arXiv:2602.03130v1 [cs.CV] 3 Feb 2026

FinMTM：用于金融推理与智能体评估的多轮多模态基准

张晨曦[1, 2*] **Ziliang Gan**[1, 3*] **朱丽云**[1*] **YouweiPang**[4] **张青**[5]
Rongjunchen Zhang[1♠]
[1]同花顺研究院 [2]武汉大学 [3]浙江大学 [4]南洋理工大学 [5]上海应用技术大学
通讯作者：
zhangrongjunchen@myhexin.com

摘要

由于金融图表格式特殊，且需要基于大量专业知识进行推理，金融领域对视觉语言模型（VLM）构成了巨大挑战。然而，现有金融基准大多采用单轮形式，且依赖有限的问题格式，因而难以在真实应用场景中进行全面评估。为弥补这一空白，我们提出 FinMTM，这是一个多轮多模态基准，从数据和任务两个维度拓展了多样性。在数据方面，我们整理并标注了 11,133 组中英双语金融问答对，内容基于金融视觉材料，包括蜡烛图、统计图和报告图表。在任务方面，FinMTM 涵盖单选题和多选题、多轮开放式对话，以及基于智能体的任务。我们还设计了针对不同任务的评估方案，包括适用于多选题的集合重叠评分规则、结合轮次级与会话级得分的加权方法（用于多轮对话），以及将规划质量与最终结果相结合的综合指标（用于智能体任务）。对 22 个视觉语言模型的广泛实验评估揭示了它们在细粒度视觉感知、长上下文推理和复杂智能体工作流方面的局限。代码见
h
t
tps://github.com/HiThink-Research/FinMTM。

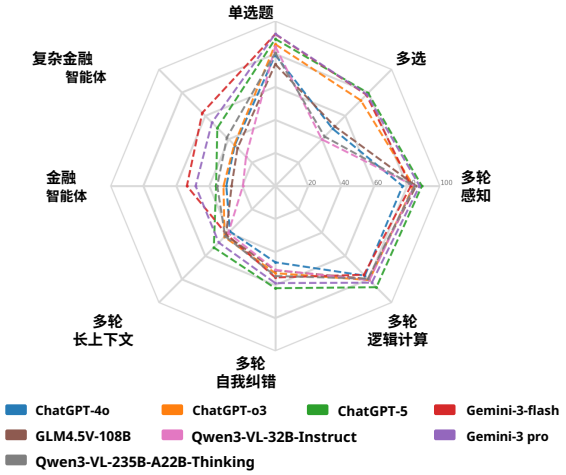

图 1：不同维度整体性能的雷达图。

1 引言

大语言模型（LLMs）和视觉语言模型（VLMs）的迅速发展，深刻影响了各行各业（Brown 等人，2020；OpenAI，2023；Huang 等人，2023）。金融领域拥有大量专业图表，因此尤其适合应用视觉语言模型。复杂且专业的金融图表，能够严格检验模型理解图表的能力以及细粒度视觉感知能力。此外，金融分析（Zhu 等人，2025b）、量化因子挖掘（Li 等人，2025b）、股票预测（Lin 等人，2025）和投资决策等复杂任务，要求模型将领域知识与推理能力相结合，从海量数据集中检索并综合关键信息，从而得出专业见解。此外，基于智能体的任务要求模型进行深入分析，并自主调用工具来收集信息、得出最终结论（Wu 等人，2023；Li 等人，2024）。总的来说，这些任务给当前的视觉语言模型带来了重大挑战。

最近，研究人员提出了一些金融基准，例如 MME-Finance （Gan 等人，2025）、FinMME（Luo etal.，2025）、FinTMM-Bench（Zhu 等人，2025a）、FinRAGBench-V（Zhao 等人，2025）和 VisFinEva（Liu 等人，2025）。尽管这些基准在一定程度上评估了视觉语言模型的感知和推理能力，但仍存在以下三个主要局限。

首先，大多数基准主要通过单轮设置评估视觉语言模型，采用二元

*共同贡献。♠通讯作者。

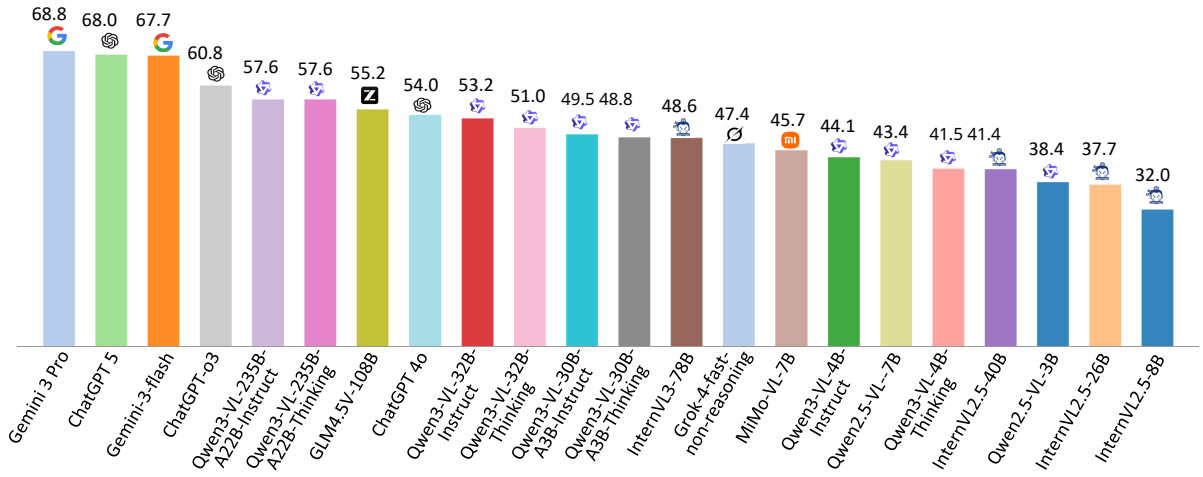

图2：领先的视觉语言模型在FinMTM基准上的通用金融能力比较。最终得分为三项任务得分的平均值：客观题、开放式问题和金融智能体。

选择题或简短的自由形式回答。这种形式往往会过度简化复杂的金融推理，无法充分体现模型的分析深度。其次，当前框架缺乏对综合能力的系统评估，尤其是在多轮交互、多源信息融合、长上下文推理和知识密集型金融分析等方面。第三，现有基准在很大程度上忽略了金融智能体行为，例如迭代使用工具、策略路径规划和结果综合；这些都是现实金融工作流程中的关键组成部分。综上，这些局限引出了一个关键问题：在静态的单轮基准中表现出色的视觉语言模型，能否在复杂的交互式多轮场景以及涉及长时程规划和模型上下文协议（MCP）交互的金融智能体任务中持续保持良好表现?

为弥合这些差距，我们推出**FinMTM**，这是一个全面的金融多模态基准，旨在真实场景中全面评估感知、推理和智能体能力。如图1和图2所示，我们在FinMTM上对多种领先的视觉语言模型进行了广泛的比较实验。结果显示，不同场景下的性能差距明显。我们的主要贡献总结如下：

• **大规模双语金融基准**。我们构建了一个包含超过 10,000 个样本的中英双语基准，基于专业金融视觉材料（例如蜡烛图、统计图表和报告图表），涵盖单选/多选问答、多轮次开放式对话以及金融智能体任务。

• **多维评估框架**。我们设计了特定任务指标和评估流程，包括针对选择题的严格基于集合的评分，以及用于开放式任务和智能体任务的双规则 LLM 评审协议（轮次级能力评分和会话级检查清单）。

• **全面评估与洞见**。我们在 FinMTM 上评估了 16 个开源视觉语言模型和 6 个专有视觉语言模型，并发现细粒度视觉定位、复杂金融推理、长上下文一致性以及工具增强型智能体规划方面仍存在持续的瓶颈。

2 相关工作

2.1 金融多模态基准

为系统评估下一代视觉语言模型在金融领域中的视觉语言能力，近年来相继出现了一系列金融多模态基准。总体演进轨迹已从单轮图表感知发展至知识密集型推理。早期具有代表性的工作，如 MME-Finance（Gan 等人，2025）、FinMME（Luo 等人，2025）和 CFBenchmark-MM（Li 等人，2025a），主要基于图表和报告截图构建单轮问答任务。这些基准侧重于对图表解读、基础数值推理以及初步“感知—推理—认知”能力进行分层评估。然而，其中的交互范式

基准	Vol.	任务类型					高级能力		
		ST	SC	多选题	MO	多轮长上下文问答	记忆能力	多步骤计算	FT
CFBenchmark (Li 等人, 2025a)	9k	✓	✓	✗	✗	✗	✗	✗	✗
FinCriticalED (He 等人, 2025)	0.5k	✗	✗	✗	✗	✗	✗	✗	✗
FinMME（Luo 等人，2025）	11k	✓	✓	✓	✗	✗	✗	✗	✗
FinMR（Deng 等人，2025）	3.2k	✓	✗	✗	✗	✗	✗	✗	✗
MME-Finance（Gan 等人，2025）	2k	✓	✗	✗	✗	✗	✗	✗	✗
FinRAGBench（Zhaoetal.，2025）	1.3万	✓	✗	✗	✗	✗	✗	✗	✗
VisFinEval（Liu 等人，2025）	15k	✓	✓	✓	✗	✗	✗	✗	✗
FinMTM（本文提出）	**11k**	✓	✓	✓	✓	✓	✓	✓	✓

表 1：本文提出的 FinMTM 与现有多模态金融基准的比较。符号：✓ 支持；✗ 不支持。MC：多选题；MO：多步骤开放式问题；MLC：多轮长上下文问答；Mem：记忆能力；MS：多步骤计算；FT：金融工具使用。

早期这些工作基本局限于单轮问答，未能涵盖涉及长文档、多页引文或工具链集成的场景。随后，FinMR（Deng 等人，2025）将重点转向知识密集型高阶推理。通过将结构化推理链引入复杂估值和风险分析任务，该基准强化了对专业领域知识和多步推理的刻画。此外，FinRAGBench-V（Zhu 等人，2025a）将多页 PDF 文档及其中嵌入的表格和图表纳入评估框架，明确衡量多模态检索增强生成在关键数值和时间数据方面的事实一致性。与此同时，Vis-FinEval（Liu 等人，2025）依据前台、中台和后台业务流程组织了约 15k 个样本，强调具备场景意识的理解和整体决策支持。

如表 1所示，现有金融视觉语言模型基准主要关注单轮任务，难以充分反映多轮可靠性或长上下文推理能力。迭代式工具使用和战略规划等关键智能体能力也尚未得到充分评估。为弥补这些不足，我们提出了 FinMTM，这是一个综合框架，用于评估客观问答、多轮对话和工具增强型智能体任务。

3 FinMTM基准

3.1 任务定义

我们将所需的金融能力归纳为三类具有代表性的评测任务，即客观题、开放式问题和金融智能体。具体而言，客观题包括单选题（SC）和多选题（MC）两种形式。开放式问题设置为多轮设置，并进一步分为四个子任务：理解、计算、自我纠错和记忆。对于金融智能体任务，我们提供了一套MCP工具，以支持工具调用和规划流程。这些任务共同构成了一个渐进式评估框架，涵盖从基础感知、深入的分析性理解，到交互式决策和智能体行为的不同层次。所有任务均提供中文和英语版本。

3.2 数据集构建

数据整理。 为构建 FinMTM，我们从真实的行业工作流程中收集了多样化的金融图像与文档，数据来自四个关键渠道：宏观经济与行业研究报告、个股研究报告、公开可得的公司财务数据，以及金融媒体图像。借助美国和中国股票市场的真实信息，我们的数据集能够支持稳健、贴近现实的基准评测。

客观题生成。 我们首先整理出一个包含1982个单轮问答对的高质量种子集，每个样本均经领域专家严格验证，并被定义为真实标签。接下来，我们使用Gemini 3 Pro（Google, 2025b）为每个种子问题生成看似合理的干扰项，并据此构建一组标准单选题（单选题）。对于多选（多选题）的构建，我们将该任务设定为负向选择范式：模型被要求选择多个错误选项，而不是选择正确答案。多选题的设计有效地降

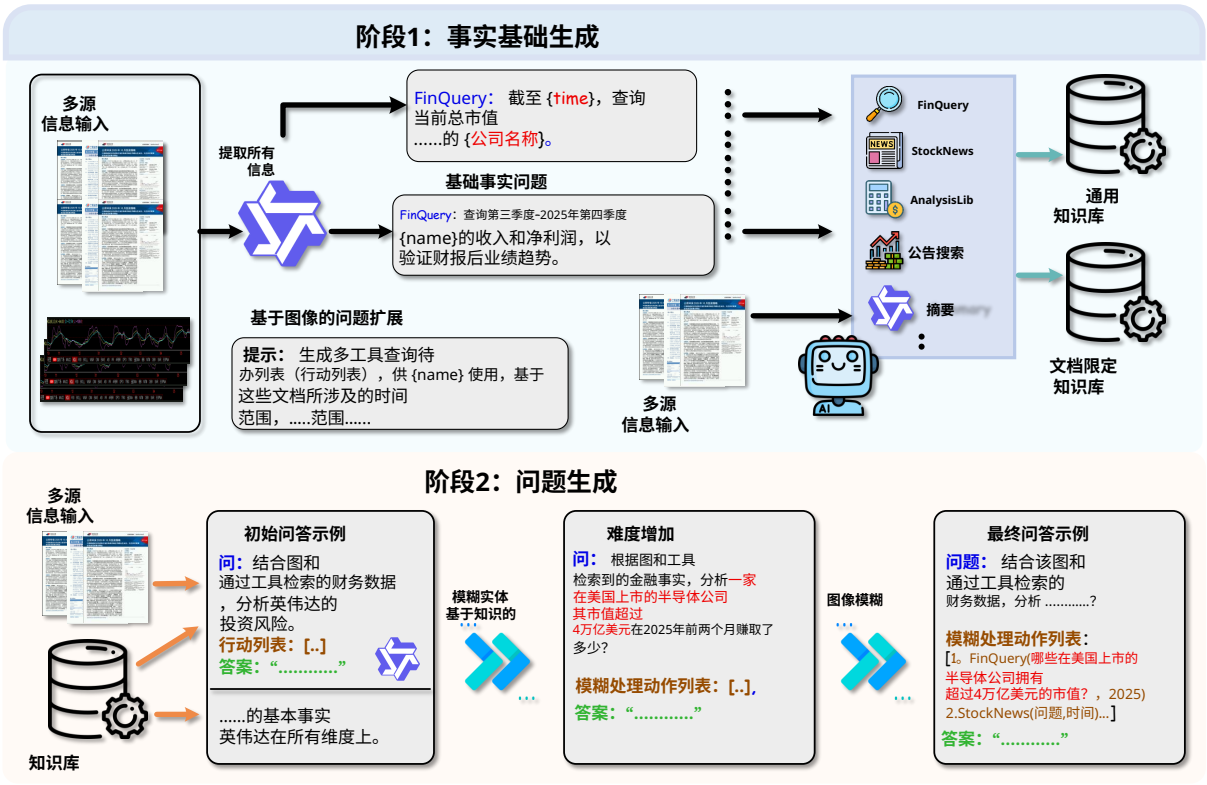

图3：针对金融智能体问题生成设计的数据合成流程。

低了猜对的概率，从而对模型提出了显著更为严格的挑战。

开放式问题生成。我们开发了一套流程，利用 Gemini 3 Pro 基于渐进式复杂度原则构建专家级多轮交互。首先，我们采用自动化流程和统一的提示规则，规范不同开放式子任务的结构、难度和评估目标。如表2所示，这些子任务分为四个渐进级别（L_1- L_4），认知需求从基础感知延伸至长期记忆。值得注意的是，该分类体系是在会话级别定义的，描述的是整个多轮对话的总体认知需求，而非任何单个轮次的认知需求。然而，基于规则的生成往往缺乏真实人机交互所需的细腻度。为此，我们加入人工润色，使对话更加自然。在保持结构可控的同时，通过引入口语表达和真实歧义，确保这些交互与真实的金融工作流程高度相似。

金融智能体数据生成。 为了构建一个包含参考工具调用和标准答案的智能体评估数据集，我们设计了一个两阶段数据合成流程，如图 3 所示。首先，我们为金融智能体数据生成构建两个知识库：一个通用知识库 K_1 和一个文档相关知识库 K_2。这些知识库存储可验证的金融事实和工具调用参数。 K_1 涵盖跨公司和时间范围的基础金融事实，并支持通过公司实体索引进行检索。 K_2 包含与文档内容相关的金融事实知识。其次，我们基于图像和知识库生成初始问答对。随后，为进一步提高任务难度，我们应用**模糊测试** 策略，系统地混淆实体先验。具体而言，我们利用知识库替换问题中的特定实体。最后，我们检查视觉内容，并遮盖泄露关键信息的区域，例如公司名称。进一步的实现细节见C.3节，载于补充材料。

3.3 数据集统计

我们的基准包含三项主要任务，共有 11,133 个问题、3,600 张图像和 400 份 PDF。它涵盖美国和中国A股市场的股票，涉及多个行业，反映了多样化的现实金融场景。该

客观题包括 1,982 道单选题和 1,982 道多选题，用于评估模型在明确约束条件下对金融多模态信息的基本理解。开放式问题包含 6,169 个样本，每个样本平均有 4.25 轮交互，平均输入词元数为 4,963。最后，金融智能体任务包含 1,000 个样本，均要求使用外部金融工具。约 73% 的任务可通过多次调用单一工具完成，其余 27% 则需要使用多个工具。这一设置反映了现实的金融分析与决策工作流程。详细统计数据见表2 及补充材料。

统计量	数量
客观题	3964
- 单选题	1982
- 多选题	1982
开放式问题	6169
理解（L_1）。	2082
- 实体识别	
- 空间感知	
计算（L_2）。	1893
- 多步数值计算	
- 图表数值估算	
自我纠错（L_3）。	1210
- 对抗鲁棒性	
- 逻辑一致性	
记忆（L_4）。	984
- 跨页面实体链接	
- 长上下文	
- 多源知识融合	
金融智能体	1000
- 单智能体	728
- 多智能体	272

表2：FinMTM 在不同任务类别及相应子任务中的统计细分。

3.4 综合金融评测

我们针对三种不同任务设计了评估框架：客观题、开放式问题和金融智能体。我们通过人工一致性实验验证了框架的稳健性，详细的实验分析见补充材料B第节。

3.4.1 客观问题评估

我们对单选题和多选题采用集合重叠评分规则。每个问题 i 都关联一个真实答案集合 $G_i \subseteq \mathcal{O}$ 和一个模型预测答案集合 $P_i \subseteq \mathcal{O}$，其中$\mathcal{O}$ 表示所有候选选项的集合。问题 i 的得分定义为

$$\text{Score}_i = \begin{cases} 0, & P_i \setminus G_i \neq \varnothing, \\ \dfrac{|P_i \cap G_i|}{|G_i|}, & \text{otherwise.} \end{cases} \tag{1}$$

此公式采用严格的禁止过度选择规则：任何错误选择都会导致得分为零；否则，按比例给予部分得分。单选题对应特殊情况 $|G_i| = 1$。

3.4.2 开放式问题评估

我们提出了一种双规则评估策略，用于评估模型在轮次级别和会话级别的表现。两个级别下基于大语言模型的评审所用的详细提示见补充材料。对于多轮对话 $\mathcal{D} = \{r_1, r_2, \ldots, r_T\}$，评估按如下方式进行。

轮次级评估。 每个回答 r_t 均依据五项基本金融能力进行评估：视觉精确率（VP）、金融逻辑（FL）、数据准确性（DA）、跨模态验证（CMV）和时间感知（TA）。由此得到每轮的能力得分向量：

$$\mathbf{s}_t = \left(s_t^{\text{VP}},\ s_t^{\text{FL}},\ s_t^{\text{DA}},\ s_t^{\text{CMV}},\ s_t^{\text{TA}}\right), \tag{2}$$

其中，每个分量 $s_t^c \in [0, 10]$ 反映模型在能力 c方面的表现。我们进一步将五项能力得分汇总为轮次级得分。

$$S_{\text{t}}(\mathcal{D}) = \frac{1}{T} \sum_{t=1}^{T} \sum_{c \in \mathcal{C}} w^c\, s_t^c, \tag{3}$$

其中，$\mathcal{C} = \{\text{VP,FL,DA,CMV,TA}\}$ 表示能力维度，且我们对所有 $c \in \mathcal{C}$使用均匀权重$w^c = 0.2$。

会话级评估。每个会话在数据集构建过程中都会被标注一个任务类型 ℓ ∈{沟通、计算、自我校正、记忆}。对于给定的会话 $\mathcal{D}$，我们首先确定其任务标签 ℓ，然后使用相应的检查清单 $\mathcal{K}_\ell$对整个会话进行评分，从而计算会话级得分 $S_{\text{e}}(\mathcal{D})$。如

方法	客观题		开放式问题			金融智能体		
	单智能体	多智能体	Com.	计算		记忆 含糊		
专有模型								
ChatGPT-4o（OpenAI，2024）	79.3	49.1	77.2	76.8	46.2	38.9	29.7	34.8
ChatGPT-o3*（OpenAI，2025b）	85.8	73.3	83.8	78.6	52.8	43.6	31.4	35.2
ChatGPT-5* (OpenAI, 2025a)	89.0	**79.6**	86.9	80.7	56.9	46.7	35.9	49.7
Gemini 3 Flash (Google, 2025a)	91.9	78.1	82.2	76.0	55.4	41.6	**53.6**	**62.6**
Grok-4-fast-non-reasoning* (xAI，2025)	71.0	46.8	66.0	61.2	39.9	24.8	30.2	39.7
Gemini3 Pro (Google，2025b)	**92.1**	78.4	**87.5**	**82.8**	**58.8**	**48.5**	48.3	54.3
InternVL 系列								
InternVL2.5-8B(Chen 等人，2024)	63.8	25.7	55.1	49.2	26.5	16.7	8.4	10.5
InternVL2.5-26B (Chen 等人，2024)	70.5	31.3	61.7	57.7	32.3	22.8	11.2	14.0
InternVL2.5-40B（Chen 等人，2024）	72.3	35.2	66.1	64.6	36.2	26.7	13.5	16.8
InternVL3-78B（Zhu 等人，2025c）	75.6	42.4	76.2	77.6	43.6	32.6	18.2	22.8
其他视觉语言系列								
MiMo-VL-7B（小米，2025）	61.1	21.4	75.1	75.4	47.2	39.9	20.2	25.5
GLM4.5V-108B（Hong 等人，2025）	73.7	51.0	85.4	79.6	51.1	42.2	26.5	32.4
*Qwen*视觉语言系列								
Qwen2.5-VL-3B（Bai 等，2025）	64.5	16.4	68.2	67.7	40.5	27.6	9.4	11.9
Qwen2.5-VL-7B（Bai 等，2025）	73.4	24.1	74.3	73.4	43.1	33.9	11.1	14.2
Qwen3-VL-4B-Instruct（Qwen，2025）	73.3	34.2	74.5	71.2	39.5	25.9	15.1	19.1
Qwen3-VL-4B-思考（Qwen，2025）	66.1	24.3	71.2	68.5	42.5	31.0	12.8	15.6
Qwen3-VL-30B-A3B-Instruct（Qwen，2025）	77.2	47.3	82.1	76.5	42.5	33.7	16.2	20.8
Qwen3-VL-30B-A3B-思考（Qwen，2025）	71.5	49.4	80.7	67.1	44.2	35.1	18.9	23.3
Qwen3-VL-32B-Instruct (Qwen, 2025)	84.5	39.9	84.3	80.7	50.8	40.3	19.6	25.1
Qwen3-VL-32B-Thinking（Qwen，2025）	83.4	46.5	80.3	68.6	43.5	33.7	23.2	28.6
Qwen3-VL-235B-A22B-Instruct（Qwen，2025）	81.3	48.5	85.5	80.9	54.5	41.5	32.1	38.7
Qwen3-VL-235B-A22B-Thinking (Qwen, 2025)	80.5	42.3	84.5	79.4	52.5	43.0	35.2	41.5

表 3：视觉语言模型在 FinMTM 上的表现。标有 * 的模型由于服务提供商政策限制，仅在 FinMTM 的子集上进行评估。**沟通**。、**计算**。、**自我校正**。、**记忆**。分别表示理解、计算、自我纠错和记忆一致性开放式任务。**有模糊测试** 和**无模糊测试** 表示金融智能体任务是否包含模糊输入条件。每个子任务的前两名结果分别以粗体和下划线标示。

总结见表 2，不同的对话类型强调不同的能力维度。

总体得分。为了为每段对话计算一个单一标量得分，我们以统一的方式汇总轮次级和会话级评估。最终的对话得分由两者加权组合得出：

$$S_{\text{final}}(\mathcal{D}) = \alpha\, S_{\text{t}}(\mathcal{D}) + (1-\alpha)\, S_{\text{e}}(\mathcal{D}) \quad (4)$$

其中，我们固定 $\alpha = 0.5$以平衡轮次级和会话级评估。

3.4.3 FinancialAgent 评估

我们提出一种基于轨迹的两阶段评估框架，用于评估金融多模态智能体的多步、多维交互。

规划阶段。模型会输出一条结构化轨迹，指定工具调用路径。遵循以往的智能体基准测试（Qin 等人，2023；Liu 等人，2024；Zhou 等人，2024；Zhu 等人，2025d），我们使用一个注重召回率的 F_β 得分来量化工具使用质量，其匹配依据是功能对齐，而非工具名称字符串匹配。我们将每次工具调用表示为$\tau = \langle$名称,args$\rangle$，其中 args 包含股票代码、日期范围或查询字符串等核心参数。令 $\mathcal{T}$ 表示预测集合，$\mathcal{T}^\star$表示参考集合。仅当预测调用$\tau \in \mathcal{T}$的工具名称及其核心参数在语义上均与 $\mathcal{T}^\star$中的某一项一致时，才将其计为真正例（TP）。我们分别将精确率和召回率定义为 $P = \text{TP}/|\mathcal{T}|$ 和 $R = \text{TP}/|\mathcal{T}^\star|$，并定义：

$$Q^{\text{t}} = w_t \cdot \frac{(1+\beta^2)PR}{\beta^2 P + R}, \quad \beta > 1, \quad (5)$$

其中 $w_t = 25$、$\beta = 2$、P，$R \in [0,1]$和 $Q^{\text{t}} \in [0, 25]$。

模型	VP	FL	DA	CMV	TA	S_e
Qwen3-VL-4B-Instruct	68.60	49.80	35.40	45.50	72.10	25.97
Qwen2.5-7B	78.00	59.90	45.60	59.60	83.50	26.53
Qwen3-VL-30B-A3B-Instruct	76.00	59.50	44.40	56.60	82.40	35.32
Qwen3-VL-30B-A3B-Thinking	69.20	54.10	40.40	51.50	75.00	33.41
Qwen3-VL-32B-Instruct	84.30	73.70	59.50	71.10	92.00	42.86
Qwen3-VL-32B-思考	76.50	59.40	44.10	56.30	82.40	40.12
Qwen3-VL-235B-A32B-Instruct	87.50	78.00	62.90	73.80	92.80	52.41
Qwen3-VL-235B-A32B-思考	85.80	77.20	61.40	74.50	92.10	50.57

表4：视觉语言模型在多轮问题上的表现，S_e 表示会话级平均分。

模型	轮次 1	轮次2	轮次3	轮次4	平均	β
Qwen2.5-VL-3B	52.9	58.7	56.0	68.8	59.1	4.5
Qwen2.5-VL-7B	61.2	61.9	65.0	72.5	65.2	3.7
Qwen3-VL-30B-A3B-Inst.	61.2	61.2	74.0	81.6	69.5	7.4
Qwen3-VL-30B-A3B-Think.	61.8	53.5	54.7	58.8	57.2	-0.8
Qwen3-VL-32B-Inst.	69.2	69.2	77.7	85.5	75.4	5.7
Qwen3-VL-235B-A32B	70.5	71.0	82.5	88.5	78.1	6.5
Qwen3-VL-235B-A22B-Think.	68.7	68.8	80.0	86.2	75.9	6

表5：每轮得分及趋势 β，其中 β 表示跨轮次性能提升的斜率。

摘要阶段。 模型综合规划轨迹和工具返回的信息，生成最终答案。推理过程和最终答案都必须以获取的视觉证据和工具输出为依据。在这种受证据约束的设置下，我们根据真实标签评估推理质量（$Q^{\mathrm{r}} \in [0, 25]$）和答案正确性（$Q^{\mathrm{a}} \in [0, 50]$）。最终得分为：

$$Q_{\text{final}} = Q^{\mathrm{a}} + Q^{\mathrm{r}} + Q^{\mathrm{t}}. \tag{6}$$

4 实验

4.1 实验设置

评测模型。我们共评测了 22 个视觉语言模型，包括专有系统和开源系统。具体而言，专有模型包括 ChatGPT-4o (OpenAI, 2024)、ChatGPT- o3 (OpenAI, 2025b)、ChatGPT-5 (OpenAI, 2025a)，以及 Gemini 3 Pro (Google, 2025b) 和 Gemini 3 Flash (Google, 2025a)，还有 Grok-4-fast-non-reasoning (xAI, 2025)。开源模型包括 InternVL 系列，即 InternVL2.5-8、26、40B (Chen et al., 2024) 和 InternVL3-78B (Zhu et al., 2025c)，以及 Qwen 系列，包括 Qwen2.5-VL-3、7B 和 Qwen3-VL-4B、30B-A3B、32B、235B-A22B-Instruct、Thinking (Bai et al., 2025; Qwen, 2025)。此外，我们还评测了其他有代表性的开源视觉语言模型，包括 MiMo-VL-7B (Xiaomi, 2025) 和 GLM4.5V-108B (Hong et al., 2025)，以进一步丰富模型的多样性。

模型	Mode	SC	MC
Qwen3-VL-32B	开启思考	83.4	46.5
	关闭思考	85.0	45.8
	Δ (开启–关闭)	-1.6	+0.7
Qwen3-VL-30B-A3B	开启思考	71.5	49.4
	关闭思考	73.0	48.6
	Δ (开启–关闭)	-1.5	+0.8

表6：Fin-MTM 推理时思考消融实验。SC = 单选题，MC =多选题。

4.2 实验结果

根据表 3 所示结果，专有模型在大多数维度上均优于开源模型。值得注意的是，Gemini 3 Pro 在开放式任务中表现出色，在理解（87.5）、计算（82.8）、自我纠错（58.8）和记忆（48.5）等子任务中名列前茅。这些结果凸显了 Gemini 3 Pro 有效处理复杂推理和复杂对话的能力。另一方面，ChatGPT-5 在多项选择题中表现最佳（79.6），同时在开放式问题中取得第二好的结果。与此同时，Gemini 3 Flash 在经模糊测试和未经模糊测试的金融智能体任务中均取得领先得分（53.6/62.6），展现出卓越的自主决策智能体能力。

专有模型与开源模型在开放式问题和金融智能体任务中的性能差距十分明显。开源模型（例如，InternVL、Qwen VL 系列）在客观任务中表现不俗（Qwen3-VL-32B-Instruct 在单选题中得分为 84.5，GLM4.5V-108B 在多项选择题中得分为 51.0），但在自我纠错和记忆密集型开放式任务中的表现则大幅下滑。例如，Qwen3-VL-32B-Instruct 不及 Gemini 3 Pro 等领先的专有模型（自我纠错：71.5对比 58.8；记忆：44.2 对比 48.5），凸显了开源模型在复杂推理和长时记忆方面的局限性。

在金融智能体任务中，开源模型的表现不及专有模型。例如，Qwen3-VL-32B-Thinking 的得分为 25.1（经过模糊测试）和 28.6（未经过模糊测试），而 Gemini 3 Flash 以 53.6 和 62.6 领先。这种表现不佳暴露了开源模型在多步骤推理和基于工具的环境中的局限性，而这两者对于金融预测和

决策。更多实验结果和分析见第 E 节，详见补充材料。

4.3 分析

任务复杂度对模型表现的影响。如表3所示，开放式问题的实验结果表明，随着任务复杂度的增加，大多数模型的表现明显下降（L_1-L_4）。模型在基础层级的沟通（L_1）和计算（L_2）任务中表现稳健。例如，Qwen3-VL-32B-Instruct 在这两个类别中均取得了较高得分（分别为 84.3 和 80.7）。相反，随着认知要求不断提高，模型表现显著下降。在自我校正（L_3）和记忆（L_4）任务中，所有模型均呈现持续下降趋势。具体而言，记忆密集型任务（L_4）涉及多图像引用、密集的知识落地以及长上下文推理，仍是开源模型难以逾越的障碍。即使是表现最佳的模型 Gemini 3 Pro，得分也仅为 48.5，而 InternVL2.5-8B 的得分更只有 16.7。这表明，在长上下文场景中进行复杂推理，仍是当前模型面临的一项重大挑战。

轮次级得分详细分析。表4展示了轮次级表现的详细情况：视觉精确度和时间感知表现突出，但在金融逻辑、数据准确性和跨模态验证方面存在不足。这表明，尽管当前模型已具备较高水平的视觉感知能力，但复杂逻辑推理能力仍有待进一步提升。表5 列出了各模型在不同轮次的得分，以及计算所得的斜率（β），用于表示得分随轮次变化的趋势。总体而言，大多数视觉语言模型都能从延长的交互中获益，这表明上下文学习有效。这一提升部分归因于后续轮次促使模型重新关注视觉证据，并在累积的对话上下文中修正先前的错误（例如，纠正误读的数值或不一致的解读）。

“思考”模式带来的性能提升。我们观察到，“思考”与指令变体之间存在任务相关的权衡。“思考”模式能提升高熵场景下的表现（例如，金融智能体任务；235B: $38.7 \rightarrow 41.5$），但会略微降低低级感知任务的表现。一种合理的解释是，需要快速视觉—语义对齐（例如，L_1 理解和单选题）可从直接映射中获益，而较长的思维链生成可能引入干扰性的中间内容，并模糊显著的视觉线索。如表 6所示，在低熵单选题中关闭思考模式会带来小幅提升，而在高熵多选题中启用思考模式则表现更佳。

错误分析。我们对得分较低的多轮实例进行了错误分析：对于 L_1，错误通常源于最初对关键数值或实体的感知错误。由于模型未能依据视觉证据重新核验其输出，这些错误会在后续轮次中不断传播并加剧。对于 L_2，失败主要源于变量值错配、参照框架定义不充分，或单位／尺度不一致（例如，将局部极值误认为全局极值）。这往往会导致推理步骤在逻辑上连贯，却仍得出错误的计算结果。对于 L_3，当后续轮次引入外部事实或改变先决条件时，模型便会遇到挑战。模型经常未能修正先前的假设，导致结论与更新后的上下文不一致。对于 L_4，失败主要源于模型难以在多页文档中进行证据溯源。模型往往难以从指定页面中检索支持性信息，从而产生不完整或虚假的引用，并用幻觉细节填补信息空白。详细分析见补充材料第D节补充材料。

专家验证。为确保数据可靠性，由 15 位金融专家和学术研究人员组成的小组对数据集进行了验证。此外，我们还开展了人类对齐实验，以验证开放式问题的评估准确性。详情见补充材料的第B 节。

5 结论

我们推出 FinMTM，这是一个多轮多模态基准，用于评估视觉语言模型全面的金融推理和智能体工具使用能力。我们对最先进的专有和开源视觉语言模型进行了广泛评估，结果表明，尽管当前模型在单轮感知和数值任务上表现出色，但在长时程多轮一致性、多文档金融推理和结构化金融智能体规划方面仍面临困难。基于本基准揭示的洞见和失效模式，未来的工作将着重开发更

稳健的金融视觉语言模型，提升其长上下文推理、证据溯源和工具驱动的决策能力。

6 局限性

FinMTM 存在若干局限性。首先，尽管我们采取了严格的质量控制措施，开放式评估的部分环节仍依赖基于大语言模型的评审，这可能带来残余的主观性，并使评估结果易受提示设计影响；我们在补充材料中公开了评分标准和提示，但提升评审稳健性和校准能力仍是未来工作的重要方向。其次，目前的金融智能体设定假定使用固定的一组 MCP工具和预定义的参考轨迹；尽管我们的指标试图评估超越工具名称匹配的功能对齐能力，但现实世界中的工具生态系统更为广泛、会不断演变，而且可能需要在工具可用性和输出均非平稳的条件下进行评估。第三，长文档任务（L_4）对引用和格式提出了严格约束；这反映了实践中以合规为导向的工作流程，但如果回答在实质上正确、却未遵循规定的流程，评估可能会低估其能力。未来版本的 FinMTM 可以纳入更强大的自动验证机制，例如结构化证据提取和引用一致性检查，并扩大工具覆盖范围，以更好地模拟开放世界金融决策环境。

参考文献

Shuai Bai 和另外 1 位作者。2025 年。Qwen2.5-vl 技术报告。*arXiv*预印本 *arXiv:2502.13923*。

Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 和另外 1 位作者。2020 年。语言模型是少样本学习者。在 神经信息处理系统进展 第 33 卷，第 1877–1901 页。

Zhe Chen 和另外 1 位作者。2024 年。拓展开源多模态大型语言模型的性能边界。*arXiv* 预印本 *arXiv:2412.05271*。

Shuangyan Deng、Haizhou Peng、Jiachen Xu、Rui Mao、Ciprian Doru Giurc˘aneanu 和 Jiamou Liu。2025 年。FinMR：面向高级金融推理的知识密集型多模态基准。在 第六届 *ACM*国际金融人工智能会议（*ICAIF*）论文集。

Ziliang Gan、Dong Zhang、Haohan Li、Yang Wu、Xueyuan Lin、Ji Liu、HaipangWu、ChaoyouFu、Zenglin Xu、Rongjunchen Zhang 及另外 1 位作者。2025 年。Mme-finance：用于专家级理解和推理的多模态金融基准。载于 第 *33* 届 *ACM* 国际多媒体会议论文集，第 12867–12874 页。

Google.2025a. Gemini3flash 系统卡 (pdf).
https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf.

Google. 2025b. Gemini3 pro 系统卡 (pdf).
https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

Yueru He、Xueqing Peng、Yupeng Cao、Yan Wang、Lingfei Qian、Haohang Li、Yi Han、Ruoyu Xiang、Mingquan Lin、Prayag Tiwari、Jimin Huang、Guojun Xiong 和 Sophia Ananiadou。2025。《Fincriticaled：用于金融事实级 OCR 评估的视觉基准》。*arXiv* 预印本 *arXiv:2511.14998*。Wenyi Hong 等 1 人。2025。《GLM-4.5V 和 GLM-4.1V-Thinking：借助可扩展强化学习实现多功能多模态推理》。 *arXiv* 预印本 *arXiv:2507.01006*。Xinyi Huang、Wei Zhang 和 Yu Li。2023。金融领域的语言模型：应用与挑战。*arXiv* 预印本 *arXiv:2307.00001*。Jiangtong Li、Yiyun Zhu、Dawei Cheng、Zhijun Ding 和 Changjun Jiang。2025a。《CFBenchmark-MM：面向多模态大语言模型的中文金融助手基准》。*arXiv* 预印本 *arXiv:2506.13055*。Xiang Li、Ming Wang、Jie Chen 等 1 人。2024。Qwen-VL：面向多模态智能的通用视觉语言模型。*arXiv* 预印本 *arXiv:2401.04088*。Yuante Li、Xu Yang、Xiao Yang、Minrui Xu、Xisen Wang、Weiqing Liu 和 Jiang Bian。2025b。R&D-Agent-Quant：用于数据中心型因子与模型联合优化的多智能体框架。预印本，arXiv:2505.15155。Xueyuan Lin、Cehao Yang、Ye Ma、Ming Li、Rongjunchen Zhang、Yang Ni、Xiaojun Wu、Chengjin Xu、Jian Guo 和 Hui Xiong。2025。《Retuning：利用大语言模型升级推理时扩展，以预测股票走势》。 *arXiv* 预印本 *arXiv:2510.21604*。Xiao Liu、Yujie Zheng、Yali Du、Ziming Li、Junxian Wang、Ziqiang Liu 和 Yue Zhang。2024。《Agent-Bench：评估作为智能体的大语言模型》。载于国际学习表征会议。Zhaowei Liu、Xin Guo、Haotian Xia、Lingfeng Zeng、Fangqi Lou、Jinyi Niu、Mengping Li、Qi Qi、Jiahuan Li、Wei Zhang、Yinglong Wang、Weige Cai、Weining Shen 和 Liwen Zhang。2025。VisFinEval：一个场景驱动的中文多模态基准，用于

全面的金融理解。载于*2025* 年自然语言处理实证方法会议论文集，佛罗里达州迈阿密。计算语言学协会。

Junyu Luo、Zhizhuo Kou、Liming Yang、Xiao Luo、Jinsheng Huang、Zhiping Xiao、Jingshu Peng、Chengzhong Liu、Jiaming Ji、Xuanzhe Liu 及其他1人。2025。FinMME：用于金融多模态推理评估的基准数据集。*arXiv*预印本 *arXiv:2505.24714*。

Philip M. McCarthy和Scott Jarvis。2010。MTLD、vocd-D与HD-D：词汇多样性评估精细方法的验证研究。行为研究方法，42(2):381–392。

OpenAI。2023。GPT-4技术报告。技术报告，OpenAI。

OpenAI。2024。GPT-4o系统卡。https://openai.com/index/gpt-4o-system-card/。

OpenAI。2025a。GPT-5系统卡（pdf）。https://cdn.openai.com/gpt-5-system-card.pdf。

OpenAI。2025b。OpenAI o3与o4-mini系统卡。https://openai.com/index/o3-o4-mini-system-card/。

Yujia Qin、Ziyu Wang、Lijun Jin、Yixuan Tang、Chen Zhang、Xu Han、Zhiyuan Liu和Maosong Sun。2023。ToolBench：面向将大语言模型作为工具学习器进行评估。见神经信息处理系统进展。Qwen。2025。Qwen3-VL：Qwen3系列中的视觉语言模型。https://github.com/QwenLM/Qwen3-VL。Sheng Wu、Tyler Mahoney、Vikas Kulkarni及其他1人。2023。BloombergGPT：面向金融的大语言模型。*arXiv*预印本 *arXiv:2303.17564*。xAI。2025。xAI Grok 4 Fast 模型卡（pdf）。https://data.x.ai/2025-09-19-grok-4-fast-model-card.pdf。

小米。2025。MiMo：释放语言模型的推理潜力——从预训练到后训练。*arXiv*预印本 *arXiv:2505.07608*。

Suifeng Zhao、Zhuoran Jin、Sujian Li和Jun Gao。2025。FinRAGBench-V：金融领域中带视觉引用的多模态检索增强生成基准。见*2025*年自然语言处理经验方法会议论文集，佛罗里达州迈阿密。计算语言学协会。

Shuyan Zhou、Zhenfang Chen、Yifei Shen、Yuchi Li、Yao Yu、Zhuo Jiang和Jiawei Han。2024。WebArena：用于构建自主智能体的真实网页环境。见国际学习表征会议。

Fengbin Zhu、Junfeng Li、Liangming Pan、Wenjie Wang、Fuli Feng、Chao Wang、Huanbo Luan 和 Tat- Seng Chua。2025a。Fintmmbench：金融领域具备时间感知能力的多模态检索增强生成基准。*arXiv* 预印本 *arXiv:2503.05185*。Fengbin Zhu、Xiang Yao Ng、Ziyang Liu、Chang Liu、Xianwei Zeng、Chao Wang、Tianhui Tan、Xuan Yao、Pengyang Shao、Min Xu 等人。2025b。Findeepresearch：评估严谨金融分析中的深度研究智能体。*arXiv* 预印本 *arXiv:2510.13936*。J. Zhu 等人。2025c。Internvl3：增强型开源多模态大语言模型。*arXiv* 预印本 *arXiv:2504.10479*。Jie Zhu、Yimin Tian、Boyang Li、Kehao Wu、Zhongzhi Liang、Junhui Li、Xianyin Zhang、Lifan Guo、Feng Chen、Yong Liu 等人。2025d。Finmcp-bench：基于模型上下文协议，对真实世界金融工具使用场景下的大语言模型智能体进行基准测试。

A 实验详情

我们分别评估专有模型和开源模型。专有模型通过商业 API 调用，而较小的开源模型则在本地部署。所有本地实验均在 NVIDIA H200 GPU 上进行，并使用 vLLM 进行高效部署和推理。本地推理时，我们启用多个 GPU 间的张量并行，并采用统一的解码配置以进行公平比较：温度 $T = 0$、top-$p = 1.0$、最大提示长度为 32,768 个词元，最大响应长度为 4096 个词元。为完整起见，我们在图10、图 11 和图 12中报告了推理提示词。

B 评估器分析

我们在双规则评估协议下开展实验，以评估不同评估器的有效性。该协议结合了轮次级能力评分和会话级检查清单验证。我们从开放式划分中随机抽取 120 个多轮实例，并使用 Qwen3-VL-30B-A3B-Instruct 作为固定响应生成器生成相应的输出。每个实例均由三位具备金融领域专业知识的资深标注人员独立评分。人工最终得分通过多数投票确定（例如，分数为 2、3、3 时，最终得分为 3），或通过均值取整确定（例如，分数为 1、2、3 时，最终得分为 2）。如果评分方差超过 2 分，则将该实例提交至裁决轮次，以确定最终得分。

随后，我们计算每个评估器的得分与人工标注分数之间的斯皮尔曼秩相关系数（Sp.）和平均绝对差Δ，以量化评估器的可靠性。如表 7所示，ChatGPT-4o 与人工判断的一致性最高，而 o3 和 Gemini-3-Pro 也表现出较强的吻合度。在开源评估器中，Qwen3-VL-235B-A22B 的一致性与之相当，且评估成本大幅降低，因此是大规模基准测试的实用替代方案。此外，我们发现，加入图像输入通常能提升评估器的一致性，这表明直接获取视觉证据有助于减少评分过程中的歧义。

我们进一步将问题划分为客观类别和主观类别。如表 8 所示，ChatGPT-4o 评估器在客观题上的一致性更高

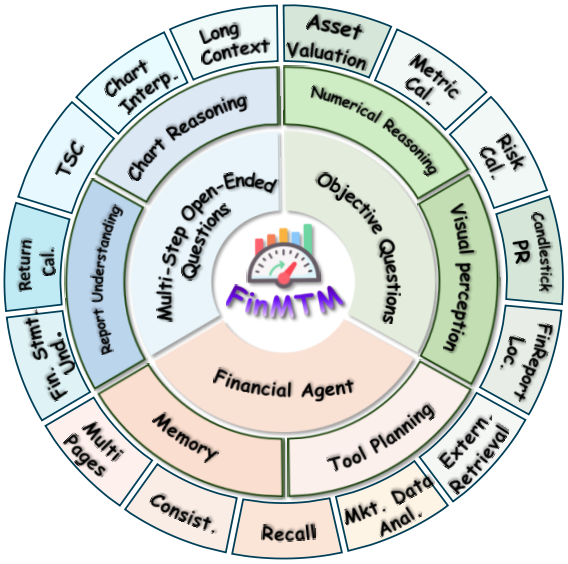

图 4：所提出的 FinMTM 概览。

表 7：评估器得分与 FinMTM 人工标注分数之间的一致性。斯皮尔曼等级相关系数（Sp.）越大、平均绝对差值（Δ）越小，表明评估器的可靠性越高。“Pic.”表示评估时将原始图像作为输入。

模型	Sp.（↑）	Δ（↓）
ChatGPT-3.5	0.46	1.42
ChatGPT-4-Turbo	0.71	0.95
o1-preview	0.62	1.12
ChatGPT-4o	0.78 (0.81)	0.76 (0.71)
Gemini-3-flash（图片）	0.74 (0.77)	0.83 (0.79)
Gemini-3-pro（图片）	0.72 (0.73)	0.92 (0.90)

（Sp. 更高且 Δ 更低），而主观问题仍然更具挑战性，也更容易受到视觉定位和评分标准解读的影响。最后，除了定量一致性研究外，我们还邀请资深金融专家和学术研究人员开展分层审计，以验证自动化评估是否符合现实金融投资研究的逻辑。

C 更多数据集详情

C.1 数据集概览

图4 展示了所提出的 **Fin-MTM**基准的总体概览。该基准旨在真实的金融推理与智能体交互场景下，系统地评估视觉语言模型（VLM）。FinMTM 将金融任务划分为三个互补类别，分别是 多步骤

表8：ChatGPT-4o 评估器与人工评分在各类别上的一致性。“w”和“w/o”分别表示使用和不使用图像输入进行评估。

	客观				主观			
	Sp. (w)	Δ (w)	Sp. （不使用）	Δ （不使用）	Sp. (w)	Δ (w)	Sp. （不使用）	Δ （不含）
ChatGPT-4o	0.86	0.55	0.84	0.60	0.54	1.18	0.49	1.26

开放式问题、客观题和金融智能体任务，涵盖了真实世界金融分析所需的广泛能力。每类任务进一步细分为细粒度的能力维度。多步骤开放式问题 侧重于图表理解、图表推理和长上下文理解，考察模型基于视觉金融证据进行迭代推理的能力。客观题 侧重于数值推理和视觉感知，包括指标计算、风险评估、资产估值和K线形态识别，这些任务要求准确结合视觉与数值信号。金融智能体任务 评估更高层次的智能体能力，例如记忆一致性、召回率、工具规划、外部检索和多页推理，体现真实的多轮金融工作流程。

C.2 数据集标注

为了降低人工标注成本并提高效率，我们引入了一种机器辅助标注方案。首先，金融领域专家设计自动化标准，用于评估问题的质量与合理性。随后，我们使用三种能力强大的视觉语言模型（VLM），即 Gemini 3 Pro、ChatGPT-5 和 QwenVL-235B-A22B，对客观题和多步骤开放式问题的合理性及可回答性进行初步评估。对于给定的问题，如果三个模型均认为其符合质量标准，则直接予以接受，无需进一步进行细致的人工标注。只有存在分歧的问题才会提交给人工标注员审核。

在人工标注阶段，我们组建了一支由 15 名标注员组成的团队，其中包括 10 名初级标注员和 5 名专家。具备基础金融知识的初级标注员负责问题审查、改写和独立解题。专家组由以下人员组成：(i) 四名来自计算机科学和金融相关领域的学术研究人员，且均至少拥有硕士学位，以及 (ii) 四名金融行业专业人士。专家负责问题筛选、质量评估和最终答案核验。总体而言，标注与审查流程约需 600 小时的人力投入，在保持较高数据质量的同时，显著降低了标注成本。

C.3 知识库构建详情

该框架包含三个步骤：知识库构建、初始轨迹合成，以及基于事实注入的实体模糊化与轨迹演化。

(i) 我们构建了一个由工业级金融工具支持的可验证知识库，其中包括通用知识库 K_1 和文档限定知识库 K_2。该知识库存储可验证事实和工具调用参数。 K_1 涵盖跨公司和时间范围的基本金融事实，并支持通过公司实体索引进行检索。在领域专家的指导下，我们定义常见的金融分析维度（如市值、估值等），从财务报告和研究文档中自动提取公司实体，并将“实体 × 维度”组合成问题模板，以批量检索和积累相应的可验证事实。

对于 K_2，我们专注于单份财务报告层面的内容扩展，并通过文档标识符支持索引和检索。我们将文档中明确陈述的事实（如季度收入、利润等）作为锚点，设定结构化扩展策略，并引导模型沿时间、范围等维度进行扩展推理。在此过程中，模型会自动生成工具查询参数，并通过 MCP 接口获取结果。

（ii）通过将基于工具调用从 K_1 和 K_2 获取的多模态信息与可验证事实相结合，我们明确约束并引导

表9：FinMTM多轮开放式问题的统计量。词汇多样性使用MTLD（McCarthy和Jarvis，2010）衡量。

统计量	数量
对话总数	6,169
不同图像	2,427
平均/最大轮数	4.25 / 5
平均/最大问题长度（字符数）	105.40 / 667
平均/最大回答长度（字符数）	75.65 / 1,146
问题多样性（词元）	11,250
答案多样性（词元）	14,890
输入长度分布（词元）	
[959, 3,717]	49.98%
[3,717, 13,505]	40.01%
[13,505, 14,124]	4.94%
[14,124, 135,055]	5.07%
中位数 / 平均词元	4,654 / 4,963

模型将这些事实组织成连贯的分析轨迹。基于该轨迹，模型通过逆向构造生成相应的金融问题和答案，从而获得一条在逻辑上保持一致且事实可验证的初始智能体轨迹。

(iii) 为了在不影响可验证性的前提下构建更具挑战性且更具泛化性的评测场景，我们逐步将工具从知识库中获取的事实纳入问题重构，在实体和表达层面重新组织原始查询。

C.4 更多数据集统计

表 9 总结了 FinMTM 多轮开放式子集的统计数据。该数据集包含 6,169 段基于 2,427 张不同图像的多轮对话，每个会话平均有 4.25 轮，反映了真实的多轮交互模式。问题和答案展现出显著的词汇多样性（以 MTLD 衡量），表明语言变化丰富且模板偏差较小。此外，输入长度分布涵盖了广泛的词元数量，中位数为 4,654 个词元，最大值超过 135k 个词元，凸显了长上下文场景的存在，这些场景会考验模型的记忆能力、上下文跟踪和跨轮次推理能力。

C.5 数据集示例

FinMTM旨在评估金融视觉语言模型在异构监督信号与交互模式下的表现，涵盖客观题选择、多轮报告依据对话以及工具增强型智能体规划。图7 展示了客观设定，包括基于技术指标图表的单选题和多项选择题。此类问题强调细粒度视觉感知与选项级判别，可作为事实依据和局部推理的受控测试平台。

除了客观评估，FinMTM 还包含多轮开放式以报告为中心的对话（图 9），在这些对话中，模型必须从多页报告中提取数字、引用跨页证据、进行中间计算，并在后续轮次中核实陈述。此设置在明确要求引用的情况下，考验长上下文稳定性、跨页证据关联和算术可靠性。

最后，图 8 展示了一个多轮规划智能体会话，其中模型必须先根据图表解读行业趋势，再规划并执行工具调用，在商业／市场约束下检索候选实体，最终针对公司层面的财务指标进行一致性检查。这种智能体设置突显了长时程规划、工具使用编排和实体消歧能力；这些能力对于现实世界中的金融工作流程至关重要，但对当前的视觉语言模型而言仍颇具挑战。

D 错误分析

尽管 Gemini 在 FinMTM 上整体表现出色，我们仍针对得分较低的示例开展了案例研究，以分析具有代表性的失败案例。如图 5所示，Gemini 存在系统性错误，例如误读图表和跨轮次不一致。

D.1 理解（L_1）

对低得分案例的分析表明，失败的主要原因在于金融视觉感知不准确。在早期轮次中，模型经常错误识别金融图表中的关键感知锚点，例如全局极值、对应的时间戳或精确的数值位置。这些错误的感知结果随后被视为固定的金融事实，并传播到后续的多步推理中。由于缺乏明确的机制来依据视觉输入重新核验或纠正感知结果，初始感知错误会在多个轮次中持续存在，并在差值估计和反事实调整等后续操作中进一步放大。此外，当出现不一致时

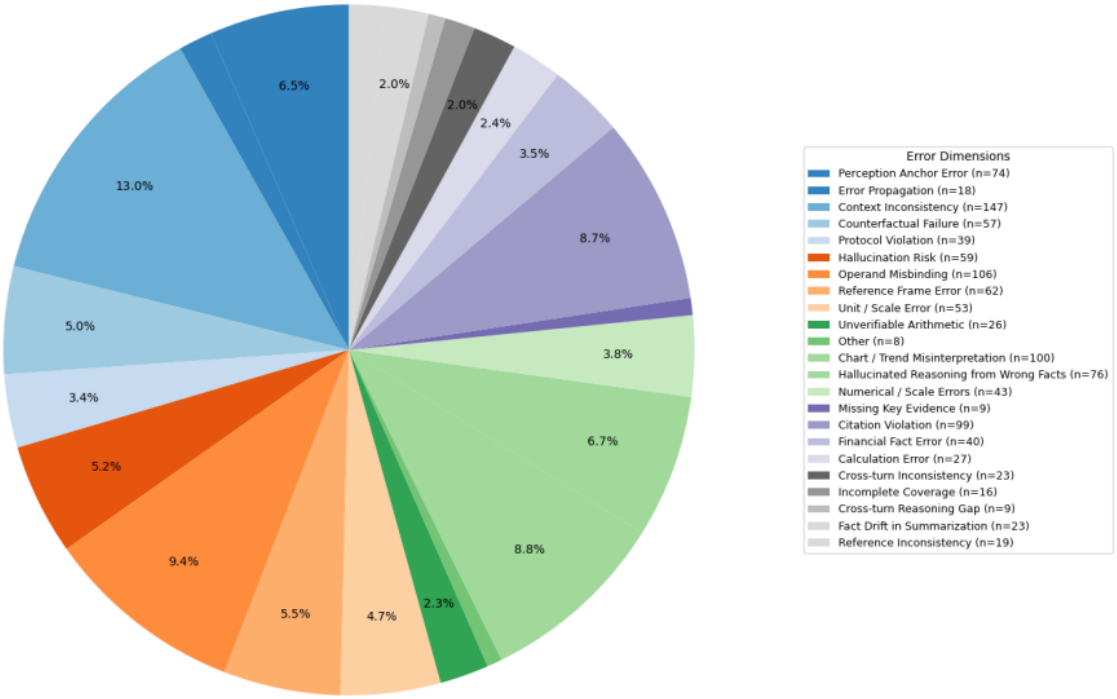

图5：此图展示了错误类型的分布情况，数据基于从 Gemini3 Pro 抽取的 1133 条回答。每条回答均经过人工审查，并被归类为不同的错误类型。

对于用户提供的描述与视觉证据之间的矛盾，模型往往无法加以调和，而是被动接受错误的上下文线索，从而加剧感知漂移。

总体而言，低得分样本揭示了模型在维持稳定且可验证的金融感知状态方面存在根本局限，这严重削弱了模型在金融图表上进行以感知为基础的多步骤推理的能力。

D.2 计算 (L_2)

对低得分案例的分析表明，失败的主要原因并非推理结构存在缺陷，而是在逻辑计算过程中错误地绑定数值与参考对象。在大多数情况下，模型能够生成连贯的推理链，却在计算前选择了错误的运算数，例如混淆不同的曲线或子图，或将局部极值误认为全局极值，导致后续的差值或比率计算无效。

此外，在处理涉及峰谷差或回撤的计算时，模型往往未能明确建立全局参考框架，而是直接进行算术运算所依据的是错误或不完整的基线。单位与尺度处理中的错误（例如，百分比与基点之间的换算不一致）会进一步放大数值偏差。在某些情况下，尽管缺少明确的坐标轴刻度或数值标注，模型仍报告精确的数值结果，从而导致无法验证的计算。

总体而言，这些低得分案例表明，模型的局限在于对计算前约束的建模不足，包括变量绑定、参考验证和单位一致性，进而导致一种典型的失败模式：推理逻辑连贯，但数值不准确。

D.3 自我纠正 (L_3)

我们分析了 L_3 多轮视觉推理任务中的低得分案例，以识别常见的失败模式。主要问题是对视觉证据的误读，包括错误的趋势方向、图例对应关系或关键拐点。由于视觉理解构成后续推理的事实基础，此类错误会使整个分析依赖于错误假设。此外，模型在跨轮次一致性方面存在不足，因为其预测在连续

模型	Tool	推理	总分	召回率 (%)	F1 (%)	EMR
Gemini _3_Flash _Preview	**11.2653**	**21.7959**	**62.9000**	**44.9245**	**42.2057**	**0.1020**
ChatGPT-5	8.0816	16.2857	44.7755	39.7694	30.3771	0.0302
Grok	9.3061	12.3469	36.9592	36.0402	34.8976	0.0408
Gemini	4.8333	20.3095	54.9048	25.1969	19.2095	0.0331
ChatGPT-3	6.6327	13.3469	35.2857	22.7796	22.4422	0.0204
Qwen3-VL-235B-Thinking	7.4231	15.1245	41.5122	36.8541	33.9102	0.0321
GLM-4.5V-108B	5.8541	11.9210	32.4353	28.1245	25.3210	0.0275
Qwen3-VL-32B-Instruct	4.5124	9.1541	25.1124	22.3120	19.7852	0.0142
InternVL3-78B	4.1245	8.3541	22.8857	20.2415	17.9214	0.0118
InternVL2.5-40B	3.0142	6.1425	16.8346	15.1241	13.2145	0.0075

表 10：金融智能体任务上的按模型评估统计量。

轮次中可能依赖于相互不一致的假设。它还表现出有限的**自我纠正能力**，即使后续轮次提供了明确的纠正线索，也常常未能更新或撤回先前的假设。总体而言，低得分主要源于视觉定位和错误纠正方面的不足，而非语言流畅性或推理结构方面的局限。

D.4 记忆（L_4）

检查 L_4场景中的低分样本后发现，大多数失败并非源于视觉感知或逻辑推理能力不足，而是违反了严格的评估约束，以及金融表达错误。最主要的失败原因是不遵守引用依赖规则：该规则要求在每个对话轮次中提供明确且结构化的引用（例如，［Source, Page X］）。即使单条回复在事实上正确，缺失或不完整的引用也会使整段多轮对话轨迹无效。

除引用相关问题外，部分错误还源于金融数据错误，包括盈亏方向混淆、单位误读，以及增长率或百分比计算错误。这些错误在多轮场景中会被放大，因为早期的数值偏差会在后续推理步骤中不断传递。此外，部分失败归因于跨轮次不一致，即后续总结遗漏或违背了此前已确立的事实。

总体而言，观察到的低得分案例主要反映了在遵循严格的验证导向协议以及跨轮次保持金融严谨性方面的挑战，而非在多模态理解或推理能力方面存在根本性局限。

E 金融智能体的进一步分析

图 6 展示了金融智能体任务的推理与评估流程。在 FinMTM 面向智能体的实验中，我们观察到，引入模糊化策略会持续降低所有受评模型的性能，凸显了在语义先验减少的情况下，决策难度有所增加。对失败案例的详细分析揭示了若干关键发现。

与智能体框架不兼容：模糊输入。 对于部分查询，智能体在模糊设置下无法返回有效输出。Gemini 等更强的模型通常能够通过重新组织查询或调整交互策略来恢复，而较弱的模型则更容易陷入不可逆的失败状态。

小型模型的智能体能力有限。 较小的模型经常出现冗余或无效的工具使用行为，例如反复调用同一工具，或通过多个工具查询相同信息。这种行为表明，智能体循环中的规划、记忆和内部状态管理存在不足。

视觉感知薄弱导致的智能体失败加剧。细粒度视觉理解不足会进一步加剧智能体层面的失败。在许多情况下，小型模型会尝试通过工具检索信息，而这些信息本可以直接从视觉输入中推断出来（例如行业图表中的趋势模式）。由于无法通过工具调用获取此类信息，这些智能体往往会触发强制摘要机制并提前终止。

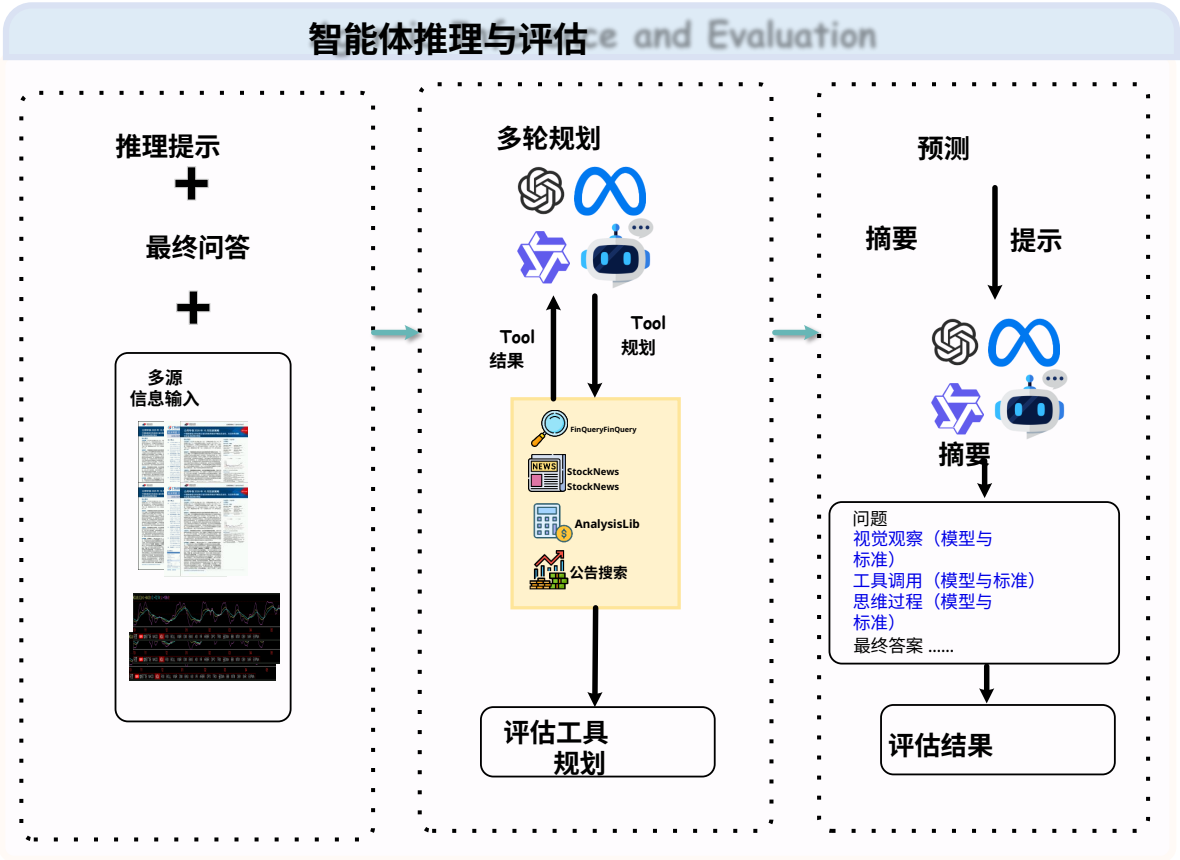

图 6：智能体评估流程示意图。

Gemini 3 Flash 的金融智能体表现更出色。在专有模型中，Gemini 3 Flash 展现出显著更强的金融智能体能力。然而，在不涉及智能体交互的纯推理场景中，闭源模型之间的性能差距要小得多，这表明它们的优势主要来自更强的智能体执行能力，而非原始推理能力。

如表 10所总结，在 FinMTM 金融智能体场景中，各模型呈现出明显的能力分层。总体而言，闭源模型的智能体综合性能显著优于开源模型。尤其是 Gemini 3 Flash-Preview，在工具、推理和总分等核心指标上领先，表明其在工具调用稳定性、推理质量和结果对齐方面具有很强的稳健性。Gemini 和 ChatGPT-5 虽然展现出较强的推理能力和信息覆盖度，但在执行对齐指标（例如召回率、F1 和 EMR）上的表现仍明显下降，反映出它们在将推理能力转化为有效智能体执行方面存在局限。在开源模型中，Qwen3-VL-235B-Thinking 表现突出，其性能可与部分闭源系统相媲美，并展现出较强的智能体执行与对齐能力。相比之下，较小规模的模型（例如 Qwen3-VL-32B 和 InternVL 系列）在工具和推理指标上均持续表现不佳，导致整体性能受到显著限制。这些结果表明，在复杂金融智能体场景中，模型规模以及智能体级规划和状态管理能力仍是决定性能的关键因素。

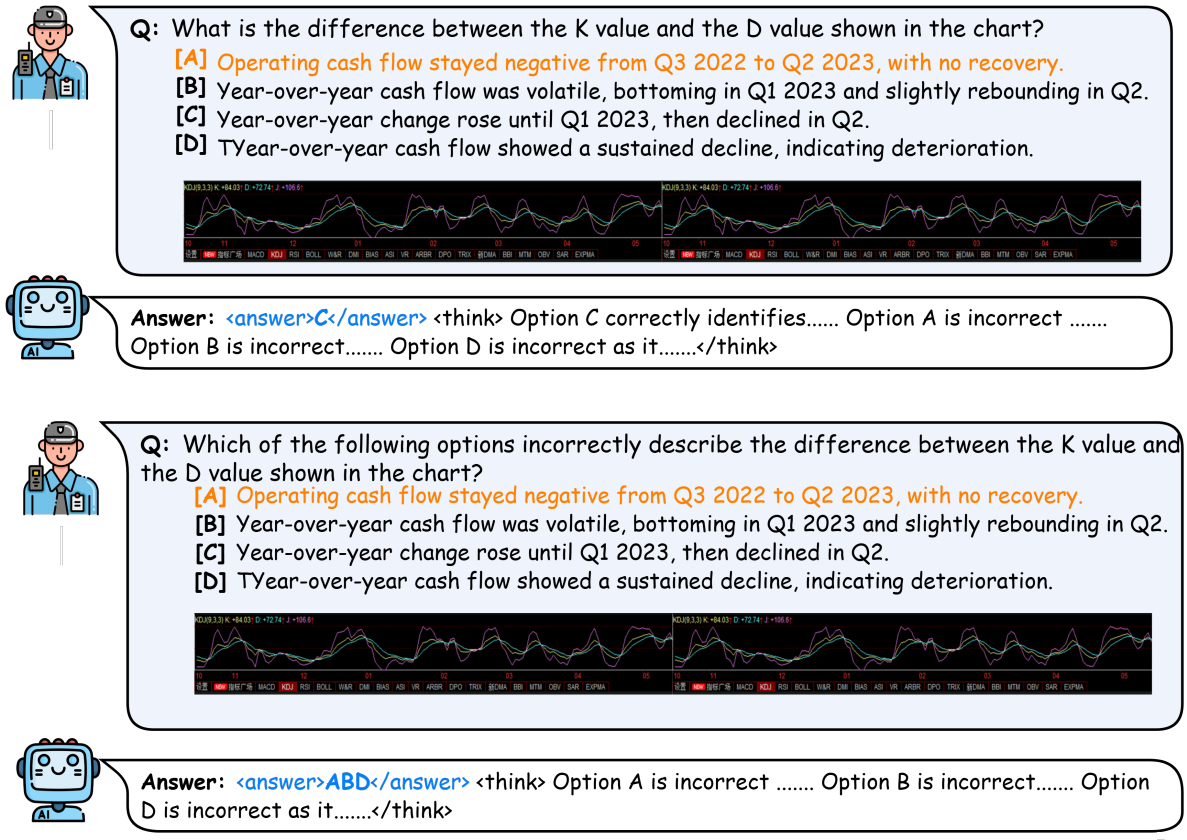

图7：FinMTM中具有代表性的客观题：基于金融图表的技术指标解读题，上方为单选题，下方为多选题（K–D曲线）。

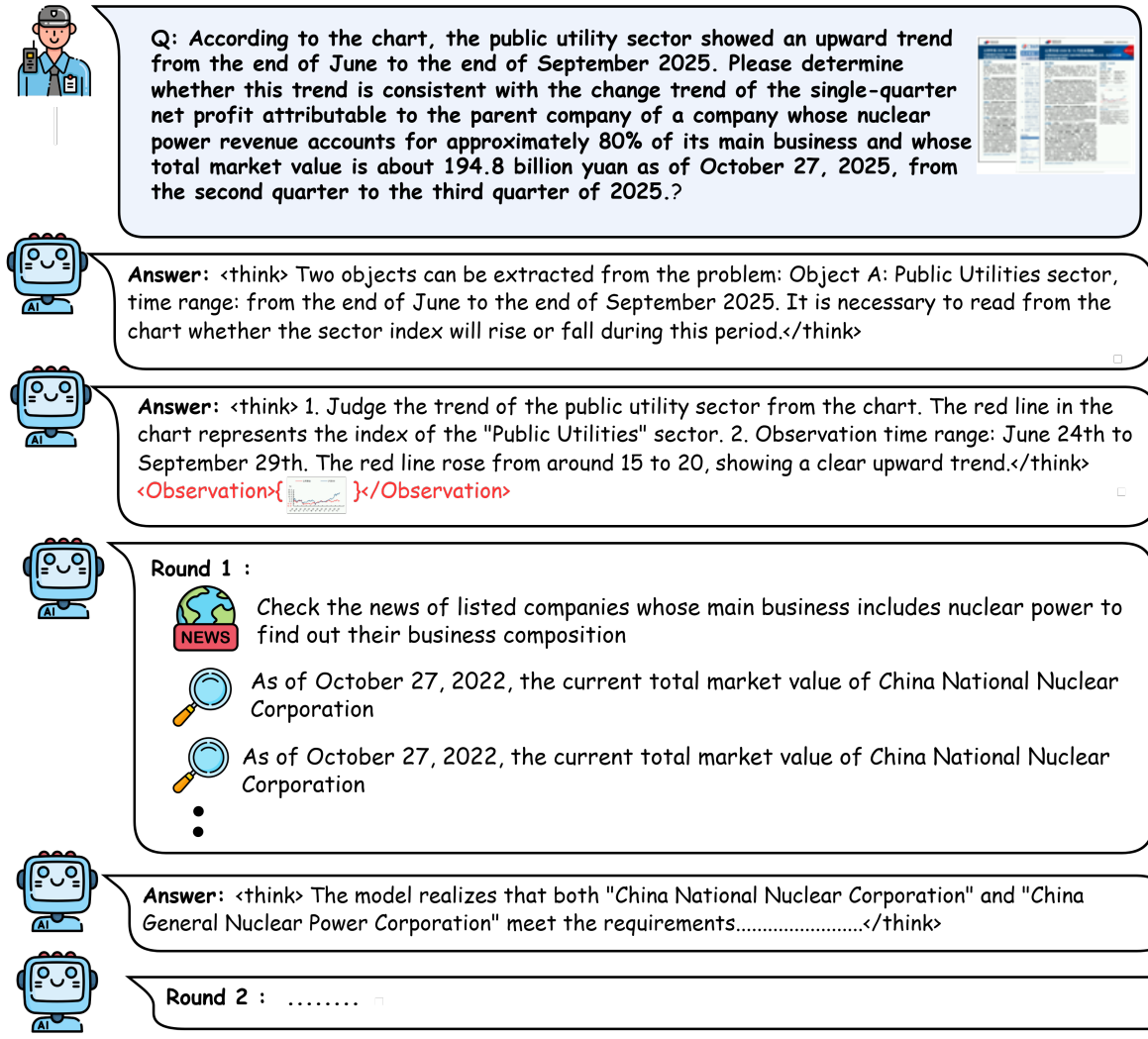

图 8：FinMTM 多轮规划智能体案例，展示了基于图表的推理、工具辅助实体搜索，以及对异构金融证据的一致性验证。

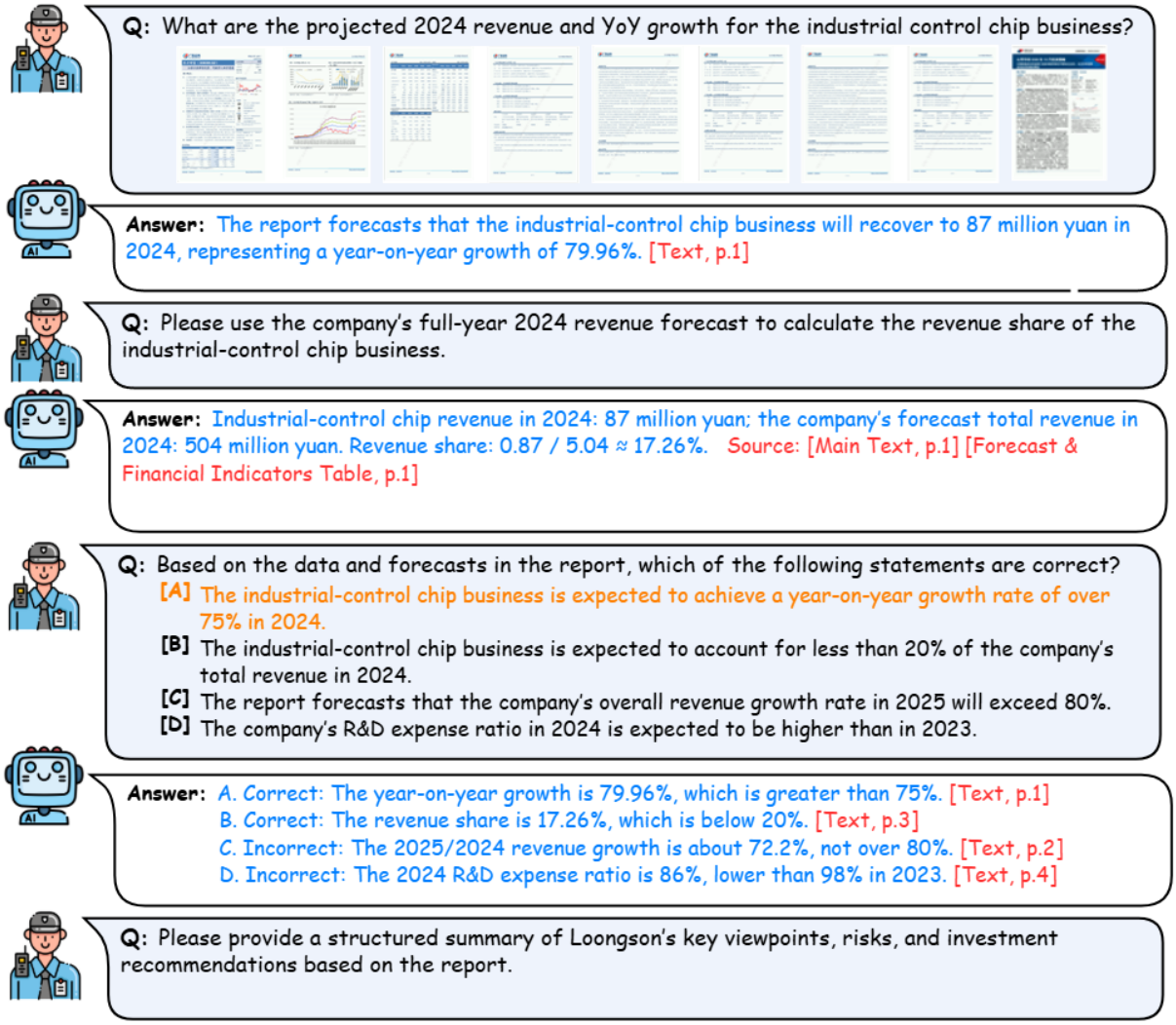

图9：一个具有代表性的 FinMTM 多轮开放式对话，以多页金融报告为依据，要求在多个轮次中进行跨页证据引用、数值提取、中间计算和陈述验证。

客观题推理提示

系统指令。 你是一位金融专家。请阅读问题和图像。仅输出一个 JSON 对象，不要添加任何额外文本、解释或反引号。

所需字段：

- 答案：选项中的最终选择
- 推理：简要说明你的推理

问题。 在 MACD 图表中，2024年5月1日之后两条曲线交叉了几次?

图像。网址

选项。

A. 24.05 之后四次交叉 B.
24.05 之后两次交叉 C.
24.05 之后三次交叉 D.
24.05 之后一次交叉

输出格式。

{" answer ":" C"," reason ":" 白线和黄线相交了三次。 "}

图10：金融图表客观题推理提示示例。

开放式问题的多轮推理提示

系统指令。 你是一名金融专家。请阅读问题、图像和对话历史。对于每一轮次，仅输出当前问题的答案。除非明确要求，否则不要包含解释。

图像。`url`

对话历史和当前轮次。

轮次 T1

- **问题：** 找出图表中红色折线的全局最大值，将其标记为点 A1，并报告其大致日期。

轮次 T2

- **背景：** 点 A1 是红色折线的全局最大值，大致日期为 2023-12-22。
- **问题：** 在点 A1 处，红色折线的值比黄色折线高多少 e?

第 T3 轮

- **背景：** 点 A1 是红色折线的全局最大值，日期约为 2023-12-22，原始红黄差值约为 60.0。
- **问题：** 假设由于数据修正，点 A1 处红色折线的实际值比观测值低 10%。请重新计算红色折线与黄色折线之间的差值。

第 T4 轮

- **上下文：** 点 A1 是红色折线的全局最大值，日期约为 2023-12-22，修正后的红黄差值约为 56.0。
- **问题：** 在与点 A1 相同的时间戳，绿色折线的值是否低于 -20.0?

预期输出（每轮）。

T1：2023 -12 -
22 T2：60.0 T3：
56.0 T4：是

图 11：开放式问题的零样本推理提示示例。

多轮金融智能体推理提示

(1) 系统指令（所有轮次共用）。该智能体被要求扮演一名金融多轮分析智能体。其目标是迭代规划 MCP 工具调用，并将工具输出与从输入图像中提取的视觉证据相结合，以回答查询。

可用工具。

- **FinQuery**：金融指标、价格和基本面。
- **Search**：一般实体或概念查询。
- **StockNews**：金融新闻检索。
- **ReportQuery**：分析师报告查阅。

输出约定。 在每个规划步骤中，智能体必须且只能输出一个 *JSON* 对象：

{ " Thought ": "高层推理与规划", " VisualObservation ": [从图像中读取的证据 "] ," ActionTrace ": [{" tool ": " FinQuery ", " query ": "..."} , {" tool ": " Search ", " query ": "..."}]}

如果任务已完成，智能体必须输出 <FINISHED> 并给出最终结论。

---

(2) 第一轮提示。 在第一轮中，智能体只会收到：(i) 系统指令，(ii) 输入的问题，以及 (iii) 图像（如果有）。不会提供先前的推理或工具反馈。

---

(3) 后续轮次的迭代提示。 对于第 $t > 1$轮，提示会补充已累积的证据来自前几轮：

<系统指令 >

前文 思考:
<思考_{t -1} >

视觉 观察:
<视觉观察 >

工具反馈:
<工具反馈 >

继续 规划 基于 上述 信息。
Do NOT 重复工具 调用。
如果 信息 已足够 , 输出 <FINISHED > 并 结束。

每轮都会重新附上同一张图像，以弥补无状态图像到文本接口的不足。

（4）强制摘要提示（终止回退）。如果智能体未能在最大规划轮次内输出 <FINISHED> ，则会触发强制摘要提示。在此阶段，不允许进行进一步的工具调用。

基于 以下 内容 视觉 观察和工具反馈，
总结 the 最终 答案。

问题: <查询 >

视觉观察: <视觉观察 >

工具反馈: <ToolFeedback >

提供 the 最终 结论并 输出 <FINISHED >。

此提示明确考查智能体在……条件下将多轮证据综合为连贯最终决策的能力严格的终止约束。

图 12：多轮金融智能体的零样本推理提示示例。