2023 年 03 月 17 日

“文心一言”发布，国内厂商距离复现 ChatGPT 有多远？

AGI（通用人工智能）专题之二

▶ “文心一言”表现一如预期，不必过度悲观

3 月 16 日“文心一言”发布，从官方 demo 来看，文心一言具备文学创作、商业文案创作、数理逻辑推算、中文理解、多模态生成能力，存在超预期亮点，但事前录屏降低了其演示的真实性，也并未对外直接开放，多因素导致公众反馈不佳。我们认为尽管上下文理解、语义逻辑、多轮对话方面尚有欠缺，“文心一言”展现了足够的文案创作能力，为 B 端降本增效的起始逻辑已经明晰，尽管尚未对公众大范围开放，企业用户已经能够申请内测邀请码，若邀请范围持续扩大，飞轮效应将推进“文心一言”表现改善，且优化空间极大。

▶ 复现 ChatGPT 的难点在哪里？AI 三要件略逊一筹，但差距并非不可逾越

1）算法上，去开源化极大增加了国内科技企业的复现难度，但学术界已有相对成功复现先例，随着人才流动、时间推移和研究进步，大模型性能很可能逐渐趋同。2）数据上，数据集质量、标注细节处理、用户真实交互是关键，尽管优质的中文标注数据集匮乏，使用英文数据进行预训练并不影响中文输出，科技企业能够参考 GPT3 的路径，利用海量用户交互提升数据质量。3）算力上，国内头部科技企业多已完成数据中心建设，能够实现算力资源部分自给，此外算力更接近于自由流通的商品，战略押注意愿明确、现金流充沛的国内厂商有希望弥合算力上的差距。

▶ 工程化和应用分发能力是隐形壁垒，头部厂商优势明显

国产 ChatGPT 的落地在技术准备之外还需要两项核心要素：工程化与分发能力。1）工程化能力，即能够利用更低的成本和更高效的迭代做出先进的大模型应用，制作更高效、廉价、贴合市场的产品，能够同时容纳亿级用户在线。2）充沛的 C 端用户及 B 端应用场景，即更低的分发触达成本、更快速的产品应用迭代。国内厂商完全具备大 DAU 场景下 AI 工程化处理的潜力，且应用分发是国内企业的长项，头部厂商本身已经建起规模及心智壁垒，且商业模式无需再探索，如要落地 AGI 相关应用，获客成本远低于新进入者。

▶ 若“文心一言”对外开放，增量成本仍可控

短期我们关注“文心一言”等产品对科技企业财务状况的影响，将增量成本拆分为训练成本、推理成本及数据标注成本（暂不考虑人力支出及维护费用），测算大模型落地搜索页面后年均增量成本约为 16 亿元。但考虑到国产模型参数量及数据集 token 数量均是未知，且 Azure 云计算价格与实际成本存在偏差，此外实际落地后各项成本均存在优化可能、具体会计处理方式还存在探讨空间，我们判断 10-20 亿元为其增量成本的合理范围（暂不考虑人力支出及维护费用），参

评级及分析师信息

行业评级：推荐

行业走势图

分析师：赵琳
邮箱：zhaolin@hx168.com.cn
SAC NO：S1120520040003

考百度 2022 年经营现金流净额 261.7 亿元，对公司正常经营影响可控。

投资建议

GPT4、Office365（Copilot）对公众的震撼只是前期技术突破后的余韵，而非 AGI 领域想象力的终点，产品的成功会驱动更多学术资源与产业投资的倾斜，人工智能必然成为产业发展长期主线，国产替代具有需求上的紧迫性。依然重点推荐百度（BIDU.US），判断“文心一言”表现符合预期，公司目前仍处于低估区间，尽管技术差距依然存在，短期内研发投入可能上行，我们看好人工智能领域投入对公司业绩及估值的长线提振。同时关注已在视频、营销、阅读等相关细分领域抢跑的重点标的，推荐当虹科技、捷成股份、蓝色光标、风语筑、浙文互联。

风险提示

“文心一言”落地效果不及预期风险；ToB 服务推进节奏不及预期风险；成本大幅增长风险；AI 产品道德及监管风险。

正文目录

1. “文心一言”答卷未知，但国产替代并不遥远 4
1.1. 细究算法、数据、算力三要件，略逊一筹但仍有追平可能 5
1.2. 工程化处理与分发能力是更高的壁垒 10
1.3. 商业化路径已经明晰，搜索场景鲜明契合 12
2. 若“文心一言”成功对公众开放，年化增量成本可控 15
2.1. 训练：前期固定投入较大，莱特定律驱动下成本必然下行 15
2.2. 推理：与用户数量成正比，成本优化路径明确 16
2.3. 数据标注：取决于人力价格，成本量级较低 17
3. 投资建议 18
4. 风险提示 18

图表目录

图 1“文心一言”生成图片 4
图 2 “文心一言”生成视频 4
图 3 2005 年起中国 AI 论文总数超美国 5
图 4 海外 AI 机构预测中国高引论文占比将超过美国 5
图 5 类 ChatGPT 产品的技术发展示意图 5
图 6 ChatGPT 的技术突破点在于引入了 RLHF（基于人类反馈的强化学习） 6
图 7 主流大模型数据集来源可分为六类 8
图 8 各类数据来源大小 8
图 9 全球前十大科技企业数据中心容量排名 9
图 10 百度昆仑一、二代芯片与英伟达 A100 参数对比 9
图 11 2013 年起公司资本开支及经营现金流情况 10
图 12 2013 年起公司现金及现金等价物充沛（亿元） 10
图 13 AGI 产业链及底层支撑示意图 10
图 14 未接入 GPT4 的 Bing 仅对搜索结果进行简单整合 12
图 15 NewBing 的 Chat 入口可以对搜索结果进行人性化整合 12
图 16 NewBing 能够帮助用户编写代码 13
图 17 NewBing 能够帮助用户进行文件阅读 13
图 18 NewBing 发布次日 Bing 下载量猛增 758% 13
图 19 百度知识图谱的首要应用场景即为搜索 13
图 20 用户输入“宝可梦朱紫”后出现游戏购买链接 14
图 21 用户输入健美运动员姓名后出现健身补剂广告 14

表 1 Github 社区中主流 AI 框架情况（2022.1） 6
表 2 类 ChatGPT 模型年均训练成本测算 15
表 3 类 ChatGPT 应用中期年均推理成本测算 16
表 4 类 ChatGPT 应用中期年均成本测算 17

1.“文心一言”答卷未知，但国产替代并不遥远

3 月 16 日百度 AI 对话模型“文心一言”发布，我们据发布会信息总结，“文心一言”基于此前 ERNIE 大模型、PLATO 对话模型训练而成，是对百度 2019 年起便已开始的 NLP 实践的延续。从技术角度看，除百度已有的知识增强、检索增强、对话增强技术外，“文心一言”引入了有监督的精调、RLHF（基于人类反馈的强化学习）、提示学习等 ChatGPT 基础技术，但具体参数量、数据量、耗能、对话时效等均未公开。

事前录屏而非实机演示，对外界信心产生负面影响。从现场发布的 demo 来看，文心一言具备文学创作、商业文案创作、数理逻辑推算、中文理解、多模态生成能力，同类问题下中文理解能力强于 GPT4，且随输入内容生成音频、视频尚属 AI 对话模型中的首例，存在超预期亮点，但发布会未能展示模型的编程能力，且事前录屏降低了其演示的真实性，也并未对公众开放，多种因素导致公众反馈不佳。

图 1 “文心一言”生成图片

资料来源：百度，华西证券研究所

图 2 “文心一言”生成视频

资料来源：百度，华西证券研究所

“文心一言”展现了足够的文案创作能力，为 B 端降本增效的起始逻辑已经明晰。出于商业角度考虑和高昂的端侧微调成本，厂商普遍放弃开源，转而以提供 API 的方式供下游用户在特定场景下进行推理使用。以 ChatGPT 为例，OpenAI 并未公布其基础模型（GPT3.5、GPT4）技术细节，用户仅能够在自身应用内通过 API 调用其模型。从第一批用户实际使用来看，“文心一言”已经展示了基础文案工作能力，除此前接入 650 家企业外，发布当日有 6.5 万家企业申请测试，签约 5 家客户，一定程度反映了企业客户的认可程度。

用户交互能够进一步改善模型表现，我们判断这也是公司急于推动模型面世的原因之一。OpenAI 自 GPT-3 便开始对外提供服务，通过开放给公众，GPT3 收集来自用户输入内容的多样性数据，从而迭代出效果更好的模型，这就决定了 GPT4 是站在用户交互飞轮的巨人肩膀上，与文心一言并不在同一起跑线。但是海量用户群也是百度的长处之一，GPT 的飞轮效应是可复制的。尽管尚未对公众大范围开放，企业用户已经能够申请内测邀请码，邀请范围若持续扩大，飞轮效应将推进“文心一言”表现改善，且优化空间极大。

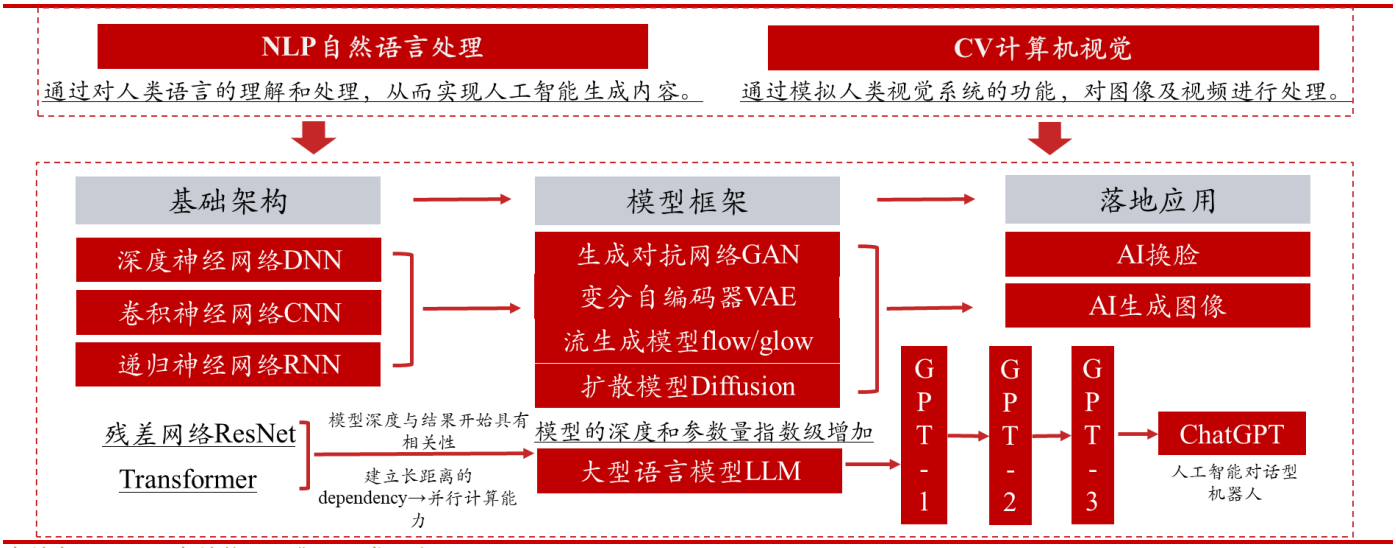
我们判断，尽管上下文理解、语义逻辑、多轮对话方面尚有欠缺，“文心一言”在部分问题处理上已经能够对标 GPT3 水平，但具体表现仍需时间和公众验证。我们依然认为，人工智能必然成为产业发展长期主线，国产替代具有需求上的紧迫性。以“文心一言”发布为契机，我们重点分析国内主流科技企业在复现 ChatGPT 领域需要克服的差距，判断对国产大模型的发展不必过度悲观。

1.1. 细究算法、数据、算力三要件，略逊一筹但仍有追平可能

1.1.1. 算法：核心差距在于方法及细节处理

大模型的技术积累已经行至一个质变节点。我们将 NLP（自然语言处理）及 CV（计算机视觉）技术视作类 ChatGPT 产品的技术底座，从深度学习的角度分析其发展历程，残差网络及 Transformer 的出现使得模型的深度和参数量指数级增加，大模型成为可能；大语言模型出现后，大模型的使用方式从预训练的单一任务模型迭代到多模态模型，微调时所需的标注数据量显著减少，从而降低了业务的使用成本。

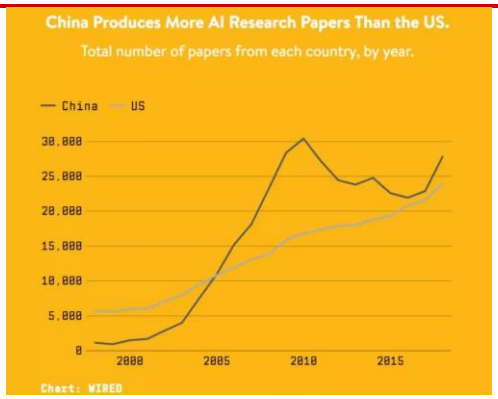
图 5 类 ChatGPT 产品的技术发展示意图



资料来源：公开资料整理，华西证券研究所

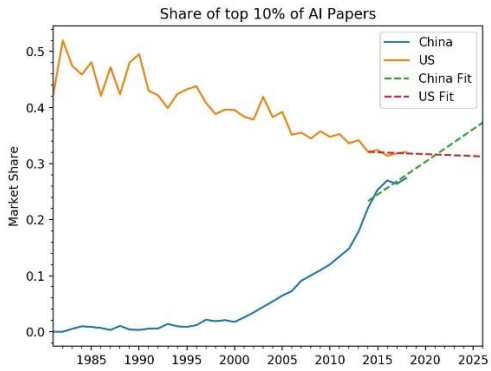
国内 AI 领域积累深厚，历年论文发表及专利申请占优。Elsevier 数据显示 2012-2021 年中国 AI 相关论文篇数始终排在首位，到 2021 年增至美国 2 倍。从论文引用次数进入前 10%的篇数来看，中国 2019 年跃居首位。2021 年达到比美国多 7 成的 7401 篇；斯坦福大学数据显示 2021 年中国提交的人工智能专利申请全球占比超 50%。

图 3 2005 年起中国 AI 论文总数超美国



资料来源：Wired，华西证券研究所

图 4 海外 AI 机构预测中国高引论文占比将超过美国



资料来源：Allen Institute for AI，华西证券研究所

从基本操作系统看，国内已经具备建立 AI 底层框架的能力。深度学习框架是实现算法的基础架构和工具，可类比为开发过程中必须使用的操作系统（如游戏制作过程中的虚幻引擎）。从技术定位看，AI 框架对下调用底层硬件计算资源，能够屏蔽底层差异并提供良好的执行性能,对上支撑 AI 应用算法模型搭建，提供算法工程化实现的标准环境，是 AI 体系的关键核心。目前海外 AI 框架领域已经形成 TensorFlow(Google)、PyTorch(Meta) 双寡头格局，国内主流 AI 框架主要有 PaddlePaddle（百度）、MindSpore（华为）、MegEngine（旷视）、OneFlow 等，从 Github 指标看，我国主体推出的 AI 框架中，华为 MindSpore、百度飞桨引用次数、点赞数、贡献者数量占优。

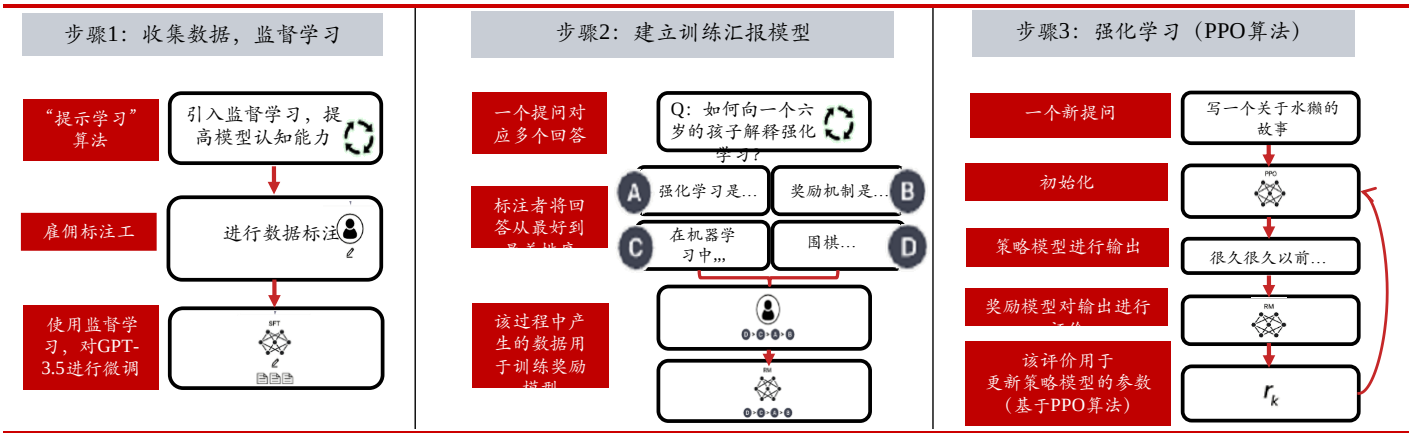
表 1 Github 社区中主流 AI 框架情况（2022.1）

地区	排名	框架	代码提交次数	引用数	点赞数	贡献者
海外	1	TensorFlow	124494	86300	163000	3056
	2	PyTorch	43390	14800	53700	2137
	3	Theano	28127	2500	9500	352
	4	CNTK	16116	4400	17100	201
	5	Maneet	11776	6900	19800	868
国内	1	PaddlePaddle	33753	4300	17500	524
	2	MindSpore	37308	514	2700	267
	3	MegEngine	2282	462	4100	32
	4	OneFlow	7621	351	3000	99
	5	Jittor	1266	235	2300	31

资料来源：AI 框架发展白皮书（2022 年），华西证券研究所

但 ChatGPT 在算法上的突破更多在于思路而非具体理论，是“菜谱”而非“食材”的创新，这成为了复现的难点之一。2022 年 11 月，OpenAI 上线了机器人对话模型 ChatGPT（GPT-3.5），引入了 RLHF（基于人类反馈的强化学习）：利用人类的标注数据去对 GPT3/GPT3.5 进行有监督训练，针对模型的多个回答进行排序标注，形成奖惩机制，让模型去拟合人的偏好，从而实现了史上最佳的输出效果。ChatGPT 并未实现任何底层理论的创新，更近于多种前沿算法理论组合，选取了大量的数据，设计了合理的标注流程，并且将这些融合，真正复杂的是这一过程。

图 6 ChatGPT 的技术突破点在于引入了 RLHF（基于人类反馈的强化学习）



资料来源：公开资料整理，华西证券研究所

OpenAI 逐步放弃开源，极大增加了国内科技企业的复现难度。相比谷歌此前公布了大量的模型原理，OpenAI 并未提供开源论文，大量的技术细节并未公开，GPT3.5 的参数规模也并不明确。尽管国内学术及业界均在 AI 领域有一定的积累，历

年论文发表及专利数占优，但在复现过程中大量细节都并不明朗，如提示学习的具体机制、算法如何泛化、算法微调的具体环节、数据标签的设置等。

从国产实践来看，学术界已有相对成功复现先例，但尚未工程化落地。百度“文心一言”外，清华智谱 ChatGLM 亦引入了监督微调、反馈自助、人类反馈强化学习等技术，尽管参数量较小，输出表现良好。2022 年 11 月，斯坦福大学大模型中心对全球 30 个主流大模型进行了全方位的评测，GLM-130B 是亚洲唯一入选的大模型，评测报告显示 GLM-130B 在准确性和公平性指标上与 GPT-3 接近或持平，鲁棒性、校准误差和无偏性均优于 GPT-3。

1.1.2.数据：数据集质量、标注细节处理、用户真实交互是关键

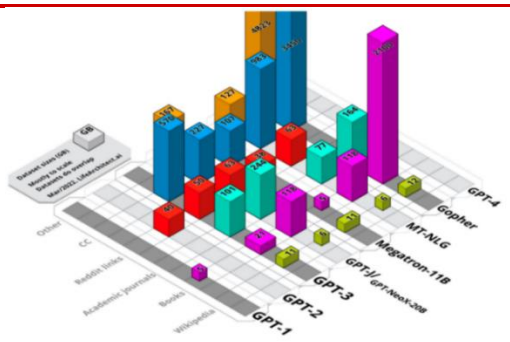
相比传统无监督学习的 GPT 模型，ChatGPT 表现更好的原因之一是在无监督学习的基础上提供了高质量的真实数据（精标的多轮对话数据和比较排序数据），主要得益于敏感词标注领域技术投入、对公众开放后形成的数据飞轮。OpenAI 并没有公开训练 ChatGPT 的相关数据集来源和具体细节，我们参考 Alan D. Thompson 文章，判断主流大模型数据集来源可分为六类，分别是：维基百科、书籍、期刊、Reddit（社交媒体平台）链接、Common Crawl（大型数据集）和其他数据集（GitHub 等代码数据集、StackExchange 等对话论坛和视频字幕数据集）。

图 7 主流大模型数据集来源可分为六类

	Wikipedia	Books	Journals	Reddit links	CC	Other	Total
GPT-1		4.6					4.6
GPT-2				40			40
GPT-3	11.4	21	101	50	570		753
The Pile v1	6	118	244	63	227	167	825
Megatron-11B	11.4	4.6		38	107		161
MT-NLG	6.4	118	77	63	983	127	1374
Gopher	12.5	2100	164.4		3450	4823	10550

资料来源：Alan D. Thompson，华西证券研究所

图 8 各类数据来源大小



资料来源：Alan D. Thompson，华西证券研究所

国内厂商在中文训练数据方面有一定优势，以百度为例，ERNIE 3.0 的中文预训练语料数量最多，主要来源为 ERNIE 2.0（包括百科、Feed 等）、百度搜索（包括百家号、知乎、铁算盘、经验）、网络文本、QA-long、QA-short、Poetry&Couplet3、医学、法律、金融等领域的特定数据以及百度知识图谱（超过 5000 万条事实）。

但中文互联网语料质量相对较差，优质的中文标注数据集匮乏，使用英文数据进行预训练更为可行。RLHF 论文中的训练数据英文占比极高，但对中文和其他小语种能力的提升非常显著，可见 RLHF 对模型能力的提升能够跨越语种，以 ChatGLM-6B 为例，该模型在 1:1 比例的中英语料上训练了 1T 的 token 量，兼顾双语能力，**我们认为中文数据集的薄弱对于国产大模型而言并不构成较大阻碍。**

精细标注、对标注人员的培训是技术难点。ChatGPT 的前身 GPT-3 已经展示了非常强大的语句串联的能力，但互联网的词汇存在负面信息，单纯凭借学习能力无法来清除这些训练数据。GPT3.5 给 AI 提供标有暴力、仇恨语言等标签，AI 工具就可以学会检测这些内容，并在它触及到用户之前将不良内容过滤掉。尽管标签主要通过人工标注，具体标注技术细节、对标注员的培训等仍需要国内科技企业探索。

科技企业能够参考 GPT3 的路径，利用海量用户交互提升数据质量。作为国内最大的搜索引擎服务商，百度在真实数据和用户需求理解方面有较多积累，能够对旗下的 AI 大模型进行充分的训练和预测，进而使得 AI 大模型的智能化水平不断进化。参考 GPT3 的经验，GPT-3 是 OpenAI 正式对外提供服务的模型。通过将模型开放给公众，GPT3 收集来自用户输入内容的多样性数据，从而迭代出效果越好的模型。若文心一言向公众开放，我们判断真实的用户调用与模型迭代之间有望形成正向循环，缩窄国产大模型产品与 ChatGPT 的差距。

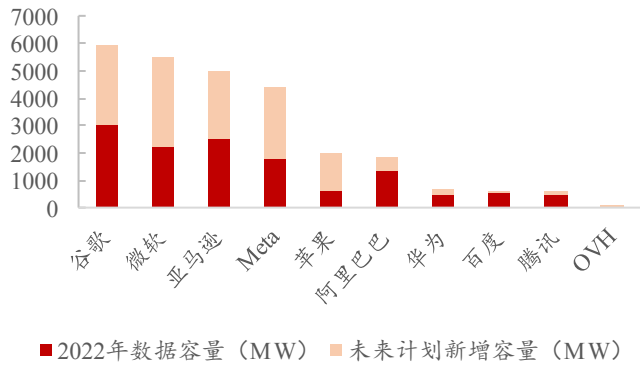
1.1.3.算力：更多关乎资金充足程度与公司战略

类 ChatGPT 产品的复现需要庞大的算力支持。浮点运算（FLOPS）的数值通常与参数数量成比例，随着 GPT、GPT-2 和 GPT-3（当前开放的版本为 GPT-3.5）的参数量从 1.17 亿增加到 1750 亿，预训练数据量从 5GB 增加到 45TB，算力需求随之增长。据 OpenAI，训练一次 13 亿参数的 GPT-3 XL 模型需要的全部算力约为 27.5PFlop/s-day（即 1PetaFLOP/s 效率跑 27.5 天），训练一次 1746 亿参数的 GPT-3 模型需要的算力约为 3640 PFlop/s-day（即 1PetaFLOP/s 效率跑 3640 天），需求呈现指数级增长。

衡量头部厂商能否支撑训练及推理环节的算力需求，我们认为更多需要考虑资金充足程度与公司战略。模型层企业更多是算力的消费者，美国芯片出口政策影响存在但并不致命。对试图复现 ChatGPT 的头部厂商而言，我们认为算力更接近于自由流通的商品，只需高额的资金投入即可完成布局。尽管美国出口限制政策影响较大，国内仍能采购性能更低的前代算力芯片，只是相对牺牲了计算速度。

国内头部科技企业多已完成数据中心建设，能够实现算力资源部分自给。Structure Research 数据显示到 2022 年全球超大规模自建数据中心总容量将达到 13177 兆瓦（MW），中国超大规模数据中心企业占亚太地区的 24%，阿里巴巴、华为、百度、腾讯和金山云均位于领先位置。

图 9 全球前十大科技企业数据中心容量排名



资料来源：Structure Research，华西证券研究所

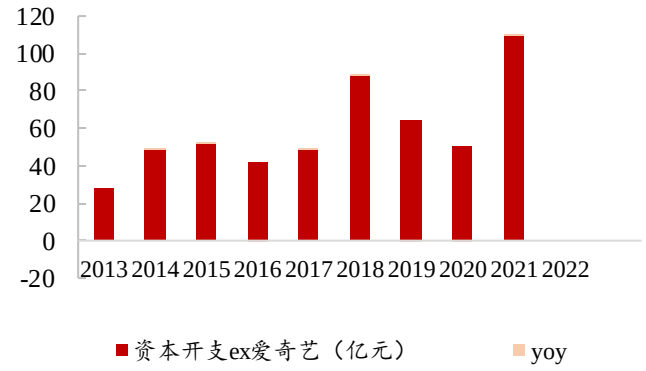
图 10 百度昆仑一、二代芯片与英伟达 A100 参数对比

参数	昆仑 1	昆仑 2	英伟达 N100
INT8	256TOPS	512-768TOPS	624/1248*TOPS
INT/FP16	64TOPS	128-192TOPS	312/624*TFLOPS
Tensor Float32	-	-	156/312*TFLOPS
INT/FP32	16TOPS	32-48TOPS	19.5TFLOPS
FP64 Tensor Core	-	-	19.5TFLOPS
FP64	-	-	9.7TFLOPS

资料来源：MIT《机器学习加速器的调查和基准测试》研究，华西证券研究所

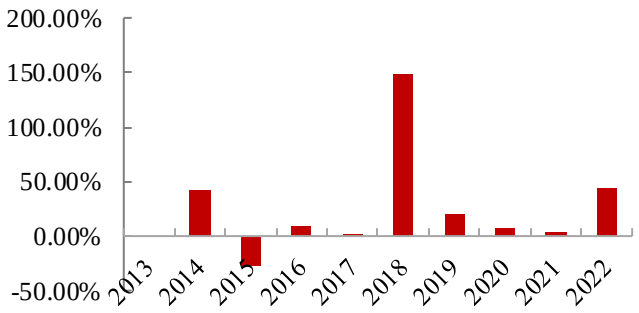
长线投入战略、充足资金预算，是复现 ChatGPT 所必须的要素。因此，战略押注意愿明确、现金流充沛的国内厂商更有希望弥合算力上的差距。以百度为例，2017 年提出“All IN AI”后资本开支波动上升，2022 全年资本开支（除爱奇艺）高达 181 亿元，同期经营现金流增长 30%至 261.7 亿元，截至 2022 年末公司用于进行资本支出的现金及现金等价物余额为 531.6 亿元，现金流稳健、流动性充足。

图 11 2013 年起公司资本开支及经营现金流情况



资料来源：公司公告，华西证券研究所

图 12 2013 年起公司现金及现金等价物充沛（亿元）



资料来源：公司公告，华西证券研究所

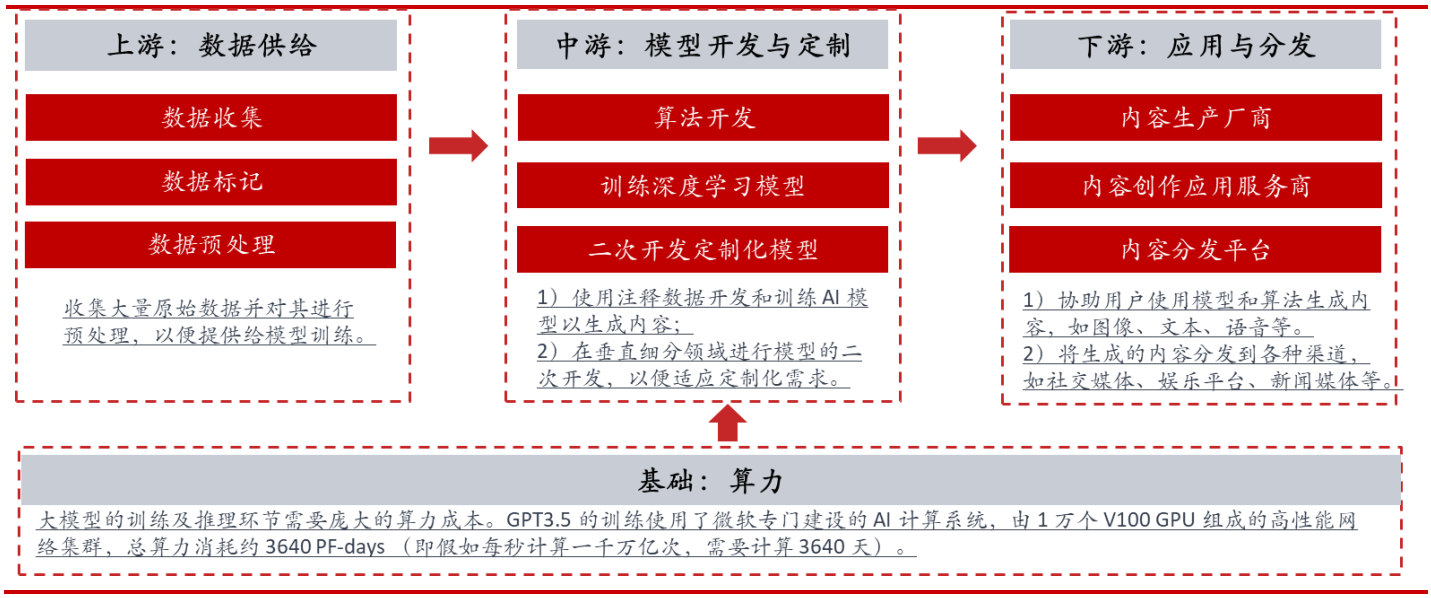
1.2. 工程化处理与分发能力是更高的壁垒

从国产替代角度看，算法、数据、算力壁垒并非不可逾越，随着人才流动、时间推移和研究进步，大模型性能很可能逐渐趋同。从落地角度来看，**ChatGPT 更多实现了工程而非技术上的成功，即完成了从底层技术到工程落地再到产品的跨越，在成本、规模和效率之间实现了正确的权衡取舍。**

1）从 B 端来看，此前的 AI 公司需要根据不同的任务训练出不同的模型，算法很难复用于其他场景。OpenAI 提供了泛化/通用的算法，让更多的使用者受益并为之消费，各行业厂商可以利用统一的预训练大模型+微调来直接应对不同的任务，后续开发工程量大幅下降，实现了 AI 开发的工业化。

2）从 C 端来看，ChatGPT 做到了在稳定输出的同时容纳破亿月活用户，具有低使用门槛、高效、强传播性的特点，开启了 AI 领域大众化的序章。

图 13 AGI 产业链及底层支撑示意图



资料来源：公开资料整理，华西证券研究所

梳理海外AGI产业链发展趋势，我们判断国内厂商相对优势存在于国内局部场景的应用。AGI 受益产业链可划分为 1）上游数据及底层算力供给；2）中游 ToB 定制化模型开发；3）下游形成 ToC 应用并进行分发。海外目前涌现的公司则集中于以下三类：1）专注于大模型开发的公司，对外允许开发者以其预训练大模型为底座，通过微调或 API 针对不同的细分领域开发应用场景，如 OpenAI；2）兼具大模型开发及垂直应用一体化能力的公司，如 Midjourney；3）单纯调用大模型 API 开发具体场景应用的公司，如 JasperAI。因此，**考虑到中外环境的显著差异、技术水平尚有差距，“模型及服务”模式下提供国内特定场景下的定制化商业模型，或面向国内C端消费者提供内容生产应用，更加有望成为国内厂商弯道超车的机会。**这意味着国产 ChatGPT 的落地在技术准备之外还需要两项核心要素：工程化与分发能力。

1）**工程化能力**，即能够利用更低的成本和更高效的迭代做出先进的大模型应用，制作更高效、廉价、贴合市场的产品，能够同时容纳亿级用户在线。大模型调优周期长、难度大，产业相关工程化经验欠缺，难以实现大模型对数据的充分吸收与利用。在 ChatGPT 之前，已有多个行业领域涌现出智能化应用，但其工程化落地情况并不理想。ChatGPT 做到了在稳定输出的同时容纳破亿月活用户，具有低使用门槛、高效、强传播性的特点，但其对于训练集群、代码编译等细节的优化处理至今未披露，而这些工程化细节正是国内厂商目前核心差距所在。

从主流科技企业用户数来看，百度、淘宝、微信等国民级应用均能承载亿级日活，我们认为国内厂商完全具备大 DAU 场景下 AI 工程化处理的潜力。

2）**充沛的 C 端用户及 B 端应用场景**，即更低的分发触达成本、更快速的产品应用迭代，此时下游流量的充沛程度便是变现的壁垒。应用分发是国内企业的长项，头部厂商本身已经建起规模及心智壁垒，全球竞争力坚实，且变现路径可以复用，商业模式无需再探索，如要落地 AGI 相关应用，获客成本远低于新进入者。

1.3. 商业化路径已经明晰，搜索场景鲜明契合

除开星辰大海的展望，商业化变现的落地更能驱动企业加大科技投入，而类 ChatGPT 产品的变现逻辑已经相当坚实。

1）为开发者直接提供 API 调用接口。此前诸如 Google、Open AI 等头部厂商将自身开发的预训练大模型开源，供下游应用者在这些模型上进行参数的微调。但随着预训练语言模型的规模急剧增长，出于商业角度考虑和高昂的端侧微调成本，大规模预训练语言模型不再被开源，转而以提供 API 的方式供下游用户在特定场景下进行推理使用。以 ChatGPT 为例，OpenAI 并未公布其基础模型（GPT3.5、GPT4）技术细节，用户仅能够在自身应用内通过 API 调用其模型。从这一角度看，“文心一言”已经以 API 形式接入 650 家企业，发布当日已有 6.5 万家企业申请测试，签约 5 家客户，为 B 端降本增效的起始逻辑已经明晰。

2）“模型即服务”，即为细分行业定制大模型。厂商以公司的预训练大模型底座结合用户的数据集，通过微调来提升模型在某一细分领域的表现，但这一模式可能随着 AGI（通用人工智能）的发展逐步式微，因为端侧精调成本昂贵，通用大模型出现后出售 API 接口会更主流。

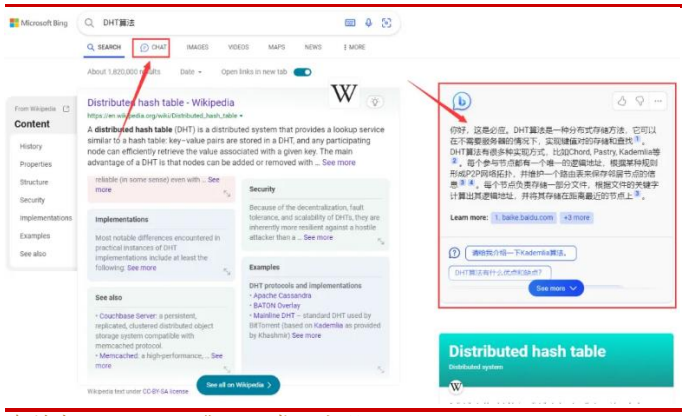
3）面向 C 端提供杀手级应用。头部产品 ChatGPT 日活超一亿人，SaaS 公司 JasperAI 通过调用 GPT-3 API 已实现超 7500 万美元收入，国内头部科技企业既具备自建应用的能力，也具备向 JasperAI 等广大创业公司提供 API 的能力。

4）将 AGI 技术嵌入到自身成熟应用中，提供更强用户体验，进而推动用户为附加服务付费。目前搜索引擎及办公软件领域已有成熟案例，2022 年 2 月微软将 GPT 模型嵌入至其搜索引擎 Bing 中，New Bing 能够与用户对话、协助用户起草文本；此外微软已经将 Office365、云计算模型 Azure 中悉数加入 AI 大模型，线上会议软件 Teams 已经推出了 AI 撰写会议纪要和邮件的付费服务。

图 14 未接入 GPT4 的 Bing 仅对搜索结果进行简单整合

资料来源：Bing，华西证券研究所

图 15 NewBing 的 Chat 入口可以对搜索结果进行人性化整合



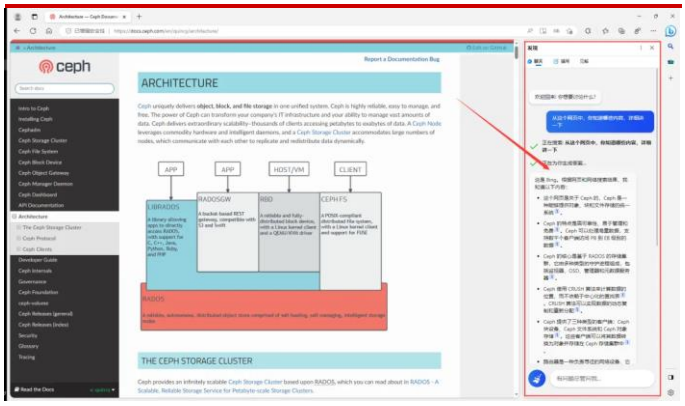
资料来源：Bing，华西证券研究所

ToC 对用户心智的冲击更为直观，我们高度重视“文心一言”为百度搜索带来的增量。搜索平台是最直观的对话模型使用场景，与类 ChatGPT 产品双向互补：1）引擎的海量数据能够弥补对话模型数据更新不及时的缺陷。ChatGPT 的训练数据集仍停留在 2021 年，因此难以回答时效性问题，而 BingChat 基于搜索库内容对内容进行回答、给出相应的搜索结果。2）GPT 的语言处理能力又使得模型能够对搜索结果进行直观集成，增进用户体验。目前微软为 NewBing 设置了三种性格状态，用户可根据偏好自行设定对话模型的回应风格。

图 16 NewBing 能够帮助用户编写代码

资料来源：Bing，华西证券研究所

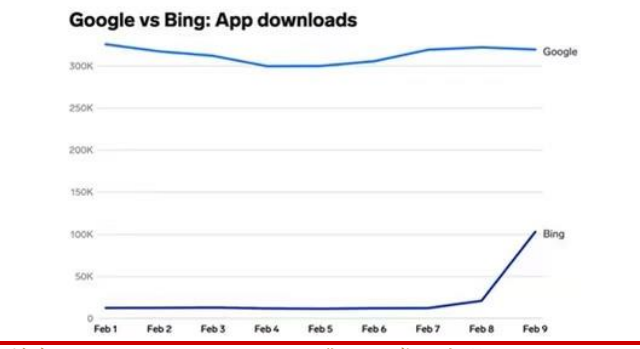
图 17 NewBing 能够帮助用户进行文件阅读



资料来源：Bing，华西证券研究所

Bing 的崛起已经证明了人工生成智能作为新介入因素能够对曾经的垄断巨头产生威胁，百度将“文心一言”嵌入搜索具有必要性。此前海内外搜索市场均呈现一家独大的格局，据 statcounter，截至 2022 年 6 月，Google 在全球搜索引擎的市占率达到 92.42%，远超第二名 bing 的 3.45%、第三名 Yahoo!的 1.32%。但 NewBing 的发布打破了这一稳态，data.ai 数据显示新功能上线当日必应 Bing 应用程序的全球下载量在一夜之间猛增十倍，蹿升至苹果 App Store 应用商店最受欢迎的免费应用榜中的第十位，并使其成为第二大最受欢迎的免费生产力应用，仅次于谷歌邮箱 Gmail。截至 3 月 10 日，Bing 活跃用户已突破 1 亿人，增幅超 600%。我们认为作为国内搜索巨头，百度在搜索领域进行人工生成智能布局已经成为战略上的必需。

图 18 New Bing 发布次日 Bing 下载量猛增 758%



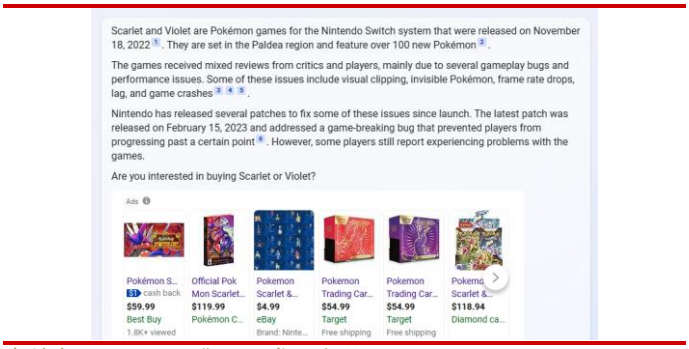
资料来源：Business Insider，华西证券研究所

图 19 百度知识图谱的首要应用场景即为搜索

资料来源：百度，华西证券研究所

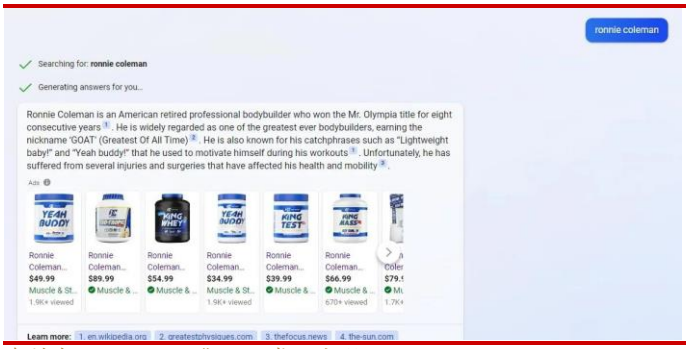
从 Bing 的成功实践来看，尽管接入 GPT4 带来了边际成本的显著增长，但商业模式已经雏形初现。随着用户数量级的提升，用户对话带来的推理成本（我们将在下文对搜索业务推理成本进行年化分析）将显著拉低搜索业务的毛利率，但人性化的对话体验是提升用户基数、用户粘性及使用时长的利器，Bing 借此实现了用户数量的快速增长和变现，广告引流已经开始。 Bing 根据用户输入的问题在下方广告位对电商产品进行展示引流，沿用了搜索广告的旧逻辑，即基于用户的搜索关键词，对应到精确的广告，Chat 界面事实上提供了新的广告位。**我们认为这只是新搜索引擎变现的起点，后续微软生态内的各类插件和增值服务均有望集成至 Bing，带来新的商业变现机会，这也是一条百度完全有能力复制的路径。**

图 20 用户输入“宝可梦朱紫”后出现游戏购买链接



资料来源：Bing，华西证券研究所

图 21 用户输入健美运动员姓名后出现健身补剂广告



资料来源：Bing，华西证券研究所

2. 若“文心一言”成功对公众开放，年化增量成本可控

短期我们关注“文心一言”等产品对科技企业财务状况的影响，将增量成本拆分为训练成本、推理成本及数据标注成本（暂不考虑人力支出及维护费用），测算大模型落地搜索页面后年均增量成本约为 16 亿元，这一金额有望随着技术进步逐步降低。

具体来看，ChatGPT 算力成本包括训练、推理及数据标注。1）训练：基于大量数据集来调整和优化模型的参数，对模型做反复迭代计算，是大模型落地前的成本环节；2）推理：大模型的运行成本，即模型基于用户输入信息进行推理计算并输出过程中产生的成本，与用户数量；3）数据标注：通过数据标注提高数据集质量。

2.1.训练：前期固定投入较大，莱特定律驱动下成本必然下行

假设所有厂商站在同一起跑线，我们将单次训练成本的测算思路总结为训练天数 x 云计算成本：

1）以达到模型预期效果所需消耗的训练 token 数量为基础，结合 GPU 在训练过程中的 token 吞吐能力，来计算在一定 GPU 数量下完成训练需要的天数。

2）根据 GPU 数量及计算出的的训练天数，假设云计算的市场价格能够代表训练过程中的硬件及能源成本（取 Azure 公开价格，但已经完成自有数据中心建设的厂商实际年化训练成本应当低于云计算市场价），从而计算出多次迭代训练的年化成本。

参考英伟达 Megatron-LM 团队在 2021 年发表的论文，完成一个 Epoch 的端到端训练时间=$\frac{8TP}{NX}$。具体来看，关键名词可以理解为：1）Epoch：将所有训练样本训练一次的过程，当一个完整的数据集通过了神经网络一次并且返回了一次，这个过程称为一次 Epoch。2）T：到模型预期效果所需消耗的训练 token 数量。3）P：大模型的参数数量。4）N：完成一次训练所需的 GPU 数量。5）X：GPU 能够达到的有效吞吐量。

根据这一计算方法，考虑到 GPT3.5 系在 GPT3 基础上微调而成，我们选取 1750 亿参数的 GPT-3 模型为样本进行测算，在包含 3000 亿 tokens 的数据集上，假设完成一组训练需要 1024 张英伟达 A100 芯片，且 GPU 在训练过程中能达到的有效计算吞吐为 140TFlops（每秒浮点运算次数），那么完成一次训练需要 34 天。

在此基础上，考虑到大模型的训练过程并不是一劳永逸的，我们假设完成实际训练至少需要两组芯片（2046 张）留出试错空间，为了充分进行模型训练需要不间断进行训练，训练过程中会产生 20%的试错成本，对厂商的年均实际训练成本做出调整，以 Azure ND A100 v4 series（8 张 A100 芯片）的服务器三年包年月租 1 万美元为基准，测算出采用云计算的前提下，大模型的年化训练成本为 2.29 亿元。但根据 OpenAI 数据，ChatGPT 离线训练成本仅为 1200 万美元，考虑到微软前期已经进行大量硬件投入，我们据此认为对于已完成自有数据中心建设的厂商，实际训练成本会更低。

表 2 类 ChatGPT 模型年均训练成本测算

训练所需 GPU 数量	2048
单台服务器 GPU 数量	8
服务器数量	256
单次训练时长（天）	34
年均训练次数	10.7
服务器月租成本（万美元）	1
试错空间	20%
年均训练成本（万美元）	3,298

美元汇率	6.94
年均训练成本（亿元）	2.29

资料来源：Azure，华西证券研究所，美元汇率为 2023 年 3 月 7 日数据

我们认为在人工智能领域，莱特定律依然有效，大模型的训练成本正随着硬件、软件的优化不断降低，目前海外已有成功商业化实践。

1）莱特定律，指飞机制造数量每累计增加一倍，制造成本就会实现固定百分比的持续下降，主要系人工熟练度的提升、生产工序的优化和原材料的节约导致。莱特定律的作用弹性取决于新技术的累积产量翻倍时间，该过程的时间越短，成本下降的作用越明显。因此在技术应用早期，产量翻倍时间短，莱特定律作用会更显著。海外机构 ARK 据此判断人工智能相对计算单元（RCU）的生产成本及软件成本将分别以每年 57%和 47%的速度下降，硬件和软件的融合可以推动人工智能训练成本在 2030 年前每年 70%的速度下降。参考 OpenAI 创始人提出每十八个月 AI 算力需求翻倍，我们认为随着应用程度加深、优化需求提升，大模型的训练成本降低是必然趋势。

2）从海外相关商业化尝试来看，专业优化公司 MosaicML 通过下调模型参数量、调升训练数据集 tokens 数量，将具备 GPT-3 同等能力的大模型的一次训练成本下调至 45 万美元，较 2020 年下降超 90%。我们认为国内厂商实现同等能力只是时间问题。

2.2.推理：与用户数量成正比，成本优化路径明确

推理成本最直观实际的估算方法是参考市场上现有基础模型 API 的标价。OpenAI 为 GPT3 及 ChatGPT 先后设置的 API 定价分别为 0.02 美元/1000tokens 及 0.002 美元/1000tokens，目前 ChatGPT 成本降低 90%的原因尚未披露。我们综合参考 SaaS 业务普遍毛利率情况、OneFlow 测算及 Azure 云计算价格，判断对于国内厂商而言，选取 GPT3 的 API 价格（而不是 ChatGPT 的超低价）并对其进行成本溢价的调整更加合理，判断合理成本约千字 0.07 人民币元(中文预训练模型将中文拆成一个个的字做学习，每一个 token 对应一个汉字)。

我们参考 NewBing 数据，上线 ChatGPT 聊天机器人功能后用户数突破一亿，约 1/3 用户每日在搜索页面使用对话功能，我们判断单用户对话数约在 5 次，单次输出 token 约在 100 个。相比嵌入 AI 后用户增长 6 倍的 Bing，国内用户基数更大、AI 认知程度低，用户使用内嵌对话模型的比例极有可能低于 1/3，**我们判断亿级日活应当是国产类 ChatGPT 应用更加合理的中期天花板，据此测算年度推理成本约在 13.62 亿元。**

表 3 类 ChatGPT 应用中期年均推理成本测算

类 ChatGPT 应用日活用户数（亿人）	1.07
单人单日对话次数	5
平均 token 输出个数	100
单日推理成本（万元）	373
年均推理成本（亿元）	13.62

资料来源：公司公告，华西证券研究所测算

尽管国内厂商难以短期内实现与 OpenAI 同等的成本优化程度，推理成本的优化目前已经有明确的路径：1）压缩模型以减少总内存占用量：通过使用模型压缩技术，如权重共享、量化和剪枝，可以降低模型的内存占用量和计算复杂度，从而降低推理成本。2）协同推理：多个设备或服务器共享计算任务，通过去中心化分摊成本。3）计算卸载：部分计算任务从一个设备卸载到另一个设备，从而实现对计算资源的优化分配。例如，将部分计算任务从 CPU 卸载到 GPU，从而提高计算效率。4）知识蒸

馏：通过让一个较小的模型（学生模型）学习一个较大的模型（教师模型）的知识，可以在保持较好性能的同时降低模型的复杂度和推理成本。

相对于独立的模型应用，搜索页面的对话模型输出效果有限，科技企业出于经济考虑对细分场景下使用的模型进行推理成本优化存在极大可能性，因此我们给出的推理成本预测仍有下调空间。

2.3.数据标注：取决于人力价格，成本量级较低

相比传统无监督学习的 GPT 模型，ChatGPT 表现更好的原因之一是在无监督基础上提升了训练数据的质量，而实现这一点的经济成本并不高。OpenAI 借鉴了 Facebook 等社交媒体公司的做法，构建一个额外的 AI 检测器，向它提供带有暴力、仇恨言论等标签的示例，让它学会识别有害内容。该检测器被内置到 ChatGPT 中，以检测输出内容是否反映了其训练数据的问题，并在它到达用户之前将其过滤。为了获得这些不良内容的标签，OpenAI 在 2021 年 11 月将标注工作交给肯尼亚外包公司 Sama，三份合同总价仅为 20 万美元。我们参考时代周刊调查数据，数据标注员团队为 30 人，每 9 小时轮班阅读和标记 150 至 250 段文字，每段 100-1000 词不等，实得工资约为每小时 1.32 美元至 2 美元。

我们据此认为，数据标注目前壁垒仍集中于技术领域，而从经济成本上看，尽管 ChatGPT 的数据标注工作并非完全由 Sama 完成（2022 年 2 月 Sama 与 OpenAI 终止合作），我们判断其完整标注成本并不高，给出年化 500 万元的估算。

表 4 类 ChatGPT 应用中期年均成本测算

年均训练成本（亿元）	2.29
年均推理成本（亿元）	13.62
年均数据标注成本（亿元）	0.05
年均成本合计（亿元）	15.96

资料来源：Azure，OpenAI，时代周刊，华西证券研究所测算

综合以上分析，我们粗略推测类 ChatGPT 应用正式运营后为公司带来的年化增量成本约为 16 亿元。但考虑到国产模型参数量及数据集 token 数量均是未知，且 Azure 云计算价格与实际成本存在偏差，此外实际落地后各项成本均存在优化可能、具体会计处理方式还存在探讨空间，我们判断 10-20 亿元为其增量成本的合理范围（暂不考虑人力支出及维护费用）。

3. 投资建议

微软新产品对公众的震撼只是前期技术突破后的余波（Copilot 的模型基础 GPT4 早在 2022 年 8 月便已训练完成），而非通用人工智能领域想象力的终点，产品的成功会驱动更多学术资源与产业投资的倾斜，更多直接应用的突破已经箭在弦上。

人工智能必然成为产业发展长期主线，国产替代具有需求上的紧迫性。以“文心一言”发布为契机，我们重点分析国内主流科技企业在复现 ChatGPT 领域需要克服的差距，判断对国产大模型的发展不必过度悲观，算法、算力、数据差距可以弥合，工程化、应用分发能力、交互飞轮可能成为厂商弯道超车的契机。

具体投资机会来看，依然重点推荐百度（BIDU.US），判断“文心一言”表现符合预期，公司目前仍处于低估区间，尽管技术差距依然存在，短期内研发投入可能上行，我们看好人工智能领域投入对公司业绩及估值的长线提振。同时关注已在视频、营销、阅读等相关细分领域抢跑的重点标的，推荐当虹科技、捷成股份、蓝色光标、风语筑、浙文互联。

4. 风险提示

“文心一言”落地效果不及预期风险；ToB 服务推进节奏不及预期风险；成本大幅增长风险；AI 产品道德及监管风险。

分析师与研究助理简介

赵琳：华西证券传媒行业首席，南开大学本硕。本科毕业后自愿到乡村学校长期支教后担任校长，期间获《中国教育报》头版头条关注报道。2017年硕士毕业后到新时代证券从事传媒行业研究，2019年加盟华西证券。

分析师承诺

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，保证报告所采用的数据均来自合规渠道，分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求客观、公正，结论不受任何第三方的授意、影响，特此声明。

评级说明

公司评级标准	投资评级	说明
以报告发布日后的6个月内公司股价相对上证指数的涨跌幅为基准。	买入	分析师预测在此期间股价相对强于上证指数达到或超过15%
	增持	分析师预测在此期间股价相对强于上证指数在5%—15%之间
	中性	分析师预测在此期间股价相对上证指数在-5%—5%之间
	减持	分析师预测在此期间股价相对弱于上证指数5%—15%之间
	卖出	分析师预测在此期间股价相对弱于上证指数达到或超过15%
行业评级标准		
以报告发布日后的6个月内行业指数的涨跌幅为基准。	推荐	分析师预测在此期间行业指数相对强于上证指数达到或超过10%
	中性	分析师预测在此期间行业指数相对上证指数在-10%—10%之间
	回避	分析师预测在此期间行业指数相对弱于上证指数达到或超过10%

华西证券研究所：

地址：北京市西城区太平桥大街丰汇园11号丰汇时代大厦南座5层

网址：http://www.hx168.com.cn/hxzq/hxindex.html

华西证券免责声明

华西证券股份有限公司（以下简称“本公司”）具备证券投资咨询业务资格。本报告仅供本公司签约客户使用。本公司不会因接收人收到或者经由其他渠道转发收到本报告而直接视其为本公司客户。

本报告基于本公司研究所及其研究人员认为的已经公开的资料或者研究人员的实地调研资料，但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。本报告所载资料、意见以及推测仅于本报告发布当日的判断，且这种判断受到研究方法、研究依据等多方面的制约。在不同时期，本公司可发出与本报告所载资料、意见及预测不一致的报告。本公司不保证本报告所含信息始终保持在最新状态。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，投资者需自行关注相应更新或修改。

在任何情况下，本报告仅提供给签约客户参考使用，任何信息或所表述的意见绝不构成对任何人的投资建议。市场有风险，投资需谨慎。投资者不应将本报告视为做出投资决策的惟一参考因素，亦不应认为本报告可以取代自己的判断。在任何情况下，本报告均未考虑到个别客户的特殊投资目标、财务状况或需求，不能作为客户进行客户买卖、认购证券或者其他金融工具的保证或邀请。在任何情况下，本公司、本公司员工或者其他关联方均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而导致的任何可能损失负有任何责任。投资者因使用本公司研究报告做出的任何投资决策均是独立行为，与本公司、本公司员工及其他关联方无关。

本公司建立起信息隔离墙制度、跨墙制度来规范管理跨部门、跨关联机构之间的信息流动。务请投资者注意，在法律许可的前提下，本公司及其所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。在法律许可的前提下，本公司的董事、高级职员或员工可能担任本报告所提到的公司的董事。