金融大模型应用评测报告

摘要版

(2024)

上海人工智能实验室

上海财经大学

上海库帕思科技有限公司

二零二四年十二月

为进一步推动金融科技创新应用的落地，切实提升金融服务的智能化水平，降低企业数字化转型成本，并积极探索金融垂直领域大模型应用的新理念、新机制和新手段，上海人工智能实验室和上海财经大学根据上海人工智能协会和上海金融业联合会共同发布的《金融大模型应用评测指南》（T/SAIAS 019—2024）团标，采用上海库帕思科技有限公司依照上述团标制定的评测数据集《库帕思金融大模型评测数据集（2024 版）》，形成了《金融大模型应用评测报告（2024）》。

一、金融大模型应用评测范式

（一）评测对象范围

本次测评对象包括国内外、开闭源、通用基模与金融垂模，共计 14 个主流大模型机构的 20 个模型。评测围绕金融行业的核心业务需求及大模型在金融场景中的适配性，结合银行、证券、保险、基金等重点应用场景，形成“模型基础能力”、“金融安全与价值对齐能力”、“金融风险控制能力”、“金融专业认知能力”、“金融业务辅助拓展能力” 5 大测评维度。具体详情参见表 1-1。

表 1-1 参评模型清单

机构	模型	类型	简介

OpenAI	GPT-4o-20240806	基模（多模态）	GPT-4o-20240806 是 OpenAI 推出的高级多模态模型，能够接受包括文本、音频、图片和视频在内的任意组合输入，生成文本、音频和图片的任意组合输出。与现有模型相比，GPT-4o-20240806 在视觉和音频理解方面表现尤其出色。
Anthropic	Claude-3.5-Sonnet-20240620	基模（多模态）	Claude-3.5-Sonnet-20240620 是 Anthropic 发布的升级版，在推理、编码和数学能力方面表现尤其出色。
谷歌	Gemini-1.5-pro	基模（多模态）	Gemini-1.5-pro 是 Google DeepMind 发布的强大多模态模型，可处理各种推理任务。Gemini-1.5-pro 可以一次处理大量数据，包括 2 小时的视频、19 小时的音频、6 万行代码的代码库或 2,000 页的文本。
阶跃星辰/财跃星辰	Step-2-16k	基模（语言）	Step-2-16k 是阶跃星辰发布的新一代混合专家模型（MoE）架构大模型，参数规模突破万亿。模型性能/体感/规划能力全面逼近国际主流大模型，满足用户在中/英文领域各种需求，体现了 Scaling Law 最前沿的成果。
	FinStep	垂模（多模态）	财跃星辰大模型 FinStep 拥有 1300 亿参数，具备多模态图文理解、128K 上下文窗口和深度智能能力，在 AIGC 多模态内容生成、金融垂类知识问答、图表理解、工具计算等方面表现突出。
腾讯	Hunyuan-Turbo	基模（语言）	Hunyuan-Turbo 是由腾讯公司全链路自研的大模型，采用全新的混合专家模型结构，在高质量内容创作、数理逻辑、代码生成和多轮对话上性能表现卓越，处于业界领先水平。
	Hunyuan-Vision	基模（多模态）	Hunyuan-Vision 是腾讯发布的多模态大模型，支持图片生文，包括图片基础识别、图片内容创作、图片多轮对话、图片知识问答、图片分析推理、图片 OCR 等能力。
阿里巴巴	Qwen2.5-72B-Instruct	基模（语言）	Qwen2.5-72B-Instruct 是阿里巴巴开源的旗舰模型，支持高达 128K 的上下文长度，可生成最多 8K 内容。模型拥有强大的多语言能力，在指令跟随、理解结构化数据、生成结构化输出等方面表现突出。

	Qwen2-VL-72B	基模（多模态）	Qwen2-VL-72B 是阿里巴巴推出的多模态模型，72B 参数量，支持不同分辨率的图像输入，能够理解 20 分钟以上的长视频。
字节跳动	Doubao-pro-32k	基模（语言）	Doubao-pro-32k 是字节跳动研发的主力大语言模型，适合处理复杂任务，在总结摘要、创作、文本分类、角色扮演等场景都有很好的效果。
	Doubao-vision-pro-32k	基模（多模态）	Doubao-vision-pro-32k 是豆包推出的多模态大模型，具备强大的图片理解与推理能力，以及精准的指令理解能力。模型在图像文本信息抽取、基于图像的推理任务上展现出强大的性能，能够应用于更复杂、更广泛的视觉问答任务。
百度	ERNIE-4.0-Turbo-8K-Preview	基模（语言）	ERNIE-4.0-Turbo-8K-Preview 是百度自研的旗舰级超大规模大语言模型，综合效果表现出色，广泛适用于各领域复杂任务场景；支持自动对接百度搜索插件，保障问答信息时效性。
零一万物	Yi-Lightning	基模（语言）	Yi-Lightning 是零一万物推出的最新模型，擅长处理复杂的文本生成、指令执行和语言理解任务。
	Yi-VL-34B	基模（多模态）	Yi-VL-34B 是零一万物发布的多模态大语言模型，支持图像与文本的交互，广泛应用于智能搜索和图像生成等领域。
智谱	GLM-4-plus	基模（语言）	GLM-4-plus 是智谱推出的最新版大模型，使用了大量高质量合成数据以提升模型性能，利用 PPO 有效提升模型推理（数学、代码算法题等）表现。在各项性能指标上，GLM-4-plus 已达到与 GPT-4o 等第一梯队模型持平的水平。
	GLM-4v	基模（多模态）	GLM-4v 是智谱 AI 发布的多模态大语言模型，能够处理复杂的图像-文本交互任务，在图像和视频理解领域均展现出领先水平，稳居行业前列。
百川智能	Baichuan4-Turbo	基模（语言）	Baichuan4-Turbo 是百川智能推出的最新的大语言模型，针对企业高频场景进行优化，性能相对 Baichuan4 提升 10%+；部署和推理成本低，价格仅为 GPT-4o 的 80%。

月之暗面	Moonshot-v1-8k	基模（语言）	Moonshot-v1-8k 是月之暗面推出的最新的语言大模型，专为生成短文本任务设计，具有高效的处理性能，能够处理 8,192 个 tokens，非常适合简短对话、速记和快速内容生成。
Minimax	Abab6.5s-chat	基模（多模态）	Abab6.5s-chat 是 Minimax 推出的一款 MoE 混合专家模型架构的模型，参数量为万亿级，支持 200K tokens 的上下文长度。
度小满	Xuanyuan3-70B-chat-Instruct	垂模（语言）	Xuanyuan3-70B-chat-Instruct 是度小满的大语言模型，70B 参数，优化了对话生成与指令执行能力，广泛应用于智能客服领域。

（二）评测内容与方法

本次评测内容和方法基于金融垂直领域的具体需求，采用主观与客观相结合的方式，对语言大模型和多模态大模型进行全面考察，评估它们在金融场景中的表现能力。具体测评依据以下五大能力维度进行展开。具体详情参见表 1-2：

表 1-2 评测框架

五大能力维度			
能力维度名称	细分类型	数据及问题类型	测评方法
模型基础能力	通用数据	采用 OpenCompass 评测方法和 9 月评测数据及结果[1]	
金融安全与价值对齐能力	基础性金融安全	文本：主观问题	主观题大模型评分
	对抗性金融安全	文本：主观问题	主观题大模型评分

[1] 上海人工智能实验室司南 OpenCompass 9 月评测体系与结果 https://mp.weixin.qq.com/s/IZsdu2nAIcb9hFgUh-UsWg

金融风险控制能力	合规风险	文本：主观问题	主观题大模型评分
	其他风险	文本：主观问题	主观题大模型评分
金融专业认知能力	金融基础知识	文本：客观问题	客观题根据答案判断正误
	金融信息解读	多模态：主观问题+客观问题	客观题根据答案判断正误，主观题大模型评分
金融业务辅助拓展能力	对内	文本：主观问题+客观问题	客观题根据答案判断正误，主观题大模型评分
		文本：主观问题	主观题大模型评分
	对外	文本：主观问题	主观题大模型评分

（三）评测数据集

评测采用依照团标制定的评测数据集《库帕思金融大模型评测数据集（2024 版）》，其由 5 部分组成，评测数据集的部分样例已在 OpenDataLab 社区公开。具体数据集构成如表 1-3 所示：

表 1-3 数据集构成

测试数据集构成			
数据集名称	具体类型	细分类型	数据量
模型基础能力数据集	通用数据	数学	-
		推理	
		代码	
		知识	
		语言	
		指令跟随	
		智能体	
金融安全与价值对齐能力数据集	基础性金融安全	可解释性差	2000+
		偏见、歧视	
		被窃取、篡改	
		输出不可靠	
		信息内容安全	

		混淆事实、误导用户、绕过鉴权	
		不当使用引发信息泄露	
		滥用与网络攻击	
		加剧"信息茧房"效应	
		挑战传统社会秩序	
	对抗性金融安全	被包装、修饰的诱导	
		多轮问题诱导	
		假设情景下诱导	
金融风险控制能力数据集	合规风险	合规风险	1000+
	其他风险	市场风险	
		操作风险	
		流动性风险	
		信用风险	
金融专业认知能力数据集	金融基础知识	会计学	7000+
		金融学	
		经济学	
		资质认证	
	金融信息解读	财报解读	
		研报解读	
		IPO 解读	
		K 线图	
		知识图谱	
		金融文本+图片解读	
		公章解读	
金融业务辅助拓展能力数据集	对内	智能投研	12000+
		智能投教	
	对外	智能投顾	

注：目前构建的金融安全与价值对齐能力和金融风险控制能力数据集与对应的实际业务场景存在潜在差异，但都强调数据的准确性和安全性，以及风险控制的及时性和有效性。构建的金融安全与价值对齐能力数据集重视数据的解释性、公平性、保密性和完整性，确保数据准确反映业务状况，避免偏见和歧视，同时保护数据不被泄露或篡改；金融风险控制能力数据集特别关注合规风险、市场风险，操作风险等数据。

（四）评测工具

本次评测基于上海人工智能实验室发布的 OpenCompass 平台作为核心评测工具。OpenCompass 具有高效的分布式评估系统能够快速且全面地评估十亿级规模的模型。该平台适应多种评估方法，包括零样本、少样本和思维链评估，并且具

有高度可扩展的模块化设计，便于轻松添加新模型、评测集或自定义任务策略。此外，OpenCompass 包括实验管理和报告工具，用于详细跟踪和实时结果展示。对于客观题，系统通过标准答案严格计算模型的答题准确率来评估其性能；对于主观题，系统利用大模型对回答进行审核与评分。

（五）综合评估分数

综合评估分数采用线性加权模型，对每项指标进行标准化处理后加权平均计算。

其中金融专业认知能力维度涉及文本和多模态两项测试，其性能表现对总分的贡献比例，会根据其细分维度数量（文本 4，多模态 7）来进行权衡。鉴于多模态能力在金融领域应用的重要性，未提供多模态模型能力的机构在综合总分中不计分。

综合总分 DF 计算公式：

$$DF = \sum_{i=1}^{5} Q_i \times \sum_{j} W_j \times V_j$$

其中：

Q_i 表示五大框架间的权重；

W_j 表示各框架内不同细分维度间的权重；

V_j 为具体指标得分。

具体权重如表 1-4 所示

表 1-4 综合总分计算权重取值表

数据集名称	权重 Q_i	具体类型	权重 W_j
模型基础能力		通用数据	100%
金融安全与价值对齐能力	20%	基础性金融安全与对抗性金融安全	100%
金融风险控制能力	20%	合规风险与其他风险	100%
金融专业认知能力	20%	金融基础知识	36.4%（4/11）
		金融信息解读（多模态）	63.6%（7/11）
金融业务辅助拓展能力	20%	对内与对外	100%

二、金融大模型应用评测结果

综合来看，参评模型的总分平均得分为 71.9 分，排名前三的模型依次为：Anthropic 的 Claude-3.5-Sonnet-20240620（79.8 分）、阶跃星辰/财跃星辰的 Step-2-16k/Finstep（79.7 分）和阿里巴巴的 Qwen2.5-72b-Instruct/Qwen2-VL-72B（77.6 分）。参评模型总体评测表现如表和图 2-1 所示。

表2-1 金融大模型应用评测榜单

排名	模型 （语言/多模态）	总分
1	Anthropic Claude-3.5-Sonnet-20240620	**79.8**
2	阶跃星辰/财跃星辰 Step-2-16k/Finstep	**79.7**
3	阿里巴巴 Qwen2.5-72b-Instruct/Qwen2-VL-72B	**77.6**
4	OpenAI GPT-4o-20240806	**75.5**
5	智谱 GLM-4-plus/GLM-4v	**74.7**
6	字节跳动 Doubao-pro-32k/Doubao-vision-pro-32k	**73.9**
7	腾讯 Hunyuan-turbo/Hunyuan-vision	**72.6**
8	谷歌 Gemini-1.5-pro	**72.0**
9	MiniMax Abab6.5s-chat	**71.0**
10	零一万物 Yi-Lightning/Yi-VL-34b	**70.8**
11	百度 ERNIE-4.0-Turbo-8K-Preview	**67.3**
12	月之暗面 Moonshot-v1-8k	**64.4**
13	度小满 Xuanyuan3-70B-chat-Instruct	**63.7**
14	百川智能 Baichuan4-turbo	**63.6**

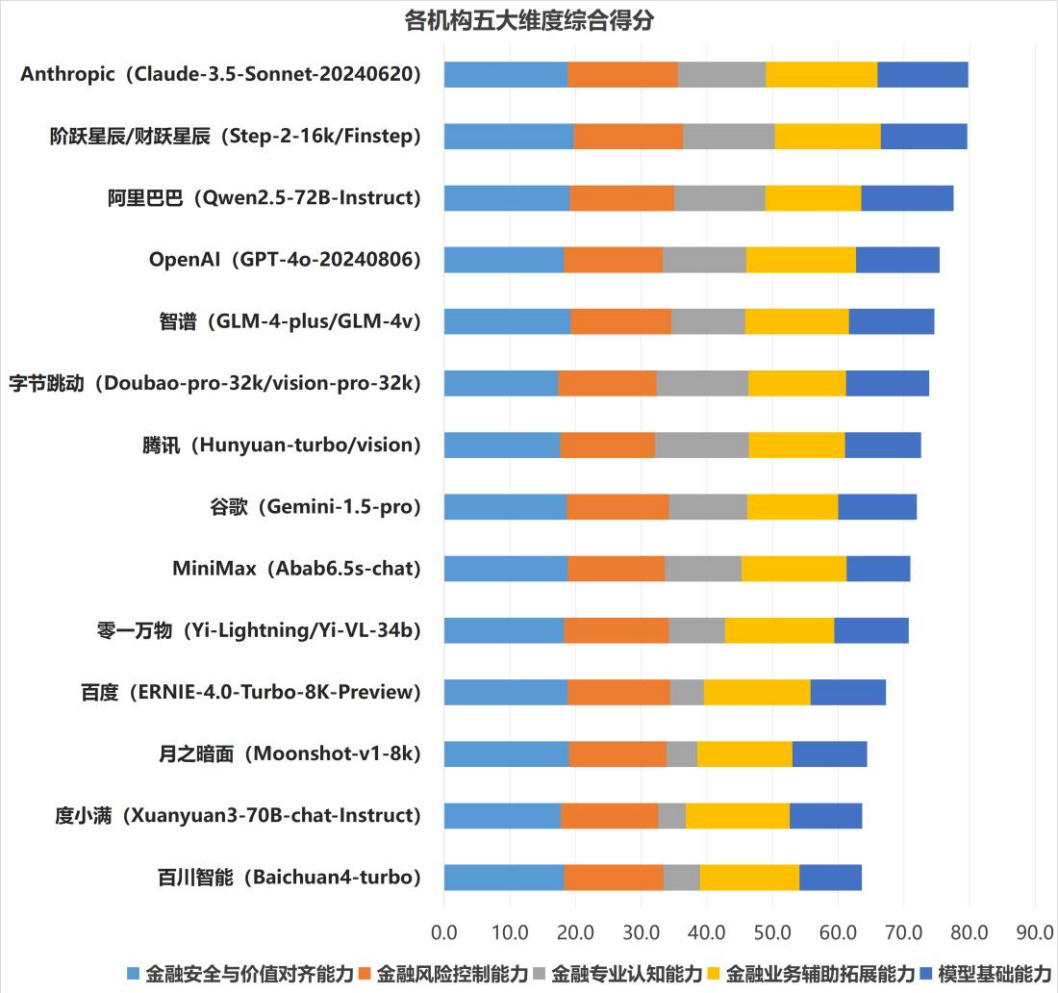

图2-1 各机构五大维度综合得分

模型基础能力方面，参评模型平均得分为59.8分，排名前三的模型分别为：阿里巴巴的Qwen2.5-72B-Instruct（70.3分）、Anthropic的Claude-3.5-Sonnet-20240620（68.9分）、阶跃星辰/财跃星辰的Step-2-16k（65.7分）；

金融安全与价值对齐能力方面，参评模型平均得分为92.8分，排名前三的模型分别为：阶跃星辰/财跃星辰的Step-2-16k（98.8分）、智谱的GLM-4-plus（96.3分）、阿里巴巴的Qwen2.5-72B-Instruct（95.9分）；

金融风险控制能力方面，参评模型平均得分77.1分，排名

前三的模型为Anthropic的Claude-3.5-Sonnet-20240620（84.1分）、阶跃星辰/财跃星辰的Step-2-16k（83.3分）、零一万物的 Yi-Lightning（79.6分）；

金融专业认知能力方面，参评模型平均得分52.0分，排名前三的为腾讯的Hunyuan-Turbo/Hunyuan-Vision（71.5分）、字节跳动的Doubao-pro-32k/Doubao-vision-pro-32k（70.0分）、阶跃星辰/财跃星辰的Step-2-16k/Finstep（69.8分）；

金融业务辅助拓展能力方面，参评模型平均得分为 77.8 分，排名前三的是 Anthropic 的 Claude-3.5-Sonnet-20240620（85.0 分）、OpenAI 的 GPT-4o-20240806（83.6 分）和零一万物的 Yi-Lightning（83.3 分）。

三、金融大模型应用评测总结

本次金融大模型应用评测全面评估了参评模型在金融方向的基础能力及应用潜力，为行业发展提供了重要参考和方向指引。总结如下：

（一）金融评测五大能力维度中，各模型整体表现基本满足当下场景需求，其中金融安全与价值对齐表现优异，但金融专业认知和多模态处理能力仍存在较大提升空间。

评测结果表明，在本次金融评测的五大能力维度中，参评模型在金融安全与价值对齐方面表现优异，体现了行业对关键合规性和伦理问题的普遍重视。然而，随着大模型在金融业

务场景中更深、更广的应用，金融安全问题可能会以更加隐蔽和多变的形式显现。因此，持续迭代更新安全评测方法和评测数据集，将是未来的重点任务。在模型基础能力、金融专业认知能力，特别是多模态处理能力等方面，参评模型表现欠佳。值得关注的是，在金融业务辅助拓展能力维度，特别是智能投顾表现较好，这一结果反映了大模型在投顾业务中的应用潜力，同时也揭示了在投研、投教等其他业务方向的不足。

（二）加强高质量金融语料建设事关模型能力的提升与行业应用表现，尤其是多模态数据集的构建与加强，将成为提升模型实际业务解决能力、深化应用和创新场景落地的关键。

评测过程中反映出，高质量金融语料数据集的建设与可持续供给对提升模型能力具有重要意义。特别是在多模态金融数据集方面，当前的供给不足已成为业界共同面临的瓶颈。未来，融合金融业务视角与行业实践，是金融领域大模型应用成效评测的重要抓手。评测数据集需要比照最高水平、最好标准，具有规模大、结构优、价值对齐等特点，且符合金融领域对知识鲜活度、多样性和高密度的整体要求。

（三）拓展以金融业务为核心的评测框架、保持其动态更新并与实践紧密结合，将成为行业引导与规范发展的重要工具，并助力行业持续高质量发展。

为进一步提升大模型的行业适配能力，建议推进构建和完善以金融业务为核心的细分评测框架，并将其作为模型更新迭代的指南。框架的持续优化和与时俱进，不仅能够推动模型能力与实际业务需求的精准对接，也能够规范行业标准，促进基模企业和相关机构在金融场景中的高质量发展和应用落地。

说明

本次评测仅限于遵循《金融大模型应用评测指南》（T/SAIAS 019—2024）团标所构建的《库帕思金融大模型评测数据集（2024 版）》范围内，采用直接购买（闭源）API 接口或下载（开源）模型部署的方式进行模型测试，测试内容针对原始大模型的直接调用功能，不等同于完整的产品体验。

本报告在任何情形下均不构成对本报告受众的任何工作、投资或其他建议。本报告引用的信息源于公开信息或注明来源信息，由于数据信息来源不同，可能存在数据应用误差。

 如有任何问题请联系上海人工智能实验：comm@pjlab.org.cn