27	115.56	5.03	12	94
	301191.SZ	菲菱科思	23.52	1.95	33.30	2.91	21	80
	601138.SH	工业富联	5118.50	200.73	6499.72	289.86	12	70
	000938.SZ	紫光股份	740.58	21.58	914.29	29.01	18	15
	000628.SZ	高新发展	65.71	1.99	--	--	--	97
	600100.SH	同方股份	237.61	-7.72	--	--	--	--
	000034.SZ	神州数码	1158.80	10.04	1319.04	14.66	13	19
智能座舱	002261.SZ	拓维信息	22.37	-10.13	41.75	3.01	65	--
	300496.SZ	中科创达	54.45	7.69	74.67	10.42	25	1
机器视觉	002920.SZ	德赛西威	149.33	11.84	264.98	21.58	26	3
	002415.SZ	海康威视	831.66	128.37	992.99	168.95	19	38
	002236.SZ	大华股份	305.65	23.24	378.99	42.98	14	58
	688003.SH	天准科技	15.89	1.52	24.79	2.74	21	1
AI+应用	300802.SZ	矩子科技	6.84	1.29	--	--	--	11
	300418.SZ	昆仑万维	47.36	11.53	55.89	9.64	48	93
	688111.SH	金山办公	38.85	11.18	61.51	17.76	63	15
	002230.SZ	科大讯飞	188.20	5.61	256.05	13.40	81	96
	600570.SH	恒生电子	65.02	10.91	94.18	21.90	21	1
	300033.SZ	同花顺	35.59	16.91	46.69	20.65	33	49
	600845.SH	宝信软件	131.50	21.86	195.59	33.12	31	13

资料来源：iFinD，上海证券研究所

**盈利预测来自 iFinD 机构一致预期；仅列举各板块部分标的；估值基于 2024 年 2 月 23 日收盘价；单位：亿元。*

2 风险提示

下游需求不及预期；人工智能技术落地和商业化不及预期；产业政策转变；宏观经济不及预期。

分析师声明

作者具有中国证券业协会授予的证券投资咨询资格或相当的专业胜任能力，以勤勉尽责的职业态度，独立、客观地出具本报告，并保证报告采用的信息均来自合规渠道，力求清晰、准确地反映作者的研究观点，结论不受任何第三方的授意或影响。此外，作者薪酬的任何部分不与本报告中的具体推荐意见或观点直接或间接相关。

公司业务资格说明

本公司具备证券投资咨询业务资格。

投资评级体系与评级定义

股票投资评级：	分析师给出下列评级中的其中一项代表其根据公司基本面及（或）估值预期以报告日起 6 个月内公司股价相对于同期市场基准指数表现的看法。	
	买入	股价表现将强于基准指数 20%以上
	增持	股价表现将强于基准指数 5-20%
	中性	股价表现将介于基准指数±5%之间
	减持	股价表现将弱于基准指数 5%以上
	无评级	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
行业投资评级：	分析师给出下列评级中的其中一项代表其根据行业历史基本面及（或）估值对所研究行业以报告日起 12 个月内的基本面和行业指数相对于同期市场基准指数表现的看法。	
	增持	行业基本面看好，相对表现优于同期基准指数
	中性	行业基本面稳定，相对表现与同期基准指数持平
	减持	行业基本面看淡，相对表现弱于同期基准指数
相关证券市场基准指数说明：A 股市场以沪深 300 指数为基准；港股市场以恒生指数为基准；美股市场以标普 500 或纳斯达克综合指数为基准。		

投资评级说明：

不同证券研究机构采用不同的评级术语及评级标准，投资者应区分不同机构在相同评级名称下的定义差异。本评级体系采用的是相对评级体系。投资者买卖证券的决定取决于个人的实际情况。投资者应阅读整篇报告，以获取比较完整的观点与信息，投资者不应以分析师的投资评级取代个人的分析与判断。

免责声明

本报告仅供上海证券有限责任公司(以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。

在法律许可的情况下，本公司或其关联机构可能会持有报告中涉及的公司所发行的证券或期权并进行交易，也可能为这些公司提供或争取提供多种金融服务。

本报告的信息来源于已公开的资料，本公司对该等信息的准确性、完整性或可靠性不作任何保证。本报告所载的资料、意见和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、价值或投资收入可升可跌。过往表现不应作为日后的表现依据。在不同时期，本公司可发出与本报告所载资料、意见或推测不一致的报告。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。

本报告中的内容和意见仅供参考，并不构成客户私人咨询建议。在任何情况下，本公司、本公司员工或关联机构不承诺投资者一定获利，不与投资者分享投资收益，也不对任何人因使用本报告中的任何内容所引致的任何损失负责，投资者据此做出的任何投资决策与本公司、本公司员工或关联机构无关。

市场有风险，投资需谨慎。投资者不应将本报告作为投资决策的唯一参考因素，也不应当认为本报告可以取代自己的判断。