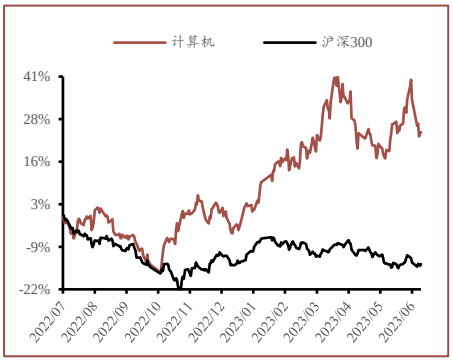
财通证券 CAITONG SECURITIES

配置 AI 算力和 C 端应用核心，关注 B 端应用 AI 进展

证券研究报告

投资评级:看好(维持)

最近 12 月市场表现



分析师 杨烨
SAC 证书编号：S0160522050001
yangye01@ctsec.com

相关报告

1. 《AI 监管走到什么阶段了？》 2023-06-25
2. 《模型成本持续降低，大规模商业变现渐行渐近》 2023-06-19
3. 《AI 带领计算机进入强比较优势阶段》 2023-06-11

核心观点

- **海外风险扰动，大模型 B 端边际加速：**本周计算机指数下跌 7.09%，跑输沪深 300 指数 7.22pct，在 31 个申万一级行业中涨幅排名 31，年初至今计算机以 27.57%的涨幅排名第 3。本周海外地缘政治不确定性加剧，板块迎来较大幅度调整。我们此前看好板块行情的持续性的三个出发点未曾改变：1）基本面提供确定性、2）流动性带来可能性、3）政策力度决定 β 强度。在此基础上，我们认为大模型在 B 端已具备基本技术条件，各公司也已陆续推出垂直领域的大模型产品与服务， B 端商业化即将进入快速落地阶段。

- **大模型在 B 端落地已具备基本技术条件：**全球的企业和开发者都在加速探索如何将 AI 大模型落地到现有的 B 端的商业场景中，包括但不限于基于基于已有垂域数据的价值挖掘，更智能化的问答体验，更高效率的自动化办公场景等等。目前开发者将垂域知识引入大模型主要采用两种思路：1）通过 Fine-Tuning 将垂域知识训练到模型的参数中、2）通过 In-Context Learning 将垂域知识放在模型的 prompt 中。前者较有代表性的范式是 Delta-Tuning，这其中最被开发者广泛关注和使用的是微软提出了 LoRA（Low-Rank Adaptation），其主要通过引入可训练的低秩矩阵，显著降低了微调模型的成本。后者较为常见的是使用 Langchain+向量数据库的组合方案，通过将用户输入的 prompt 与向量数据库中相关的内容一起输入给大模型，成为了另一种可实现“大模型的通用能力+垂直领域的专业知识”的技术路径。

- **AIGC 加速赋能 B 端用户，注重数据质量与专业性：**由于安全合规以及大模型在细分行业回答精准度等问题，B 端用户对 AI 大模型的接受节奏略慢于 C 端，当前利用通用数据+行业专业数据训练的面向垂直领域大模型逐步落地，将加速企业广泛使用大模型技术，AI 在 B 端应用迎来加速。恒生电子将 AI 技术与金融业务 know-how 结合，打造金融行业大模型和全新数智产品，推出金融智能助手“光子”和智能投研平台 WarrenQ，以及底层金融行业大模型 LightGPT；拓尔思拓天大模型聚焦优势行业，利用自有的高质量数据进行预训练，推出适用于媒体、金融、政务的三大行业大模型；Glean 定位基于 AI 的企业搜索与知识管理平台，协助用户跨应用、个性化搜索，能够更快、更准确地找到企业内部的知识和数据，打通各类 SaaS 化应用。

- **投资建议：**见正文。

- **风险提示：**AI 技术迭代不及预期的风险，商业化落地不及预期的风险，政策支持不及预期风险，全球宏观经济风险。

内容目录

图表目录

1 本周回顾：海外风险扰动，大模型 B 端边际加速

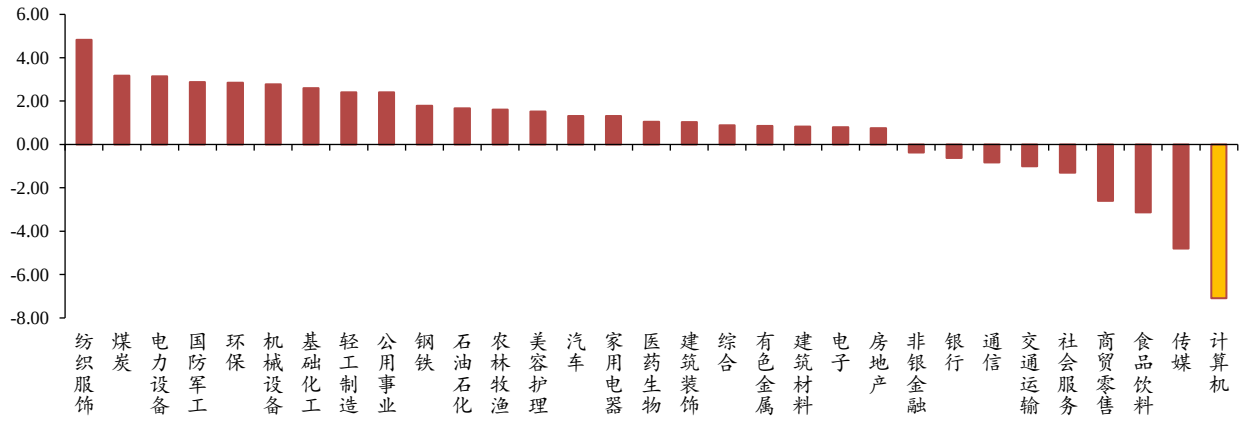
行情短期波动，AI 大模型 B 端应用边际加速。本周计算机指数下跌 7.09%，跑输沪深 300 指数 7.22pct，在 31 个申万一级行业中涨幅排名 31，年初至今计算机以 27.57%的涨幅排名第 3。本周海外地缘政治不确定性加剧，板块迎来较大幅度调整。我们此前看好板块行情的持续性的三个出发点未曾改变：1）基本面提供确定性、2）流动性带来可能性、3）政策力度决定β强度。在此基础上，我们认为大模型在 B 端已具备基本技术条件，各公司也已陆续推出垂直领域的大模型产品与服务，B 端商业化有望逐步进入快速落地阶段。

图1.计算机板块相对各指数涨跌幅统计（2023.6.26-2023.6.30，单位:%）

代码	名称	近 5 日涨跌幅	年初至今涨跌幅	周相对涨跌幅	年初至今相对涨跌幅
801750.SI	计算机（申万）	-7.09	27.57	-	-
000001.SH	上证指数	0.13	3.65	-7.22	23.92
000300.SH	沪深 300	-0.56	-0.75	-6.53	28.33
399006.SZ	创业板指	0.14	-5.61	-7.23	33.19

数据来源：Wind，财通证券研究所

图2.本周各行业涨跌幅统计（2023.6.26-2023.6.30，单位:%）



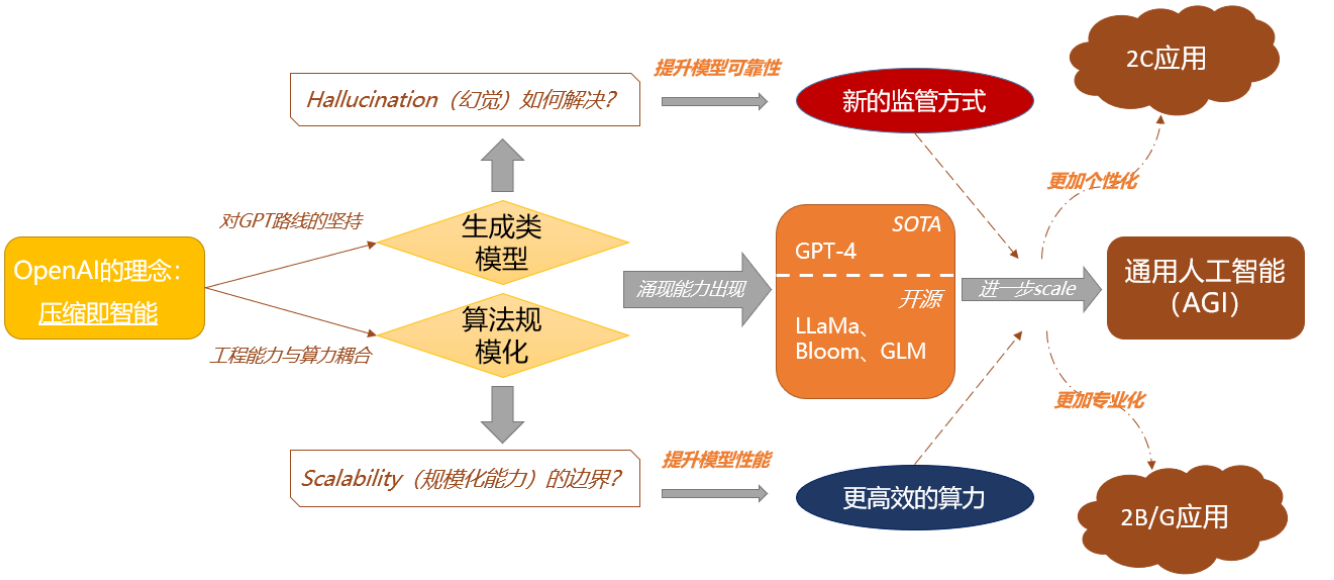
数据来源：Wind，财通证券研究所

2 大模型在 B 端落地已具备基本技术条件

大模型在 B 端商业场景有望边际加速。ChatGPT 的横空出世，已让学术界和工业界充分意识到，OpenAI 对生成类模型（GPT）和算法规模化（Scalability）的两个基础技术路线的持续押注，可能正是打开通用人工智能（AGI）这个终极理想的金钥匙。在这令人兴奋的技术奇点上，全球的企业和开发者都在加速探索如何将 AI 大模型落地到现有的 B 端的商业场景中，包括但不限于基于已有垂域数

据的价值挖掘，更智能化的问答体验，更高效率的自动化办公场景等等。无论是本周恒生电子发布的金融行业大模型 LightGPT、拓尔思发布的拓天大模型（媒体、金融、政务），亦或是上周腾讯云发布的 2B 行业大模型的 MaaS 解决方案，都指向着“大模型的通用能力+垂直领域的专业知识”的结合正在成为大模型在 B 端落地的标准范式，我们有望持续看到大模型在 B 端商业化的边际加速。

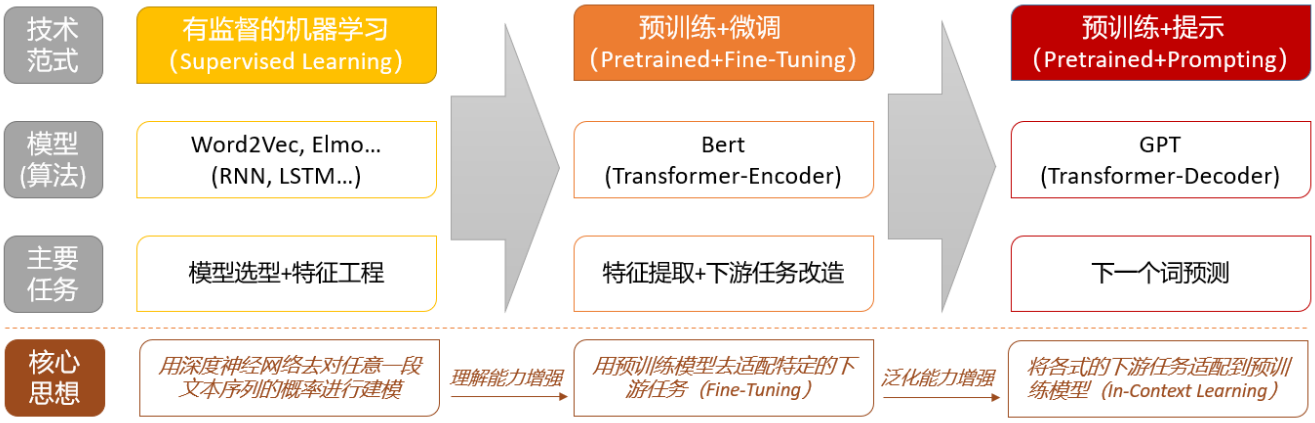
图3.大模型的通用能力助力 B 端商业场景快速落地



数据来源：《Emergent Abilities of Large Language Models》（Jason Wei, Yi Tay 等），财通证券研究所

Fine-Tuning vs In-Context Learning，各有利弊的两种引入垂域知识的方式。如何将企业多年积累的垂域知识嫁接到大模型的通用能力上，是大模型在 B 端商业化落地的核心技术问题。目前开发者主要采用两种思路：1）通过 Fine-Tuning 将垂域知识训练到模型的参数中、2）通过 In-Context Learning 将垂域知识放在模型的 prompt 中。目前这两种方案各有优劣，前者更有利于私有化部署，但训练成本高，且模型会出现原有知识/能力的遗忘；后者使用方便快捷，但受制于目前 LLM 对 prompt 中输入文本长度的瓶颈。通过下图回顾 NLP 领域的三次范式转移可以看到，微调（Fine-Tuning）实际上是更属于 Bert 时代的产物（其必须对下游任务进行改造，且模型更小），而本轮以 OpenAI 的 GPT 系列为代表的大模型更多是强调通过 prompt 去做 In-Context Learning ，即“用下游任务去适配模型”而非 Bert 主张的“用微调的模型去适配任务”。我们认为上述两种方案将会在一定时间内并存，下文中我们也将对一些主流的训练方案做简要介绍。但随着大模型的参数量进一步增大，对模型微调，甚至直接用垂域知识预训练模型的成本会进一步提高。与此同时，随着 embedding model 的持续降本和支持更长的上下文（例如 OpenAI 近期发布的 text-embedding-ada-002），In-Context Learning 有望逐渐成为更经济可靠的方案。

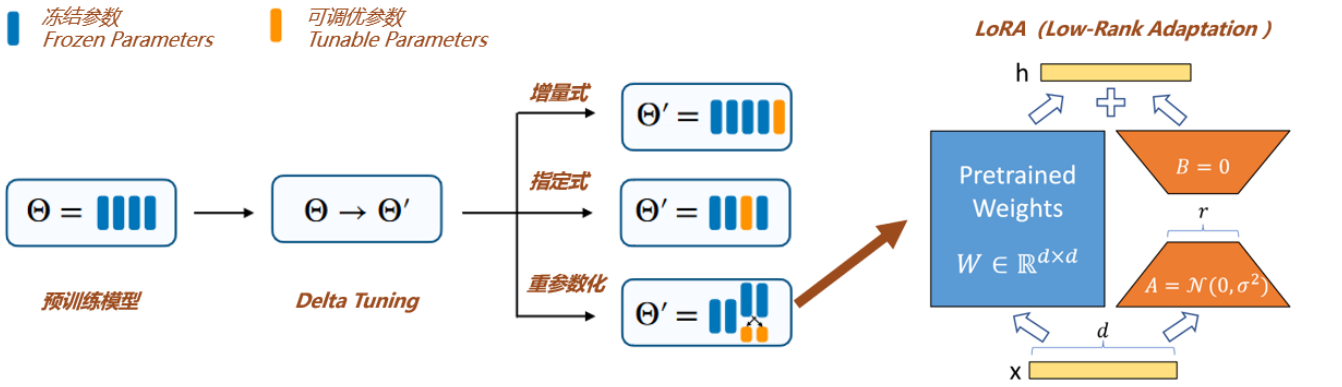
图4.NLP 技术发展的三次范式转移



数据来源：《BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding》（Jacob Devlin, Ming-Wei Chang 等），《Language Models are Few-Shot Learners》（Tom B. Brown, Benjamin Mann 等），财通证券研究所

Delta-Tuning 是对 LLM 参数高效的微调范式。当大模型的规模越来越大时，做全局的微调，即重新训练所有的模型参数无疑会变得愈发不可行，亟需一种参数高效（Parameter-efficient）的新范式。清华与智源研究院在论文中对解决上述问题的方法进行了总结，这些方法本质上都是在尽量不改变原有模型参数的情况下引入一个增量参数（Delta Paremters）进行微调，因此将它命名为 Delta-Tuning。在众多 Delta-Tuning 的实践中，最被开发者广泛关注和使用的，当属微软提出了 LoRA（Low-Rank Adaptation of Large Language Models）。LoRA 的原理是冻结预先训练好的模型参数，在 Transformer 架构的每一层注入一个可训练的低秩矩阵，并在模型训练过程中只训练降维矩阵 A 与升维矩阵 B（下图橙色部分），其本质是基于 LLM 内在的低秩特性，增加旁路矩阵来模拟全参数微调。以微调 175B 参数的 GPT-3 为例，与 Adam 调优的 GPT-3 相比，LoRA 可训练参数量减少了 1 万倍，GPU 内存需求减少了 3 倍，显著降低了微调模型的成本。

图5.Delta-Tuning 是对 LLM 参数高效的微调范式

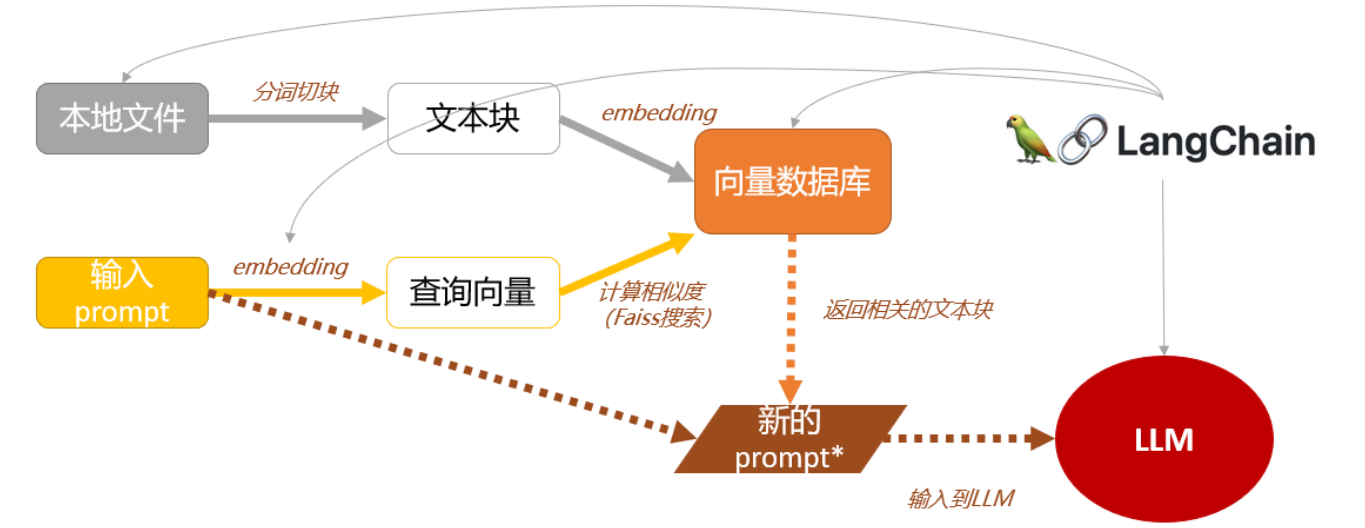


数据来源：《LoRA: Low-Rank Adaptation of Large Language Models》（Edward J. Hu, Yelong Shen 等），《Delta Tuning: A Comprehensive Study of Parameter Efficient Methods for Pre-trained Language Models》（Ning Ding, Yujia Qin 等），财通证券研究所

Langchain+向量数据库打造企业专属知识库问答系统。LangChain 是一套强大的大模型应用开发框架，集成了模型 I/O、数据连接、链、代理、内存、回调等模

块，赋予了大模型：1）数据感知（可连接其他的外部数据源）、2）代理能力（允许大模型与环境互动）。在 LangChain 的帮助下，开发者可以更加便捷的将大语言模型这一“大脑”装上“四肢”，赋予其访问本地文件、调用 API 接口、访问互联网等进阶能力，快速打造知识库问答、聊天机器人、结构化数据分析等功能。因此，使用 LangChain 将大模型与企业的垂域知识库连接（通常以向量数据库的形式），将用户输入的 prompt 在向量数据库中检索最相关的内容，再将返回的内容和输入的 prompt 本身一起成为输入给大模型的最终 prompt，成为了另一种可实现“大模型的通用能力+垂直领域的专业知识”的技术路径。

图6.Langchain+向量数据库打造企业专属知识库问答系统



数据来源：LangChain 官网，财通证券研究所

3 AIGC 加速赋能 B 端用户，注重数据质量与专业性

3.1 恒生电子：发布大模型 LightGPT，构建金融行业 AI 生态

恒生电子将 AI 技术与金融业务 know-how 结合，打造金融行业大模型和全新数智产品，为金融行业应用大模型提供新范式。6 月 28 日，恒生电子和旗下子公司恒生聚源正式发布基于大语言模型技术打造的数智金融新品——金融智能助手“光子”和智能投研平台 WarrenQ，以及底层金融行业大模型 LightGPT。WarrenQ 面向投研人员，“光子”面向金融业务人员，LightGPT 作为底层技术支撑，共同构建人工智能应用生态，助力金融数智化变革。

- **WarrenQ 推出 WarrenQ-Chat 和 ChatMiner，“大模型+数据+工具”打造新一代投研。**WarrenQ 是恒生聚源推出的面向投研投资场景打造的专业一体化投研工具平台，赋能“搜读算写”投研全流程场景，新产品将其升级为“Chat 读算写”，通过智能对话的方式，帮助投研人员提高工作效率。WarrenQ-Chat 利用大模型叠加搜索和聚源金融数据库，通过对话获得金融行情、资讯和数据，且支持原文溯源，还可以生成金融专业报表；ChatMiner 是一款金融文档挖掘器，基于大模型和向量数据库构建，可以根据用户对话指令对指定文档进行快速解读，将关键信息进行有效的整合归纳。

- **金融智能助手“光子”实现业务系统的智能化升级和重构。**在恒生大模型产品技术生态的布局中，光子是串联了“通用工具链+金融插件工具+金融数据+金融业务场景”的智能应用服务。基于金融行业大模型 LightGPT 能力，光子可以为金融机构的投顾、客服、运营、合规、投研、交易等业务系统注入 AI 能力，通过互动问答与智能生成、筛选，成为金融从业人员的助手。

- **更专业、更合规、更轻量，LightGPT 提供 AI 生态底层支持。**LightGPT 在语料和训练方式方面具备优势，使用了超 4000 亿 tokens 的金融领域数据（包括资讯、公告、研报等）和超过 400 亿 tokens 的语种强化数据（包括金融教材、金融百科、政府报告、法规条例等），并作为大模型的二次预训练语料，支持超过 80+金融专属任务指令微调，使 LightGPT 具备金融领域的准确理解能力。LightGPT 可实现金融专业问答、逻辑推理、超长文本处理、多模态交互、代码能力等，为投顾、客服、投研、运营、风控、合规、研发等金融业务场景提供底层 AI 能力支持。

图7.恒生电子发布新产品与大模型

数据来源：恒生电子微信公众号，财通证券研究所

3.2 拓尔思：拓天大模型聚焦优势行业，高质量数据赋能媒体、金融与政务领域

拓天大模型聚焦优势行业，利用自有的高质量数据进行预训练，推出适用于媒体、金融、政务的三大行业大模型。6 月 30 日，拓尔思发布拓天大模型，该模型拥有内容生成、多轮对话、语义理解、跨模态交互、知识型搜索、逻辑推理、安全合规、数学计算、编程能力和插件扩展十大基础能力，且具有中文特性增强的可控生成技术、融合搜索引擎的生成结果可信核查、融合稠密向量的跨模态能力加强以及支持外界知识及时更新四大创新点。依托于在自然语言处理领域 30 年的技术积淀，以及优质客户基础和行业 know-how，公司推出适用于媒体、金融、政务的行业大模型。未来，公司还将陆续推出网络舆情、公安、知识产权、法律、审计等行业大模型。

- **媒体大模型主要覆盖内容生产智能助手、新一代搜索与推荐、多模态传播与服务三大业务场景。**基于党媒、商媒、微信公号等合规数据生成预训练模型，再面向通用 NLP 任务进行模型训练，SFT（监督指令微调）后可应用于真实与及时的新闻生产、高质量内容生成以及内容安全具有主流意识形态的知识服务。

图8.拓天媒体行业大模型的训练、微调与对齐

数据来源：2023 拓尔思拓天大模型成果发布会，财通证券研究所

图9.拓天媒体行业大模型应用领域

数据来源：2023 拓尔思拓天大模型成果发布会，财通证券研究所

- **金融大模型主要覆盖智能风控、智能客服、智能投研、自动业务批处理等业务场景。**在金融领域已有国外 BloombergGPT、国内恒生电子 Light - GPT、农行 ChatABC 等大模型的推出，B 端金融用户存在广泛需求；在面对安全性、可解释性、适配性、实时性问题仍存在挑战。拓天大模型+金融知识图谱+工具+数据的融合与互补将为金融领域多业务场景提质增效。

图10.拓天金融大模型技术能力与应用领域

数据来源：2023 拓尔思拓天大模型成果发布会，财通证券研究所

- **政务大模型主要覆盖公文辅助写作、政策大脑和新一代政务互动等业务场景。公司主要聚焦数字政府门户、政务舆情、产业招商、金融监管和数字机关等领域。**公司积累了大量优质头部用户和高质量数据，并了解涉及政务场景的需求和痛点。拓天政务大模型可面向智库、研究机构、高校、投资人员，以及媒体和政府制定政策的人员进行政策研读；协助中小微企业了解政府出台的扶持政策，便于企业享受补贴、减免税等福利。

图11.拓天政务大模型助力政务行业应用质效提升

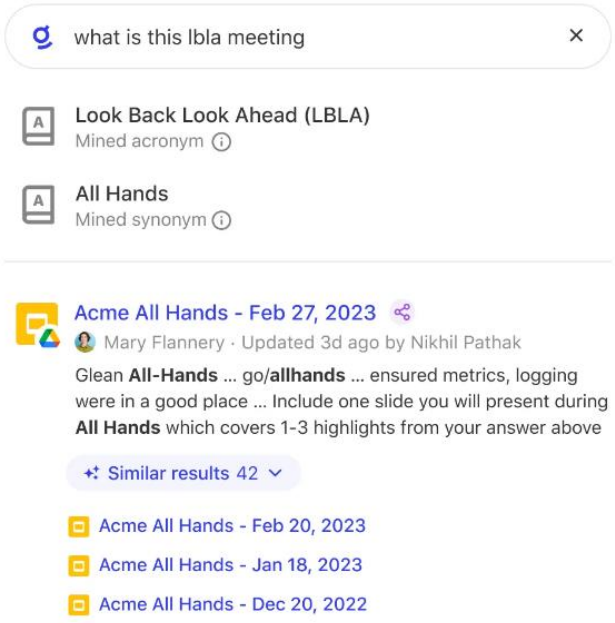
数据来源：2023 拓尔思拓天大模型成果发布会，财通证券研究所

3.3 Glean：打通 SaaS 化应用，成为企业场景入口

Glean 定位基于 AI 的企业搜索与知识管理平台，协助用户跨应用、个性化搜索，能够更快、更准确地找到企业内部的知识和数据。Glean 的跨应用搜索相当于在所有 SaaS 产品之上构建通用平台，用户不需要再逐一打开应用，而是在 Glean 上就可以查到企业数据，完成部分高频工作。Glean 掌握企业内部知识的同时又了解每位员工的偏好，进而形成了企业整体与员工个体之间的沟通桥梁，为办公提质增效。Glean 产品主要功能包括：

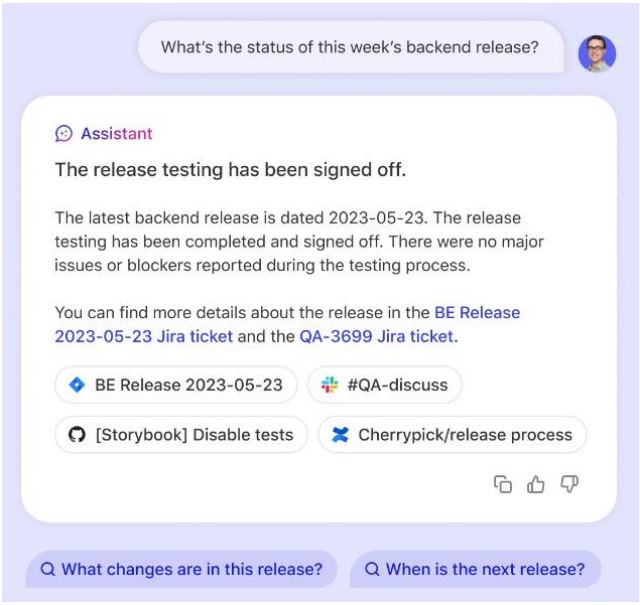
- **搜索：**Glean 结合客户公司的语言和背景等为每一个客户自动训练定制化的 AI 模型，并构建有关公司人员、内容和互动的知识图谱；API 支持跨应用程序搜索，因此用户可以通过 Glean 搜索公司所有应用程序的内容。
- **知识管理：**用户可以分享和整合相关的文档或链接、使用全新简短形式的 URL 进行界面跳转等。
- **工作空间：**在首页个性化显示用户在工作过程中常用的一些功能模块，比如公司公告、公司文件资源、日历和活动等。
- **安全保障：**用户可以通过运行 DLP 报告以发现过度暴露的敏感内容；Glean 会进行用户访问审查以执行最小特权原则，也对所有的数据都进行了安全加密；用户可以自行选择本地或者云端部署等。

图12.Glean 产品的搜索功能



数据来源：Glean 官网，财通证券研究所

图13.Glean 产品的 AI 助手



数据来源：Glean 官网，财通证券研究所

4 投资建议

AI 大模型赋能下游应用，C 端标准化工具类产品有望率先享受产业红利，建议关注金山办公、万兴科技、同花顺、科大讯飞、福昕软件等。

AI 在 B 端加速落地，具备细分行业数据与客户资源卡位的企业有望优先受益，建议关注恒生电子、拓尔思等。

算力是 AI 大模型产业化落地的必备环节，建议关注 AI 服务器相关厂商以及国产 AI 芯片厂商：浪潮信息、中科曙光、优刻得、紫光股份、海光信息、寒武纪、拓维信息、神州数码以及在向量数据库及垂直大模型领域有技术优势的星环科技等

生成式 AI 的诞生促使网络安全防护迎来范式转移，AI+安全建议关注：启明星辰、美亚柏科、三未信安、深信服、安恒信息、奇安信、中孚信息、中新赛克等。

5 风险提示

AI 技术迭代不及预期的风险：若 AI 技术迭代不及预期，NLP 模型优化受限，则相关产业发展进度会受到影响。

商业化落地不及预期的风险：ChatGPT 盈利模式尚处于探索阶段，后续商业化落地进展有待观察。

政策支持不及预期风险：新行业新技术的推广需要政策支持，存在政策支持不及预期风险。

全球宏观经济风险：垂直领域公司与下游经济情况相关，存在全球宏观经济风险。

信息披露

● 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，并注册为证券分析师，具备专业胜任能力，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。本报告清晰地反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响，作者也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。

● 资质声明

财通证券股份有限公司具备中国证券监督管理委员会许可的证券投资咨询业务资格。

● 公司评级

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。

● 行业评级

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。

● 免责声明

本报告仅供财通证券股份有限公司的客户使用。本公司不会因接收人收到本报告而视其为本公司的当然客户。
本报告的信息来源于已公开的资料，本公司不保证该等信息的准确性、完整性。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他投资标的邀请或向他人作出邀请。
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价格、价值及投资收入可能会波动。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。
本公司通过信息隔离墙对可能存在利益冲突的业务部门或关联机构之间的信息流动进行控制。因此，客户应注意，在法律许可的情况下，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。在法律许可的情况下，本公司的员工可能担任本报告所提到的公司的董事。
本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。在任何情况下，本公司不对任何人使用本报告中的任何内容所引致的任何损失负任何责任。
本报告仅作为客户作出投资决策和公司投资顾问为客户提供投资建议的参考。客户应当独立作出投资决策，而基于本报告作出任何投资决定或就本报告要求任何解释前应咨询所在证券机构投资顾问和服务人员的意见；