arXiv:2601.05039v1 [cs.MA] 8 Jan 2026

FINDEEPFORECAST：用于对金融预测领域深度研究智能体进行基准测试的实时多智能体系统

项目网站： https://OpenFinArena.com/

李翔宇*♣†、姚轩*♠、齐国豪*♣、朱凤斌*‡♣、Kelvin J.L. Koa♠、Xiang Yao Ng♡、刘子扬♡、倪星宇♠、刘畅♠、杨永辉♣、张扬♣、WENJIEWANG♢、FuliFeng♢、王超♡、Huanbo Luan♡、Xiaofen Xing†、Xiangmin Xu†、Tat-Seng Chua♣、黄克伟♠,

♣新加坡国立大学，新加坡♠亚洲数字金融研究院，
新加坡♡6Estates Pte Ltd，新加坡♢中国科学技术大学，中国

†中国华南理工大学

由先进大语言模型（LLM）驱动的深度研究（DR）智能体，已从根本上改变了完成复杂研究任务的范式。然而，在高风险领域（例如金融）中，对其在现实世界研究导向任务上的预测表现进行全面、实时的评估，仍未得到充分探索。我们推出 FINDEEPFORECAST，这是首个端到端多智能体系统，可通过持续生成研究导向的金融预测任务，自动评估深度研究智能体。该系统配备了双轨分类体系，能够在企业和宏观层面动态生成周期性与非周期性预测任务。借助该系统，我们构建了 FINDEEPFORECASTBENCH：一项为期十周的每周评估基准，涵盖 8 个全球经济体和 1,314 家上市公司，并对 13 种代表性方法进行了评估。大量实验表明，尽管深度研究智能体的表现始终优于强基线，但仍未达到真正前瞻性金融推理的水平。我们期待所提出的 FINDEEPFORECAST 系统能够持续推动深度研究智能体在研究导向的金融预测任务方面取得进展。该基准及排行榜已在 OpenFinArena平台公开。

ACM参考文献格式：
Xiangyu Li, 齐国豪, Xuan Yao, 朱凤斌, Kelvin J.L. Koa, Xiang Yao Ng, 刘子扬, 倪星宇, 刘畅, 杨永辉, 张扬, 王文杰, Fuli Feng, 王超, Huanbo Luan, Xiaofen Xing, Xiangmin

---

*同等贡献。
‡**项目负责人及通讯作者：朱凤斌，fengbin@nus.edu.sg**。作者联系方式：Xiangyu Li*♣†，Xuan Yao*♠，齐国豪*♣，朱凤斌*‡♣，Kelvin J.L. Koa♠，Xiang Yao Ng♡，刘子扬♡，倪星宇♠，ChangLiu♠，杨永辉♣，张扬♣，王文杰♢，FuliFeng♢，ChaoWang♡，HuanboLuan♡，XiaofenXing†，Xiangmin Xu†，Tat-Seng Chua♣，黄克伟♠，♣新加坡国立大学，新加坡♠亚洲数字金融研究院，新加坡♡6Estates私人有限公司，新加坡♢中国科学技术大学，中国

†中国华南理工大学。

特此授予许可，允许为个人或课堂使用而免费制作本作品的全部或部分数字或纸质副本，但前提是不得出于营利或商业利益制作或分发副本，且副本首页必须附有本声明及完整引文。对于本作品中由非作者所拥有的组成部分，必须尊重其版权。注明出处后进行摘要转载是允许的。如需以其他方式复制、重新出版、发布到服务器或再分发至列表，必须事先获得具体许可和/或支付费用。请发送邮件至 permissions@acm.org 申请许可。© 2026年 版权归所有者/作者所有。出版权已授权给 ACM。ACM XXXX-XXXX/2026/1-ART https://doi.org/10.1145/nnnnnnn.nnnnnnn

Xu†，Tat-Seng Chua♣，黄克伟♠。2026年。FinDeepForecast：用于金融预测中深度研究智能体基准测试的实时多智能体系统：**项目网站**：**https://OpenFinArena.com/**。1，1（2026年1月），44页。https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 引言

深度研究（DR）智能体是自主人工智能（AI）系统，可通过迭代规划、证据获取、推理和报告撰写来执行复杂研究任务 [7, 21]。它们的出现重塑了处理复杂任务的方式，并吸引了越来越多的研究关注。除了开发更强大的深度研究智能体之外，可靠地评估其研究能力也是推动该领域发展的同等重要的根本目标，但目前对此研究不足。传统的静态评估基准不可避免地会泄漏到训练语料库中，导致其失效（即数据污染） [1, 35]。近年来，人们开始探索实时基准，通过持续生成新实例，确保训练数据与评估数据在时间上相互分离 [31]。

现有实时基准分为对时间不敏感和对时间敏感两类。前者侧重于真实标注静态且确定的领域，例如代码生成 [9, 38] 和数学问题求解 [2], ，这往往会导致评估不够严格——模型可能依赖于回忆已有的答案模式，而非进行真正的推理。时效性基准 [10, 37] 围绕未来数据、未披露内容或未知结果构建，因此在评估前无法得知正确答案，从而能够严格评估预测推理。然而，这些基准中的任务通常取自现有问答网站，或使用固定模板生成。过度依赖外部来源或预先设定的模板可能会带来固有偏见，并限制真正研究导向型任务的可用性，进而制约评估的广度和深度。要真正对深度研究智能体的能力进行基准测试，就需要一个动态评估环境——持续提供**面向未来、研究导向型**的任务，并施加严格的时间约束，同时确保结果可客观验证。

金融领域为此类前瞻性、研究导向型任务提供了理想环境 [3, 13], ，其特性支持对深度研究智能体预测能力的持续评估。1）它提供定期信息披露，持续产生可验证的数据点，例如企业财务指标（例如，图 1（a）中的每股收益）和宏观经济（宏观）指标（例如，图 1（b）中的消费者价格指数）。2）它具有多样化市场事件，需要从海量、噪声繁多的数据中提炼有价值的信号和市场动态，以准确预测各种关键事件，包括企业行动（例如，图 1（c）中的新合作关系）和宏观变化（例如，图 1（d）中的出口管制）。3）它保证严格时间隔离，答案严格在信息披露或相应事件发生后才会出现。在此类金融市场中，金融专家必须收集并分析信息，通过推理预测未来结果，从而完成图 1 所示的复杂任务。

为持续评估深度研究智能体应对研究型金融预测挑战的能力，我们提出了一种名为 FINDEEPFORECAST 的新型实时多智能体系统。如图 2 所示，该系统采用双轨分类体系（更多详情见附录 A 和 B），以有效管理涵盖企业层面和宏观层面任务的周期性与非周期性预测情境。系统由六个专用智能体驱动，包含四个关键阶段，可自动端到端评估深度研究智能体，流程从数据收集和任务生成开始，历经模型预测，直至真实值获取和性能评估。为验证该系统，我们生成了 FINDEEPFORECASTBENCH 基准测试，其中宏观任务涵盖 8 个主要全球经济体，企业任务则涉及来自 9 个主要指数的 1,314 家上市公司。该基准测试总计包含 1,394 项任务，包括 296 项周期性宏观任务、723 项周期性企业任务、128 项非周期性宏观任务和 247 项非周期性企业任务。

我们评估了 13 种用于完成 FINDEEPFORECASTBENCH 每周任务的代表性方法，包括 3 个深度研究智能体、5 个兼具思考与搜索能力的大语言模型，以及 5 个具备思考能力的大语言模型。我们得出了几项重要发现。1）深度研究智能体的表现始终优于对比方法，但它们在 FINDEEPFORECASTBENCH 的任务上仍面临很大困难，最高得分为

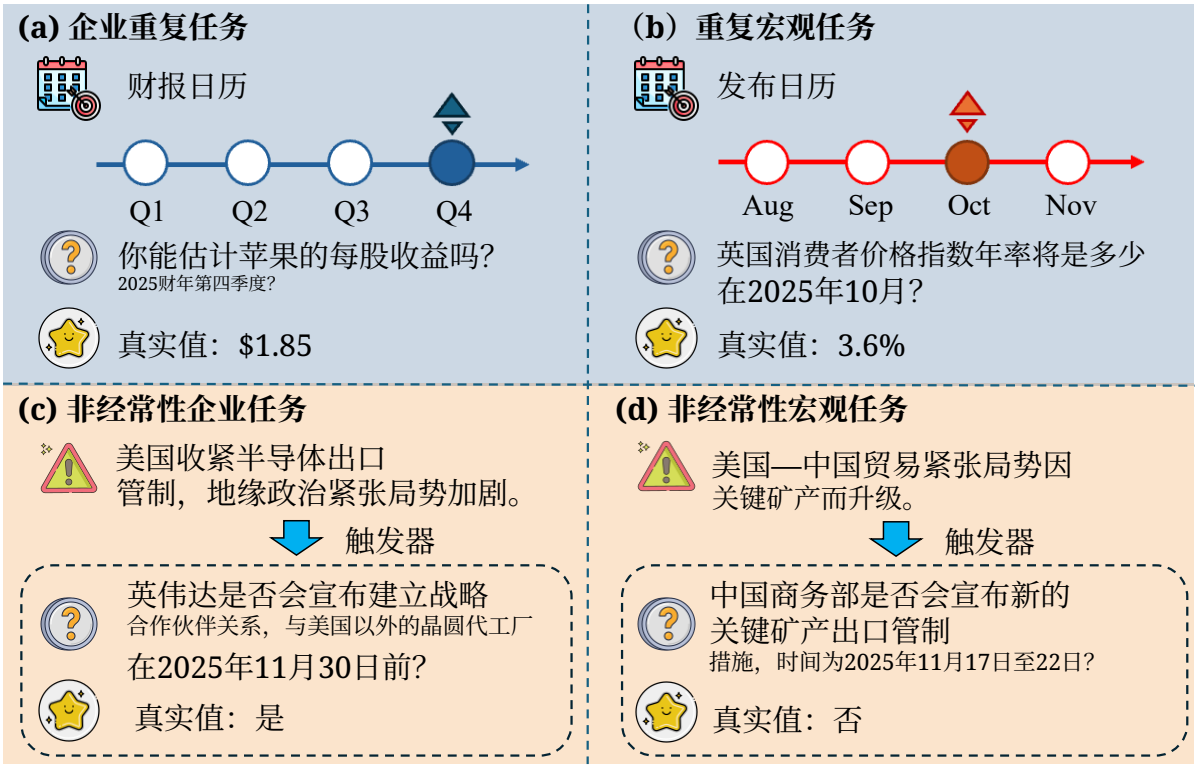

图 1. 用于定期披露的重复性任务，以及用于事件驱动预测的非重复性任务。

39.5 分（满分 100 分）。2）大多数方法在信息丰富的市场（例如，美国和中国）中表现最佳，但在数据或语言多样性相对有限的市场（例如，日本）中表现不佳。3）模型在非重复性任务上准确率较高，但在重复性任务上的表现则急剧下降。这种差异凸显了在定期披露的情况下，相较于二元事件预测，实现精准数值预测本身难度更大。

我们的贡献总结如下：

• 我们开发了 FINDEEPFORECAST，**这是首个端到端多智能体系统**，旨在**持续生成面向未来的研究导向型任务**，用于对深度研究智能体进行无污染评估。

• 我们提出了一种**双轨分类体系**，用于动态生成重复性和非重复性金融预测任务，涵盖实时市场环境中的企业层面和宏观层面预测（涉及数百项指标和事件类别）。

• 借助 FINDEEPFORECAST，我们生成了 FINDEEPFORECASTBENCH，这是一个**每周评估基准测试**，覆盖十周时间跨度，目前涵盖 8 个主要全球经济体和来自 9 个主要指数的 1,314 家上市公司，同时可轻松扩展至更多市场和公司。

• 对 13 种代表性方法进行的大量评估表明，尽管深度研究智能体显著优于其他方法，但仍有很大的改进空间，这凸显了当前方法在解决这些任务时的局限性。这充分确立了 FINDEEPFORECAST 的及时性和必要性，也使其成为持续推动深度研究智能体未来发展的重要贡献。

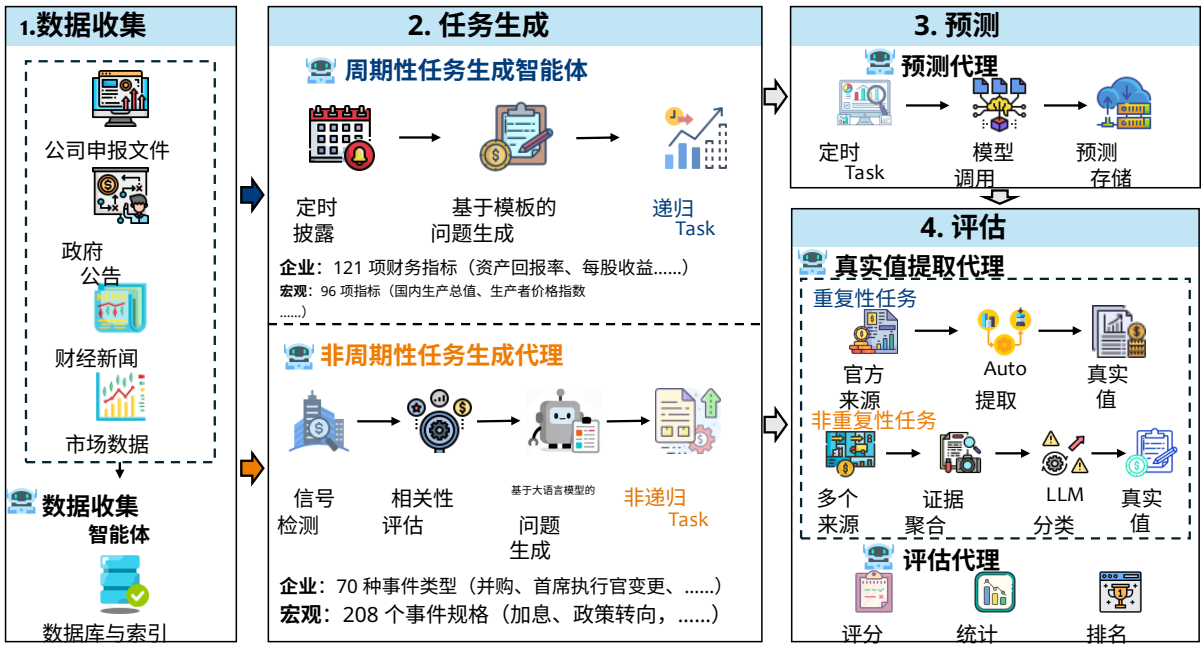

图 2. FinInDeepForecast 系统包含四个阶段：（1）数据收集将财务信息汇总至带时间戳的数据库；（2）任务生成通过基于模板的问题生成产生重复性任务，并通过基于大语言模型的流程生成非重复性任务；（3）预测在时间隔离条件下调用模型；（4）评估提取真实值并计算性能指标。

2 FINDEEPFORECAST 系统

本节介绍 FINDEEPFORECAST，这是一个实时多智能体系统，通过面向研究的任务生成、严格时间隔离和严谨的真实值验证，评估深度研究智能体在金融预测方面的实际能力，如图 2 所示。在 FINDEEPFORECAST 中，我们设计了双轨分类体系，以区分针对定期披露进行数值估算的重复性预测，以及针对不确定的新兴事件进行二元分类的非重复性预测。具备时间隔离的持续生成 通过实时创建任务防止数据污染，同时确保不同模型遵循统一的信息边界。

2.1 预测问题定义

预测问题定义为一个元组 $P = (q, t_g, t_d, t_e, y)$，其中 q 表示预测问题，t_g 表示任务生成时间，t_d 表示预测截止时间，t_e 表示评估时间，y 表示真实结果。时间顺序 $t_g < t_d < t_e$ 确保在结果变得可观测之前作出预测。给定一个预测问题，模型会生成一个预测结果 $\hat{y} = f(q, I_{t_d})$，其中 f 表示模型的预测函数，I_{t_d} 表示截至截止时间 t_d 可用的信息集。

我们的双轨分类体系主要依据时间可预测性来区分预测问题：周期性预测针对时间已知但数值结果不确定的定期披露，而非周期性预测则针对其发生本身无法根据日历预判的事件。在每条轨道中，任务还会按预测范围 进一步划分为企业层面和宏观层面，从而形成四个互为补充的评估维度。

2.2 数据收集

依托亚洲数字金融研究院（AIDF）提供的数据基础设施，数据收集智能体持续监测并收集四类信息：1）来自监管数据库的公司申报文件；2）来自统计机构的政府公告；3）来自实时信息流的财经新闻；4）来自交易所的市场数据。所有收集的数据均整理至带时间戳的数据库和索引中，从而在评估期间实现时间隔离：限制模型只能访问在预测截止时间之前发布的内容。

2.3 任务生成

任务生成采用两个专用智能体，分别对应我们的双轨分类体系。

周期性任务生成。 周期性任务生成智能体通过两阶段流程构建重复性任务：首先从官方日历中识别定期披露事项，然后应用基于模板的问题生成，生成带有时间参数的任务（t_g， t_d， t_e）。宏观层面任务监测 14 项指标（例如，GDP增长、PPI变化），企业层面任务则针对 121 项财务指标（例如，资产回报率、每股收益）。完整规格见附录 A。

非周期性任务生成。 非周期性任务生成代理采用三阶段的基于大语言模型的流程：(1) 信号检测从新闻信息流中识别指标；(2) 相关性评估判断其预测显著性；(3) 基于大语言模型的问题生成根据明确的事件定义生成任务。宏观任务采用包含 9 个高层类别和 26 个细粒度子类别的分类体系，并通过核心—自适应接口将其具体化为经济体特定事件规范（例如，加息、政策转变）。企业任务围绕 70 种精选事件类型定义（例如，并购、首席执行官变更），这些事件具有明确的预测语义和客观可验证的结果，并具体落实到各个上市公司层面。完整分类体系见附录 B，生成流程的实现细节见附录 C。

2.4 预测

预测代理通过三步工作流程征集预测。首先，`任务调度` 根据生成任务的预测截止时间 t_d 对任务进行组织，并将其分配到每周评估批次。其次，`模型调用` 通过应用程序接口，使用标准化提示词调用每个待评估模型；提示词中包含预测问题和截止时间。为确保时间隔离，搜索增强模型和深度研究模型仅配置为访问 t_d之前发布的内容。第三，`预测存储` 将所有模型输出连同时间戳记录到结构化数据库中，以供后续评估。

2.5 评估

为确保评估的客观性并模拟现实世界中的财务问责，我们采用由两个智能体管理的确定性、以结果为导向的协议。

真实标签提取。 对于非重复性任务，我们采用人在回路协议以避免评估偏差：大语言模型智能体首先汇总多源证据，提出可能的结果；随后，**由领域专家对所有样本进行严格核验**，以确定最终真实值。该核验协议的详情见附录 D。

评分指标。 每项预测任务均使用二元评分函数进行评估，预测正确得 1 分，否则得 0 分：

$$\text{Score}(y, \hat{y}) = \begin{cases} \mathbf{1}\left[\left|\frac{\hat{y}-y}{y}\right| < \epsilon_k\right] & \text{if recurrent} \\ \mathbf{1}\left[\hat{y} = y\right] & \text{if non-recurrent} \end{cases} \tag{1}$$

其中 $\mathbf{1}[\cdot]$ 表示 指示函数， y 是真实值， $\hat{y}$ 是模型预测， ϵ_k 是指示 阈值。对于重复性任务，如果相对误差在阈值范围内，则预测结果正确 ld; 函数特定的容差

对于非重复性任务，则必须完全匹配。总体准确率按正确预测结果所占的百分比计算：

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \text{Score}(y_i, \hat{y}_i) \times 100\% \tag{2}$$

其中 N 是评估任务的总数。结果按任务类别、预测时域和市场区域细分。

3 FINDEEPFORECASTBENCH

为验证 FINDEEPFORECAST，我们推出了 FINDEEPFORECASTBENCH，这是一个为期十周、覆盖 8 个主要市场的每周评估基准。本节介绍 FINDEEPFORECASTBENCH 的构建与质量控制，将其与现有基准进行比较，并将详细统计数据留待附录 E 介绍。

3.1 FINDEEPFORECASTBENCH 生成

我们在以下设置下，使用所提出的 FINDEEPFORECAST 系统实例化 FINDEEPFORECASTBENCH。

市场与公司选择。 根据经济重要性和数据可用性选择市场。目前的实例覆盖三大洲和**八个主要经济体**：美国、中国、中国香港、日本、英国、德国、法国和新加坡。为建立标准化的评估范围，我们依据所涵盖经济体中的九个主要股票指数来选择公司（标准普尔500指数、纳斯达克100指数、富时100指数、DAX 40指数、CAC 40指数、日经225指数、沪深300指数、恒生指数和海峡时报指数）。由此形成的公司池包含1,314家成分股公司，其定义依据固定快照日期（2025年10月2日）的指数成分资格。

每周任务生成。 自2025年10月27日起，基准测试持续运行，每周四发布一个新的任务批次。对于周期性企业任务，我们采用动态分层抽样策略，每周从各市场的披露公司中最多抽取30%，以平衡各地区的任务密度；对于重复宏观任务，我们涵盖八个经济体的所有预定指标发布。对于非重复性任务，候选任务根据实时新闻流生成，并每周接受专家审查；只有质量最高的事件才会被选中纳入，以确保其预测价值和可回答性。

真实值获取。 真实值获取按每周滚动周期运行，每周一处理评估时间 t_e 已到的任务。对于非重复性任务，领域专家会核验自动分类结果，以确保其有效性。为应对延迟披露等现实世界中的不规则情况，结果不确定的任务会被标记为待定，并每周重新审查。在2周有效期窗口后仍未解决的任务将被标记为无效，并从性能评估中排除。

3.2 质量控制

我们通过在流程的每个阶段实施系统性的质量控制，确保基准可靠性。

专家参与。 领域专家参与基准生命周期的各个环节。一支接受过金融和经济学博士级培训的研究人员团队参与多个阶段：为周期性任务生成设计标准化模板，为非重复性任务定义事件分类体系和早期信号标准，审查生成的问题，并核验真实值。

任务质量保证。 对于重复性任务，由领域专家验证的标准化模板可确保指标定义的一致性，并与监管报告标准保持一致。对于非重复性任务，专家每周审查所有候选问题，并应用严格的筛选标准，确保最终任务集兼具质量和跨市场平衡。

真实值验证。 对于重复性任务，真实值通过自动化方式从官方来源中提取d解析。抽样对一手来源进行验证后确认，准确率达到 99.8%，少数差异可归因于le

表 1. FINDEEPFORECASTBENCH 与现有基准的比较。“T-S”和“T-IS”分别表示“时间敏感型”和“时间非敏感型”。

	领域	Type	问题	频率
金融预测				
FLUE	金融	T-IS	精选	一次性
PIXIU	金融	T-IS	精选	一次性
FinanceBench	金融	T-IS	标注	一次性
FinBen	金融	T-IS	精选	一次性
FinCall	金融	T-IS	基于规则	一次性
实时基准				
LiveBench	Misc	T-IS	精选	每月
LiveCodeBench	编码	T-IS	基于规则	每月
SWE-bench	软件	T-IS	基于规则	季节性
MathArena	Math	T-IS	精选	季节性
CryptoBench	加密货币	T-IS	标注	每月
LiveXiv	科学	T-IS	基于规则	每月
ForecastBench	通用	T-S	基于规则	每周
FutureX	通用	T-S	基于规则	每周
Ours				
FINDEEPFORECAST	**金融**	**T-S**	**动态**	**每周**

至后续数据 发行方重述。对于非重复性任务，真实值的判定结合了自动 d
证据汇总 并经专家验证，标注者间一致性达到 95%。

3.3 与现有基准的比较

表 1 从几个关键维度比较了 FINDEEPFORECASTBENCH 与现有基准。我们有以下观察：1）现有金融预测基准主要侧重于不受时间影响且周期性 的预测任务。这些任务通常在历史数据上以一次性方式进行评估，此时测试阶段的真实值已为已知，因此容易受到数据污染的影响。2）大多数现有实时基准都不受时间影响 ，涉及编码、软件工程和数学等确定性领域，因此评估不够严格，因为模型可能依赖对既有答案模式的记忆。3）具有时效性 的基准是评估预测任务的重要进展；然而，它们很大程度上依赖固定模板或基于规则的提取流程来生成任务，这限制了任务多样性，也制约了基准反映不断演变的现实情境的能力。

关键差异。 FINDEEPFORECASTBENCH 通过以下三个特点区别于以往基准：1）采用双轨分类体系，同时支持周期性和非周期性预测任务，涵盖定期披露和事件驱动预测；2）专注于金融领域，真实值可根据官方申报文件、统计发布和权威披露进行客观验证；3）在实时市场环境中动态生成任务并每周更新，从而能够持续、以研究为导向且无污染地评估深度研究智能体。

4 个实验

我们全面评估了最先进方法的金融预测能力。

4.1 评估模型

我们评估了涵盖三种范式的 13 个模型，这些范式具有不同的信息获取能力。**具备思考能力的大语言模型（T）**。OpenAI GPT-5（T） [27],Claude-Sonnet-4.5（T） [25],Gemini 2.5Pro（T） [11],Deepseek- v3.2（T） [26]和 Grok 4（T） [32]。**具备思考能力的大语言模型 + 搜索（T+S）**。OpenAI GPT-5（T+S） [27], Claude-Sonnet-4.5（T+S） [25], Gemini 2.5 Pro（T+S） [11],Deepseek-v3.2（T+S） [26],和 Grok 4（T+S） [32]。**深度研究**。OpenAI o3-deep-research [16], Perplexity Sonar Deep Research [28]和通义 Deep Re- search [30]。

4.2 实施细节

时间隔离。 为确保公平比较，我们在任务生成、模型预测和性能评估之间实施严格时间隔离。所有受评估模型只能访问预测截止时间 t_d 之前发布的内容，从而避免获取预测时尚不可用的信息。

任务生成。 任务每周四生成，预测截止时间 t_d 设为接下来星期日 23:59（UTC+8）。对于周期性企业任务，我们会选择预测窗口内安排发布财报的公司。对于非重复性任务，领域专家会审核候选问题，并根据预测质量和市场平衡情况，选取约 20% 纳入任务。

模型预测。 所有模型都会收到标准化提示词，其中指定预测问题、截止时间和所需输出格式。模型会生成结构化输出：重复性任务为数值估计 $\hat{y} \in \mathbb{R}$，非重复性任务为二元预测 $\hat{y} \in \{\text{Yes}, \text{No}\}$。输入和输出示例见附录 H。预测任务的模型设置会在适用情况下启用思考和网络搜索功能。除非另有说明，所有参数均采用默认值。详细的模型配置见附录 F。

答案评估。 对于重复性任务，我们根据单位类型和指标类别应用指标特定阈值 ϵ_k。按单位类型划分，百万级财务指标的阈值设为 5%，百分比和比率指标的阈值设为 1%。按指标类别划分，利率和汇率的阈值设为 0.1%，其他宏观指标的阈值设为 1%。对于非重复性任务，真实值通过汇总多个来源的证据确定，并经专家验证，标注者间一致性达到 95%。

4.3 主要结果

我们在为期 10 周的时间段内生成的近 1,400 项预测任务上评估模型性能，准确率以预测正确的比例衡量。图 3 的实证结果揭示了三项关键发现：1）与另外两类模型相比，深度研究模型展现出明显的性能优势。OpenAI o3-deep-research（39.5%）和 Perplexity Sonar 深度研究（39.4%）的表现优于所有其他架构方法；二者的成绩几乎持平，表明这一范式在前沿领域正趋于收敛。2）虽然将搜索功能融入推理有所助益，但仍无法复现深度研究的表现。领先的思考 + 搜索（T+S）模型 GPT-5（36.0%）、Claude-Sonnet-4.5（35.9%）和 Gemini 2.5 Pro（35.0%）表现集中在一个较窄区间内，但始终比深度研究系统低 3–4 个百分点，这表明仅靠搜索增强无法充分发挥专门设计的深度研究架构的全部优势。3）外部信息检索能力对基准测试性能至关重要。对 T 和 T+S 配置进行模型内比较后发现，移除搜索会导致准确率系统性下降，降幅从 11.0 个百分点（GPT-5：36.0% → 25.0%）到 14.2 个百分点（Gemini 2.5 Pro：35.0% → 20.8%）不等。所有受评估模型均呈现这一一致模式，充分证明获取外部信息是此基准测试中取得成功的基本决定因素，仅凭纯推理能力并不足够。

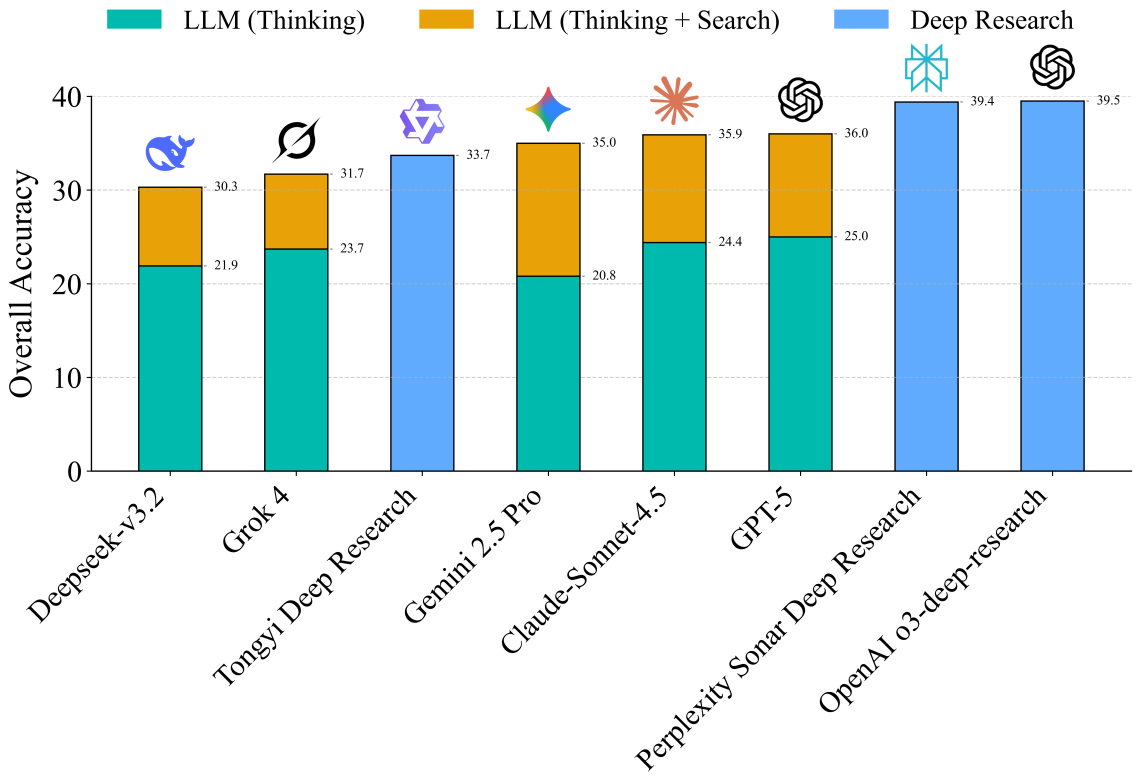

图 3. 主要结果。整个十周时间跨度内的模型性能对比 .

4.4 深入分析

不同任务上的性能分析。我们分析模型在不同情境（即重复性和非重复性）及不同层级（即企业和宏观）下的性能。结果见表 2，由此得出以下主要发现：1）仅依靠思考的大语言模型在非重复性任务上能达到尚可的准确率，但在重复性场景中的表现却急剧下滑（整体准确率通常低于 10%），这表明仅靠内部推理不足以应对具有时间依据、细粒度的金融预测。2）具备思考与搜索能力的大语言模型在所有任务类型上的表现均优于仅依靠思考的模型，说明获取外部信息十分重要。然而，它们在重复性任务上的提升仍然有限，这意味着单靠信息检索不足以应对 FINDEEPFORECASTBENCH 中的挑战。4）深度研究智能体在非重复性和重复性任务上均表现最佳，尤其是在企业和宏观重复性预测方面。这表明，多步规划、证据综合和结构化推理共同作用，有助于在严格时间隔离下实现更准确的预测。4）在所有范式中，模型在非重复性任务上的准确率较高（最高达 81.4%），而在重复性任务上的表现则大幅下降，最佳方法的整体准确率也仅为 25.5%。这凸显了在定期披露情境下进行精确数值预测，相较于二元事件预测具有内在难度。为更好地理解其失败模式，我们在附录 G 中提供了一项错误案例研究。

表 2. 非递归（Non-rec.）和递归（Rec.）情境下的性能分析。报告的数值为总体准确率。

	非递归			Rec.		
方法	**公司**	**Mac.**	**Ovr.**	**公司**	**Mac.**	**Ovr.**
大语言模型（思考）						
OpenAI GPT-5（T）	68.4	65.6	67.5	6.8	4.1	6.0
Claude-Sonnet-4.5（T）	66.8	70.3	68.0	8.4	1.0	6.3
Grok 4（T）	73.7	71.2	73.1	11.5	2.7	9.0
Deepseek-v3.2（T）	61.9	58.6	60.8	6.9	2.0	5.6
Gemini 2.5 Pro（T）	73.3	68.8	71.7	8.5	1.0	6.4
大语言模型（思考 + 搜索）						
OpenAI GPT-5（T+S）	78.1	72.7	76.3	22.8	11.1	19.5
Claude-Sonnet-4.5（T+S）	79.8	73.4	77.6	20.7	19.6	20.4
Grok 4（T+S）	74.5	77.3	75.5	15.1	18.2	16.0
Deepseek-v3.2（T+S）	76.9	70.3	74.7	13.4	14.6	13.7
Gemini 2.5 Pro（T+S）	78.5	**77.5**	78.3	23.3	17.6	21.7
深度研究						
Perplexity Sonar	**81.4**	75.0	**79.2**	26.2	**23.7**	**25.5**
通义深度研究	79.8	74.2	77.9	23.5	15.5	21.2
OpenAI o3-deep	80.6	75.8	78.9	**26.7**	21.3	25.2

不同市场的性能分析。 随后，我们分析不同市场中的模型性能，并将结果呈现在表 3 中，由此可以观察到：1）深度研究智能体在几乎所有市场中都取得了最高准确率，表明其在异质监管制度、披露标准和信息环境中的泛化能力更强。2）在每个市场中，增加搜索能力都会带来显著提升，这凸显了在实时金融预测场景下获取最新市场特定信息的重要性。3）大多数方法在美国和中国等信息丰富的市场中表现最佳，而在可用数据相对较少或语言多样性较低的市场（如日本）中，准确率始终较低。

每周性能分析。 我们分析了全部十周内的模型性能，并将结果呈现在图 4 中。我们可以观察到：1）随着披露期结束后重复性任务占比下降，准确率逐周稳步提高，这与表 2 所示所有模型在非重复性任务上的更强表现一致。2）深度研究智能体在所有周次中始终优于其他方法，表明其整合观测信号并随时间进行适应的能力更强且更稳定。

5 相关工作

5.1 深度研究智能体

深度研究（DR）智能体旨在通过规划、信息收集和多步推理来解决复杂任务，并且近期已广泛应用于工业界和开源大语言模型系统 [4, 15, 17,18, 29]。早期研究 [22, 36] 首次确立了将推理与工具使用及环境交互交织结合的核心智能体范式。随后，这些范式被拓展至更贴近现实的领域，例如基于网络的信息检索和软件工程 [6, 40]。与此同时，也有研究提出基准 [14, 41] ，用于评估

表3。8个金融市场的性能分析。所报告的数值表示总体准确率。

方法	US	CN	HK	JP	UK	DE	FR	SG
大语言模型（思考）								
OpenAI GPT-5（T）	19.0	32.3	26.7	14.4	28.6	47.6	39.1	32.6
Claude-Sonnet-4.5（T）	18.1	28.5	28.3	16.7	31.0	44.0	32.7	32.6
Grok 4（T）	20.2	26.9	20.8	17.5	31.0	39.3	26.4	31.5
Deepseek-v3.2（T）	16.5	26.9	25.8	12.5	27.8	42.9	30.9	28.1
Gemini 2.5 Pro（T）	18.7	24.6	19.2	12.8	26.2	39.3	21.8	25.8
大语言模型（思考 + 搜索）								
OpenAI GPT-5（T+S）	36.7	40.0	35.8	22.6	38.9	47.6	44.5	39.3
Claude-Sonnet-4.5（T+S）	33.4	42.3	42.5	23.7	39.7	48.8	40.0	43.8
Grok 4（T+S）	29.2	40.8	34.2	16.7	37.3	45.2	37.3	44.9
Deepseek-v3.2（T+S）	26.4	36.9	40.3	17.1	33.3	45.8	37.3	39.3
Gemini 2.5 Pro (T+S)	35.7	38.0	30.8	22.0	34.7	47.5	44.5	**47.7**
深度研究								
Perplexity Sonar	40.1	**45.7**	**47.9**	24.3	**40.0**	50.0	40.9	44.9
通义深度研究	33.9	44.6	29.4	20.7	38.7	47.0	36.4	36.4
OpenAI o3-deep	**41.0**	45.4	40.0	**25.3**	38.9	**51.2**	**48.2**	41.6

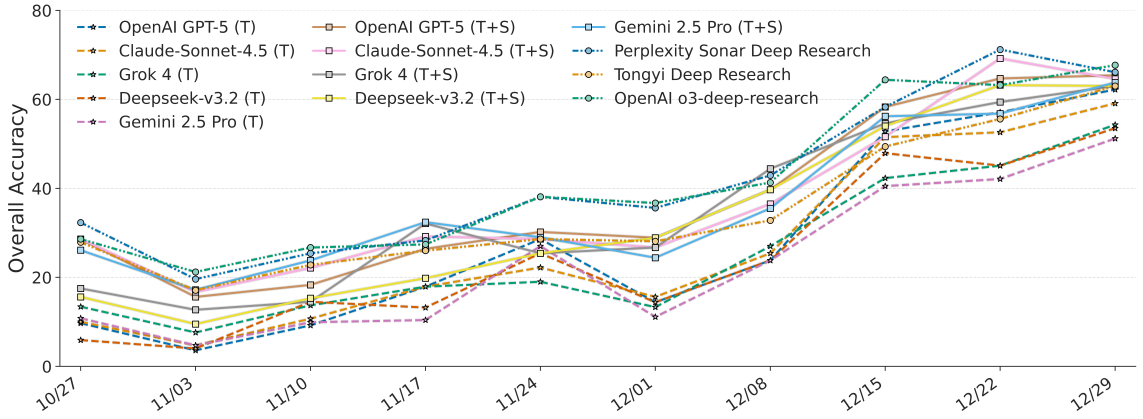

图4. 每周性能比较。

智能体在受控环境中的能力。然而，现有的大多数评估都依赖静态任务集，因此难以捕捉智能体在不断变化的环境中的行为。随着深度研究智能体日益广泛地运行于不断演变的现实环境中，如何在动态且受时间制约的条件下可靠地评估其表现，已成为亟待解决的关键挑战，而这正是本研究旨在应对的问题。

5.2 实时基准

大语言模型的实时基准测试已成为缓解评估中数据污染问题的一个重要方向[1]。现有的实时基准大致可分为时间不敏感和时间敏感任务。时间不敏感基准 [2, 9, 31, 38] 旨在通过持续更新来缓解数据污染，但其任务依赖于不受未来结果影响的确定性真实值。相比之下，时间敏感基准[10, 37] 评估模型针对测试时答案尚未知晓的问题进行预测推理的能力。然而，这些基准通常依靠人工整理或基于规则的提取流程来构建任务，这限制了任务的多样性和适应性。我们的工作提出了一种时间敏感的金融预测基准，其任务根据不断变化的现实市场环境动态生成。

5.3 金融预测基准

当前的金融预测基准 [8, 20] 通常聚焦于递归事件，即定期发生的目标，例如股价变动 [33, 34] 或公司收益 [23]。这些基准基于静态历史数据集构建，因此引发了人们对基于大语言模型的解决方案产生数据污染 [19] 的担忧。另一个重要但较少见的基准针对非递归任务，即同样会影响金融市场 [5, 12], 的离散事件，例如新的合作关系或关税。这些基准还可按企业层级 [42]或宏观层级 [39]任务分类，两者在事件规模和粒度方面有所不同。我们的工作在统一的评估框架下处理这些任务，该框架基于实时数据构建。

6 结论

本文介绍了 FINDEEPFORECAST，这是首个用于评估金融预测领域深度研究智能体的实时端到端多智能体系统。它能够在严格时间隔离的条件下，持续生成面向未来的研究导向型任务，并将任务创建、模型调用和真实值验证整合到统一的全自动化流程中。基于该系统，我们构建了 FINDEEPFORECASTBENCH，这是一个每周运行的基准测试，涵盖企业和宏观经济层面的递归数值披露与非递归事件驱动预测。我们评估了 13 个具有代表性的系统，结果表明，当前的深度研究智能体在真正的研究导向型金融预测任务上仍面临挑战，尤其是在精确的递归数值预测方面。FINDEEPFORECAST 为实时市场环境中的深度研究智能体建立了一种动态、无污染的评估范式，并为未来系统和基准的持续构建与扩展奠定了基础。

未来，我们计划将该系统扩展至更丰富的任务形式，包括概率预测、多步预测和投资组合层面预测，并引入基于过程的评估，以更好地理解深度研究智能体在实时预测情境中的搜索、推理和失败方式。

7 贡献

- **项目负责人：** 朱凤斌
- **主要贡献者：** 李翔宇、姚轩、齐国豪
- **次要贡献者：** Kelvin J.L. Koa、Xiang Yao Ng、刘子扬、倪星宇、刘畅、杨永辉、YangZhang
- **顾问：** 王文杰、Fuli Feng、王超、Huanbo Luan、Xiaofen Xing、Xiangmin Xu、Tat-Seng Chua 和黄克伟。

参考文献

[1] Simone Balloccu、Patrícia Schmidtová、Mateusz Lango 和 Ondrej Dusek。2024年。《泄露、作弊、重复：闭源大语言模型中的数据污染与评估不当行为》。载于《计算语言学协会欧洲分会第18届会议论文集》。67–93页。[2] Mislav Balunovi´c、Jasper Dekoninck、Ivo Petrov、Nikola Jovanovi´c 和 Martin Vechev。2025年。《Matharena：在未受污染的数学竞赛上评估大语言模型》。arXiv预印本 arXiv:2505.23281（2025年）。[3] Yuemin Chen、Feifan Wu、Jingwei Wang、Hao Qian、Ziqi Liu、Zhiqiang Zhang、Jun Zhou 和 Meng Wang。2024年。《知识增强的金融市场分析与报告生成》。载于《2024年自然语言处理经验方法会议：工业赛道论文集》。1207–1217页。[4] Dave Citron。2025年。《Gemini 2.5 Pro Experimental现已支持深度研究》。https://blog.google/products/gemini/deep- researchgemini-2-5-pro-experimental/ 访问日期：2025年。[5] Ziniu Hu、Weiqing Liu、Jiang Bian、Xuanzhe Liu 和 Tie-Yan Liu。2018年。《倾听混沌的低语：用于新闻导向的股票趋势预测的深度学习框架》。载于《第十一届ACM国际网络搜索与数据挖掘会议论文集》。261–269页。[6] Dong Huang、Jie M Zhang、Michael Luck、Qingwen Bu、Yuhao Qing 和 Heming Cui。2023年。《Agentcoder：采用迭代测试与优化的多智能体代码生成》。arXiv预印本 arXiv:2312.13010（2023年）。[7] Yuxuan Huang、Yihang Chen、Haozheng Zhang、Kang Li、Huichi Zhou、Meng Fang、Linyi Yang、Xiaoguang Li、Lifeng Shang、Songcen Xu、Jianye Hao、Kun Shao 和 Jun Wang。2025年。《深度研究智能体：系统性考察与路线图》。arXiv:2506.18096 [cs.AI] https://arxiv.org/abs/ 2506 18096 。Pranab Islam、Anand Kannappan、Douwe Kiber、Zachary Walters、Scott Kantor[8]、Tom Sun 和 Nils Holmes。2023年。《FinanceBench：金融问答的新基准》。arXiv预印本 arXiv:2311.11944（2023年）。[9] Naman Jain、King Han、Alex Gu、Wen-Ding Li、Fanjia Yan、Tianjun Zhang、Sida Wang、Armando Solar-Lezama、Koushik Sen 和 Ion Stoica。2024年。《LiveCodeBench：面向代码的大语言模型整体性且无污染评估》。载于《计算语言学协会第62届年会论文集》。[10] Ezra Karger、Houtan Bastani、Chen Yueh-Han、Zac hary Jacobs、Danny Halawi、Fred Zhang 和 Philip E Tetlock。2024年。《ForecastBench：人工智能预测能力的动态基准》。arXiv预印本 arXiv:2409.19839（2024年）。[11] Koray Kavukcuoglu。2025年。《Gemini 2.5：我们最智能的人工智能模型》。https://blog.google/technology/google-deepmind/gemini-model- thinking-updates-march-2025/。访问日期：2025-10-07。[12] Kelvin JL Koa、Yunshan Ma、Ritchie Ng 和 Tat-Seng Chua。2024年。《利用自我反思型大语言模型学习生成可解释的股票预测》。载于《ACM网络会议2024论文集》。4304–4315页。[13] Ross Koval、Nicholas Andrews 和 Xifeng Yan。2024年。《基于文本和表格时间序列的金融预测》。载于《计算语言学协会论文集：EMNLP 2024》。8289–8300页。[14] Xiao Liu、Hao Yu、Hanchen Zhang、Yifan Xu、Xuanyu Lei、Hanyu Lai、Yu Gu、Hangliang Ding、Kaiwen Men、Kejuan Yang 等。2023年。《Agentbench：将大语言模型作为智能体进行评估》。arXiv预印本 arXiv:2308.03688（2023年）。[15] OpenAI。2025年。《推出深度研究》。https://openai.com/index/introducing-deep-research/ 访问日期：2025年。[16] OpenAI 团队。2025年。《推出深度研究》。https://openai.com/index/introducing-deep-research/。访问日期：2025-10-07。[17] Perplexity。2025年。《推出Perplexity 深度研究》。https://www.perplexity.ai/ja/hub/blog/introducing-perplexity-deepresearch 访问日期：2025年。[18] 通义千问。2025年。《深度研究（Qwen-Deep-Research）》。https://www.alibabacloud.com/help/en/model-studio/qwen-deep-research 访问日期：2025年。[19] Oscar Sainz、Jon Ander Campos、Iker García-Ferrero、Julen Etxaniz、Oier Lopez de Lacalle 和 Eneko Agirre。2023年。《自然语言处理评估陷入困境：为何需要针对每个基准测量大语言模型数据污染》。arXiv预印本 arXiv:2310.18018（2023年）。[20] Raj Sanjay Shah、Kunal Chawla、Dheeraj Eidnani 等。2022年。《当FLUE遇上FLANG：金融领域的基准与大型预训练语言模型》。arXiv预印本 arXiv:2211.00083（2022年）。[21] Zhengliang Shi、Yiqun Chen、Haitao Li、Weiwei Sun、Shiyu Ni、Yougang Lyu、Run-Ze Fan、Bowen Jin、Yixuan Weng、Minjun Zhu 等。2025年。《深度研究：系统综述》。arXiv预印本 arXiv:2512.02038（2025年）。[22] Noah Shinn、Federico Cassano、Ashwin Gopinath、Karthik Narasimhan 和 Shunyu Yao。2023年。《Reflexion：采用语言强化学习的语言智能体》。载于《神经信息处理系统进展》第36卷（2023年），8634–8652页。[23] Dong Shu、Yanguang Liu、Huopu Zhang 和 Mengnan Du。2025年。《FinCall-Surprise：大规模多模态盈利意外预测基准》。arXiv预印本 arXiv:2510.03965（2025年）。[24] 标普全球市场情报。2024年。《S&P Capital IQ关键动态：数据方法论与事件分类体系》。https://www.marketplace.spglobal.com/en/datasets/key-developments-(15)。专有数据库。完整的分类体系与模式文档需要订阅。。[25] Claude团队。2025年。《推出Claude Sonnet 4.5》。https://www.anthropic.com/news/claude-sonnet-4-5。访问日期：2025-10-07。

[26] DeepSeek 团队。2025年。推出 DeepSeek-V3.2-Exp。https://api-docs.deepseek.com/news/news250929。访问日期：2025-10-07。[27] OpenAI 团队。2025年。推出 GPT-5。https://openai.com/index/introducing-gpt-5/。访问日期：2025-10-07。[28] Perplexity 团队。2025年。推出 Perplexity 深度研究。https://www.perplexity.ai/ja/hub/blog/introducing-perplexity-deep-research。[29] 通义深度研究、李柏轩等。2025年。通义深度研究技术报告。*arXiv* 预印本 *arXiv:2510.24701* (2025)。[30] 通义团队。2025年。通义深度研究：开源 AI 研究智能体的新时代。https://tongyi-agent.github.io/blog/introducing- tongyi-deep-research/。访问日期：2025-10-07。[31] Colin White、Manley Dooley 等。2025年。LiveBench：一项具有挑战性且污染受限的大语言模型基准测试。载于国际学习表征会议论文集。焦点论文。[32] xAI 团队。2025年。Grok 4。https://x.ai/news/grok-4。访问日期：2025-10-07。[33] 谢倩倩、韩伟光等。2023年。PIXIU：面向金融的大语言模型、指令数据与评估基准。载于神经信息处理系统进展。[34] 谢倩倩、韩伟光、Yanzhao Lai、Min Peng 和 Jimin Huang。2024年。FinBen：面向大语言模型的综合金融基准。载于神经信息处理系统进展。[35] 徐成、关书豪、李元、贾伟、王睿、严涵宇和张宏鑫。2024年。大语言模型的基准数据污染：综述。*arXiv* 预印本 *arXiv:2406.04244* (2024)。[36] Shunyu Yao、Jeffrey Zhao、于典、杜楠、Izhak Shafran、Karthik Narasimhan 和曹源。2023年。ReAct：让语言模型中的推理与行动协同。载于国际学习表征会议论文集。[37] 曾志远、刘家硕、陈思远、何天赐、廖雅丽、田逸潇、王金鹏、王再元、杨洋、殷凌月等。2025年。FutureX：面向大型语言模型智能体未来预测的先进实时基准。*arXiv* 预印本 *arXiv:2508.11987* (2025)。[38] 张凌昊、何世林、张朝云、康宇、李博文、谢成兴、王俊豪、王茂全、黄宇帆、傅胜宇等。2025年。SWE-bench 实时版！*arXiv* 预印本 *arXiv:2505.23419* (2025)。[39] 张扬、杨文博、Jun Wang、马强和熊杰。2025年。CAMEF：通过整合时间序列模式和显著宏观经济公告，实现因果增强的多模态事件驱动金融预测。载于第*31*届 *ACM SIGKDD* 知识发现与数据挖掘会议论文集第*2*卷。3867–3878。[40] 周书言、徐凡等。2023年。Webarena：用于构建自主智能体的真实网络环境。*arXiv* 预印本 *arXiv:2307.13854* (2023)。[41] 朱凤斌、Xiang Yao Ng、刘子扬、刘畅、曾宪伟、王超、谭天辉、姚轩、邵鹏扬、徐敏、王子轩、王静、林鑫、李俊锋、朱景贤、张扬、王文杰、Fuli Feng、洪日昌、Huanbo Luan、黄克伟和 Tat-Seng Chua。2025年。FinDeepResearch：评估深度研究智能体的严谨金融分析能力。arXiv:2510.13936 [cs.CL] https://arxiv.org/abs/2510.13936。朱卓航、陈昊东、屈强和 Vera Chung。2025年。FinCast：用于金融时间序列预测的基础模型[42]。载于第*34*届 *ACM* 信息与知识管理国际会议论文集。4539–4549。

A 重复性任务规格

A.1 宏观指标

我们监测96项宏观指标，这些指标源自14种指标类型，覆盖八个经济体，并辅以两个全球市场指数，详情见表4。筛选标准涵盖宏观分析的四大基本支柱：实际经济活动（如国内生产总值、失业率）、价格稳定（如消费者价格指数、生产者价格指数）、货币状况（如利率、股票指数）以及外部平衡（如外汇汇率、经常账户余额）。尤为重要的是，我们还纳入全球风向标——具体而言是大宗商品和波动率指数——以检验模型对跨境供应冲击和系统性风险情绪的敏感度。要准确预测这些指标，金融专家必须进行广泛的信息收集和多步推理，因此这些指标是评估深度研究能力的理想代理。

表4. 重复性任务的宏观指标。

No.	指标	描述	经济体
		全球指数（2）	
1	标普高盛商品指数	全球商品价格指数	全球
2	芝加哥期权交易所波动率指数	市场波动率指数	全球
		经济体特定指标 (94)	
3	股票指数	主要股票市场指数	全部8个
4	利率（1年期）	1年期政府债券收益率	不含新加坡
5	利率（3个月期）	3个月期国库券利率	全部8个
6	外汇汇率	兑美元汇率	不含美国
7	GDP	国内生产总值	全部8个
8	CPI	消费者价格指数	全部8个
9	PPI	生产者价格指数	全部8个
10	失业率	失业率	全部8个
11	HPI	房价指数	全部8个
12	NEER	名义有效汇率	全部8个
13	银行间利率	3个月银行间贷款利率	全部8个
14	CAB	经常账户余额	全部8个

经济体： 美国、中国、中国香港、日本、英国、德国、法国、新加坡**注：** 美国不包括外汇汇率；新加坡不包括利率（1年期）

A.2 企业财务指标

我们涵盖121项企业财务指标，并将其分为9个类别，如表5和表6所示。这一广泛的指标选取体现了专业基本面分析所采用的综合框架。通过涵盖三大主要财务报表（资产负债表、利润表、现金流量）以及六类派生比率，我们要求模型不仅能检索原始数据，还能进行算术推理，以评估流动性、偿付能力和运营效率。要准确预测这些指标，金融专家通常需要进行详细的信息收集和多步骤分析推理，这为深度研究能力提供了严格的测试。

B 非重复性任务规格

B.1 非经常性宏观事件分类体系

为对非周期性宏观经济冲击进行分类，我们采用了**“稳定核心、自适应接口”**设计理念。该层级框架由一个确保一致性的固定语义层（第一级），以及一个确保相关性的动态锚定层（第二级）组成。

表5. 企业财务指标——第1部分（资产负债表、利润表、现金流量）。

期	指标	描述	期	指标	描述
资产负债表项目（25 项指标）					
1	总资产	所有资产的总价值	14	应付账款	欠供应商的款项
2	负债总额	负债总额	15	应计费用	已发生但尚未支付的费用
3	总权益	股东权益总额	16	递延收入	已收但尚未赚取的收入
4	流动资产总额	一年内可变现的资产	17	留存收益	累计未分配收益
5	流动负债总额	一年内到期的负债	18	库存股	回购公司股份
6	长期债务	一年以上到期债务	19	少数股东权益	非控股权益
7	短期债务	一年内到期债务	20	优先股	优先股股息权益
8	短期和长期债务合计		21	普通股	基本所有权股份
9	贷款总额	金融机构持有的贷款	22	存款总额	客户存款（金融机构）
10	现金及现金等价物	流动资产及现金	23	储蓄存款	储蓄账户存款
11	应收账款	客户欠款	24	不动产、厂房及设备	实物资产价值
12	存货	待售商品	25	无形资产	专利、商誉等。
13	商誉	已支付的收购溢价			
利润表项目（22 项指标）					
26	收入	销售总额和收入	37	利息收入	利息收入
27	收入成本	销售商品直接成本	38	其他收入	营业外收入
28	毛利润	收入减去收入成本	39	非经常性项目	异常收益或损失
29	营业收入	核心业务利润	40	终止经营	已关闭分部的业绩
30	息税前利润	利息和税项前利润	41	基本每股收益	基本每股净收益
31	息税折旧摊销前利润	息税前利润加折旧与摊销	42	稀释每股收益	稀释后每股净利润
32	净利润	扣除所有费用后的最终利润	43	每股股息	每股派发的股息
33	利息支出	债务利息支出	44	营收同比增长	营收同比变化
34	研发费用	研发成本	45	净利润同比增长	年度收入变化
35	销售、一般及管理费用	销售、一般及管理成本	46	营业费用	营业总成本
36	所得税费用	企业所得税	47	税前收入	所得税费用前收入
现金流项目（15项指标）					
48	经营活动现金流	经营活动产生的净现金	56	偿还债务	用于偿还债务的现金
49	投资活动现金	投资活动产生的现金净额	57	债务发行	发行债务所得现金
50	融资活动现金	融资活动产生的现金净额	58	股票回购	用于股票回购的现金
51	自由现金流	经营现金流减资本支出	59	股票发行	发行股票所得现金
52	折旧与摊销	非现金资产减值	60	股息支付	已支付的股息现金
53	资本支出	固定资产投资	61	现金净变动	现金状况总变动
54	收购	收购公司的现金支出	62	营运资本变动	经营性资产／负债变动
55	资产剥离	出售单位所得现金			

第一级：稳定语义分类体系（不可变层）。第一级定义了一套标准化的宏观金融事件本体，旨在确保其跨时间和地区保持不变。我们将事件划分为 9 个类别（A 至 I）和 26 个子类别（表 7）。该分类体系遵循相互排斥且完全穷尽（*MECE*）原则。通过保持语义层静态，我们确保模型性能在不同时期之间具有可比性，从而为纵向评估提供一致的基准测试。

第二级：经济体特定锚定（自适应层）。不同于静态的第一级，锚定层旨在实现动态扩展。该层将通用概念映射到每个经济体特定且可证伪的市场指标。

我们的专家小组 设计这一层时，我们采用了两级灵活性，以适应金融领域不断变化的特性 al 市场：

表6. 企业财务指标——第2部分（盈利能力、流动性、杠杆、效率、偿付能力、估值）。

No.	指标	描述	No.	指标	描述
盈利能力比率（15 项指标）					
63	资产回报率（ROA）	净利润 / 总资产	71	销售回报率	营业收入 / 营收
64	股东权益回报率（ROE）	净利润 / 总权益	72	现金资产回报率	经营现金流 / 总资产
65	投入资本回报率	税后净营业利润 / 投入资本	73	现金权益回报率	经营现金流 / 总权益
66	毛利率	毛利润 / 收入	74	不良贷款率	不良贷款 / 贷款
67	营业利润率	营业收入 / 营收	75	净息差	净利息收入 / 生息资产
68	EBITDA利润率	EBITDA / 营收	76	效率比率	非利息支出 / 营收
69	净利率	净利润 / 营收	77	成本收入比	营业费用 / 营业收入
70	利润率	(营业收入 – 折旧与摊销) / 营收			
流动性比率（8 项指标）					
78	流动比率	流动资产 / 流动负债	82	营运资本	流动资产 – 流动负债
79	速动比率	（流动资产 – 存货）/ 流动负债	83	营运资本比率	营运资本 / 总资产
80	现金比率	现金及现金等价物 / 流动负债	84	防御区间比率	流动资产 / 每日运营费用
81	经营现金流比率	经营现金流 / 流动负债	85	现金转换周期	存货周转天数 + 应收账款周转天数 – 应付账款周转
杠杆比率（12 项指标）					
86	债务权益比率	总债务 / 总权益	92	长期债务资产比	长期债务 / 总资产
87	债务资产比率	总债务 / 总资产	93	短期债务总债务比	短期债务 / 总债务
88	负债资产比率	总负债 / 总资产	94	净债务	总债务 – 现金及现金等价物
89	权益比率	总权益 / 总资产	95	净债务权益比	净债务 / 总权益
90	权益乘数	总资产 / 总权益	96	净债务与 EBITDA 比率	净债务 / EBITDA
91	长期债务权益比	长期债务 / 股东权益总额	97	财务杠杆	平均资产 / 平均股东权益
效率比率（12 项指标）					
98	总资产周转率	营业收入 / 总资产	104	股东权益周转率	营业收入 / 股东权益总额
99	固定资产周转率	营业收入 / 固定资产	105	存货周转天数	365/存货周转率
100	存货周转率	销售成本 / 平均存货	106	应收账款周转天数	365 / 应收账款周转率
101	应收账款周转率	收入 / 平均应收账款	107	应付账款周转天数	365 / 应付账款周转率
102	应付账款周转率	销售成本 / 平均应付账款	108	贷款总额同比增长	贷款总额同比变化
103	营运资本周转率	收入 / 营运资本	109	存款同比增长	存款同比变化
覆盖率（6 项指标）					
110	利息覆盖率（息税前利润）	息税前利润 / 利息支出	113	利息覆盖率（营业收入）	营业收入 / 利息费用
111	利息覆盖率（息税折旧摊销前利润）	息税折旧摊销前利润 / 利息支出	114	债务偿付覆盖率	营业收入 / 债务偿付
112	利息覆盖率（净利润）	净利润 / 利息费用	115	固定费用覆盖率	(息税前利润+租赁) / (利息..+租赁)
估值与市场指标（6 项指标）					
116	每股账面价值	总权益 / 流通股数	119	每股现金流	经营现金流 / 流通股数
117	每股有形账面价值	有形权益 / 流通股数	120	企业价值	市值 + 债务 –现金
118	每股收入	收入 / 流通股数	121	市值	价格 ×已发行股份

(1) **动态校准：** 可量化触发条件（例如特定的基点阈值）需要定期重新校准。随着市场机制的变化（例如从低利率环境转向高通胀时代），可以更新这些参数，以保持其区分能力，而无需更改第一级定义。

(2) **未来任务的可扩展性：**该框架支持无缝整合新的经济体或其他事件类型。未来版本的基准测试只需定义相应的第二级依据逻辑，即可引入新的任务子类别或扩展至新兴市场，同时保持整体分类体系的完整性。

在当前版本中，我们为8个主要经济体定义了“真实值”。对于每个经济体—子类别对，我们建立了**权威来源**，严格指定官方来源（例如联邦公开市场委员会、中国人民银行、英国预算责任办公室），并设定了**可量化触发条件**及严格的量化阈值（例如加息 ≥ 25 基点、国内生产总值财政刺激 > 1%），这些阈值根据本地市场结构（例如日本的“春斗”、德国的“债务刹车”）进行了定制。详细的依据表见表8至表14。这一科学设计确保基准测试始终是“动态的”评估标准，能够随着现实世界金融格局的发展而不断演进。

表 7. 非经常性宏观事件分类。

类别	代码和详细描述	类别	代码和详细描述
A. 货币与金融状况（3 种）	**A1 货币政策转向**：中央银行政策利率加息/降息、政策立场变化、量化宽松或缩减购债决策。 **A2 金融市场流动性冲击**：债券或货币市场承压、融资紧缩、回购市场错位、做市能力受损。 **A3 宏观审慎监管变化**：宏观审慎监管的变化工具，例如贷款价值比/债务收入比限制、逆周期性资本缓冲或杠杆 caps.	**E. 实体经济活动**（4 种类型）	**E1 工业生产 / 制造业冲击**：工业生产冲击或制造业活动冲击，包括急剧收缩或激增。 **E2 零售 / 消费 / 服务冲击**：家庭消费、零售销售或服务业活动受到的冲击由收入或情绪变化驱动。 **E3住房／房地产周期冲击**：房地产市场的衰退或繁荣 mar-建筑活动或相关政-政策变化。
B. 财政政策 & 公共财政（2种）	**B1财政刺激／紧缩**：政府预算决策显著扩大或缩减支出、转移支付计划或税收负担。		**E4技术、数字经济与人工智能驱动的产业活动**：实时经济体因重大技术／人工智能发展、采用浪潮，或半导体限制而产生的影响。
	B2 主权债务压力：表明主权债务压力的事表明主权信用压力，包括评级下调、再融资压力、或违约风险。	**F. 金融稳定与信贷周期**（3 种类型）	**F1 信贷周期转变（繁荣/萧条）**：家庭或企业私人信贷的快速扩张或收缩。家庭或企业。
C. 贸易与外部部门（3 种）	**C1 贸易政策变化／制裁／关税**：引入或取消关税、配额、出口管制、制裁或反倾销税／反补贴税。 **C2 货币/外汇压力冲击**：夏普汇率变动、储备损失或资本外流，表明外汇市场压力。		**F2 银行体系压力／不良贷款冲击**：银行资产质量恶化，风-ing 不良贷款，或流动性/-偿付能力方面的担忧。 **F3 资产价格冲击（股票-债券/波动率）**：大幅调整，发生于股票或债券市场，波动率飙升，或广泛的市场重新定价事件。

续见下页.....

表 7——续自上一页

类别	代码与详细描述	类别	代码与详细描述
	C3 外部融资 / 经常账户失衡冲击：与外部融资、经常账户失衡有关的突然停止或债务展期风险。	**G. 结构性 & 监管政策**（3种）	**G1 气候 / 碳 / ESG政策**：政策目标、碳定价、排放交易或ESG披露规则相关的政策变化。
D. 大宗商品、能源与供应链链条（3种类型）	**D1 能源价格冲击**：大幅且迅速能源价格（石油、天然气、电力）的变化影响生产成本和通货膨胀。 **D2 大宗商品价格冲击**：显著关键非能源大宗商品的波动率例如金属、食品或农业投入品品。		**G2 科技/数据/隐私监管**：新或修订后的数据隐私、网络安全、跨境数据流动或数字治理。 **G3 结构性/制度性改革**：劳动力市场、养老金方面的改革，以及社会保障、法律或制度框架。
	D3 全球供应链中断：物流瓶颈、航运中断导致损害的贸易瓶颈全球供应链。	**I. 地缘政治 & 系统性冲击**（3种）	**I1 冲突/制裁冲击**：军事冲突、地缘政治升级或制裁制度，并产生宏观/部门影响。
H. 劳动力市场与家庭行业（2种）	**上半年劳动力市场冲击**：对就业、失业率、劳动力参与率或工资动态的冲击。 **下半年 家庭收入/消费压力**：家庭资产负债表承压，包括实际收入下降、债务困境或需求疲软。		**I2 自然灾害/大流行冲击**：重大自然灾害或健康危机扰乱经济活动。 **I3 全球金融传染**：全球金融危机的溢出效应、跨境银行业压力或系统性流动性冲击。

表8. 美国（US）市场的事件判定标准。

代码	主管机构/来源	可量化触发条件/定义
	A. 货币与金融状况	
A1	美联储（联邦公开市场委员会）	联邦基金目标利率区间上限变动 ≥ 25基点；或官方声明明确转变政策立场。
A2	美联储 H.4.1 报告 / 纽约联储	远期利率协议—隔夜指数掉期利差位于第 > 95 个百分位数；或逆回购工具使用量在一周。
A3	美联储理事会 / 美国联邦存款保险公司	实施新的资本规则（例如巴塞尔协议 III 最终阶段），或改变 CCAR 压力测试情境。
	B. 财政政策与公共财政	

续于下一页.....

表8——接上页

代码	主管机构/来源	可量化触发条件/定义
B1	美国国会预算办公室 / 白宫	通过立法（例如《冠状病毒援助、救济和经济安全法案》、《通胀削减法案》），对可自由支配支出产生影响 ≥ 1% 占国内生产总值的比例。
B2	美国财政部 / 信用违约互换市场	美国主权信用违约互换（5年期）利差 > 50 基点；或“非常规措施”耗尽日期将在30天内临近。
		C. 贸易与外部部门
C1	美国贸易代表办公室 / 美国商务部	实施新的第301条关税或出口管制（实体清单），影响关键行业。
C2	美国财政部 / 美联储	贸易加权美元指数（美元指数）在 3 个月内变动 ≥ 10%。
C3	BEA	经常账户赤字同比扩大 > 2% 的国内生产总值；或外国资本净流出超过历史水平 2σ。
		D. 大宗商品、能源与供应链
D1	美国能源信息署（能源信息）	西德克萨斯中质原油或亨利中心天然气现货价格在6个月内变动 ≥ 30%。
D2	美国农业部 / 美国地质调查局	主要农产品或金属商品价格偏离6个月移动平均值·
D3	纽约联邦储备银行 / 美国人口普查局	全球供应链压力指数 (GSCPI) 超过 2 个标准差。
		E. 实体经济活动
E1	美联储理事会 (G.17)	工业生产指数连续 2 个月同比下降 ≥ 3%。
E2	美国人口普查局	零售销售（不含汽车）同比下降 ≥ 2%；或密歇根大学消费者信心降至底部10%。
E3	联邦住房金融局 / 标普 CoreLogic	凯斯-希勒全国房价指数同比转负；或新屋开工量下降 ≥ 20%同比。
E4	美国经济分析局 / 国会	科技行业增加值偏离 ≥ 2σ 趋势；或重大产业政策通过（例如，《芯片与科学法案》。）
		F. 金融稳定与信贷周期
F1	美联储理事会 / 国际清算银行	私人非金融部门信贷与GDP缺口超过 +10%（繁荣）或低于 -5%（萧条）。
F2	美国联邦存款保险公司 / 美联储	受保存款机构的不良贷款率上升 ≥ 1.0%；或系统重要性金融机构银行倒闭/获救。
F3	纽约证券交易所 / 纳斯达克	标准普尔500指数或纳斯达克100指数进入技术性熊市（从峰值回撤 ≥ 20%）。
		G. 结构与监管政策
G1	美国环境保护署 / 国会	通过重大气候立法（例如《通胀削减法案》补贴）；或美国证券交易委员会出台新的气候披露规—确保任务。

续见下页.....

表8 – 接上页

代码	主管机构/来源	可量化触发条件/定义
G2	联邦贸易委员会 / 联邦通信委员会	针对大型科技公司的重大反垄断诉讼提起；或颁布新的联邦数据隐私行政命令命令。
G3	国会	对社会保障、医疗保险或移民法进行重大改革。
	H. 劳动力市场与家庭部门	
H1	BLS	失业率变化 ≥ 0.5%（萨姆规则）；或非农就业人数偏离市场共识 > 50k。
H2	BEA	实际个人可支配收入同比收缩 ≥ 2%。
	I. 地缘政治与系统性冲击	
I1	美国国务院 / 美国财政部外国资产控制办公室	美国成为武装冲突的一方；或对某个二十国集团经济体实施重大制裁经济体。
I2	美国联邦紧急事务管理局 / 美国疾病控制与预防中心	针对损失达 > 100亿美元的事件发布总统灾难声明；或全国公共卫生紧急状态声明。
I3	美国财政部 / 美联储	VIX指数 > 35 与外国净抛售美国国债相结合。

表9。中国（CN）市场的事件判定标准。

代码	主管机构/来源	可量化触发条件/定义
	A. 货币与金融状况	
A1	中国人民银行（中央银行）	7天期逆回购或1年期中期借贷便利利率变动 ≥ 5基点；或存款准备金率（RRR）下调 ≥ 25 基点。
A2	中国外汇交易中心/国家金融与发展实验室	银行间存款类机构7天期质押式回购利率较政策利率偏离 > 50 个基点，持续 5+天；或中国人民银行每周净流动性投放5000亿元人民币。
A3	中国人民银行 / 国家金融监督管理总局	调整宏观审慎评估（MPA）参数；或调整房地产行业“三道红线”指标。
	B. 财政政策与公共财政	
B1	国务院 / 财政部	发行超长期特别国债；或地方政府专项债券额度增加 > ￥1万亿。
B2	财政部 / 市场数据	10年期中国国债收益率在一个月内飙升 ≥ 20基点；或重大地方政府融资平台债券违约事件。
	C. 贸易与外部部门	
C1	商务部 / 海关	实施对战略材料（如镓/锗）的出口管制；或对主要贸易伙伴征收新关税。

下页继续.....

表9——接上页

代码	主管机构/来源	可量化触发条件/定义
C2	国家外汇管理局 / 中国人民银行	美元/人民币每日中间价与市场收盘价偏差 $>$ 500 个基点（逆周期要素使用）；或外汇储备每月下降 $>$ 500亿美元。
C3	SAFE	资本账户净流出超过国内生产总值的2%（年化）；或对跨境资本流动实施重大限制。
D. 大宗商品、能源与供应链		
D1	国家发展和改革委员会 / 国家环境局	国家发展和改革委员会调整成品油指导零售价格；或动力煤现货价格超过监管价格上限区间。
D2	大连商品交易所 / 上海期货交易所	铁矿石或螺纹钢的国内期货价格偏离6个月移动平均线 $\geq$ 20%。
D3	交通运输部 / 财新	财新制造业采购经理指数的供应商配送时间分项指数降至45.0以下。
E. 实体经济活动		
E1	国家统计局 / 财新	官方制造业采购经理指数或财新采购经理指数连续2个月收缩（<50.0）。
E2	NBS	社会消费品零售总额同比增速转负；或青年失业率（16-24岁）超过20%。
E3	国家统计局 / 住房和城乡建设部	70个大中城市新建商品住宅价格指数同比下降；或百强房企销售额下滑 $\geq$ 20% 同比。
E4	工业和信息化部 / 国家发展和改革委员会	启动重大专项工程（例如“东数西算”）；或战略性新兴产业增加值 $\geq 2\sigma$ 较趋势。
F. 金融稳定与信贷周期		
F1	PBoC	社会融资规模增速与名义国内生产总值增长率的差距 $\geq \pm$5%。
F2	国家金融监督管理总局 / 中国人民银行	接管/处置一家中型银行（例如包商银行式事件）；或商业银行不良贷款率上升 $\geq$ 0.5%。
F3	上海证券交易所 / 深圳证券交易所	沪深300指数快速下跌 $\geq$ 20%（技术性熊市）或触发交易限制。
G. 结构与监管政策		
G1	国家发展和改革委员会 / 生态环境部	发布“双碳”（1+N）政策文件；或启动新的全国碳市场交易规则。
G2	国家互联网信息办公室 / 国家市场监督管理总局	对平台经济企业处以新的反垄断处罚；或国家互联网信息办公室启动网络安全——对重要数据处理者的安全审查。
G3	中共中央	“三中全会”或“两会”宣布重大改革（例如户籍改革，共同富裕举措）。
H. 劳动力市场与家庭部门		
H1	NBS	城镇调查失业率上升 $\geq$ 0.5%；或农民工人数合约同比。

续见下页.....

表9——接上页

代码	主管机构/来源	可量化触发条件/定义
H2	国家统计局/中国人民银行	居民存款同比激增5万亿元人民币（超额储蓄）；或居民杠杆率-年龄比率下降（去杠杆）。
	I. 地缘政治与系统性冲击	
I1	外交部 / 中央军委	台湾海峡或南海紧张局势升级，引发军事演习；或对中国实体实施外国制裁。
I2	国家卫生健康委员会 / 应急管理部	启动一级公共卫生应急响应；或自然灾害影响 > 1% 的全国耕地。
I3	PBoC	股票通/债券通资金净流出超过历史第99百分位数。

表10. 日本（JP）市场的事件判定标准。

代码	主管机构/来源	可量化触发条件/定义
	A. 货币与金融状况	
A1	日本银行（日本银行）	政策利率（无担保隔夜拆借利率）变动 ≥ 10 基点；或修改收益率曲线控制（YCC）区间（例如，扩大区间）。
A2	日本银行 / 日本证券业协会	10年期日本国债收益率突破参考区间上限；或“轮番购债”（日本国债购买）操作显著增加。
A3	日本银行 / 日本金融厅	修改交易所交易基金/日本房地产投资信托购买计划指引；或宏观审慎措施针对地方银行房地产贷款的措施。
	B. 财政政策与公共财政	
B1	内阁府 / 财政部	批准支出为 > ¥10 万亿日元的“补充预算”（补正预算）；或宣布新的经济方案。
B2	MoF	预算预测中的日本国债偿债成本大幅上升；或主权评级因债务占国内生产总值比率而下调展望。
	C. 贸易与外部部门	
C1	METI	对战略技术材料（如光刻胶）实施出口限制；或从贸易伙伴“白名单”中移除。
C2	财政部 / 日本银行	财政部确认官方外汇干预（买入日元/卖出美元）；或美元/日元汇率单周内变动 ≥ 3% 。
C3	MoF	经常账户盈余大幅收窄或转为赤字（由于能源进口成本）。
	D. 大宗商品、能源与供应链	

续见下页.....

表10——接上一页

代码	主管机构/来源	可量化触发条件/定义
D1	经济产业省 / 东京电力公司	核电站重新启用获批；或公用事业公司申请电力-电价加息 > 10%。
D2	MAFF	消费者价格指数中的“食品价格指数”同比上涨 ≥ 5%；或政府对汽油的补贴——小麦价格受到触发。
D3	经济产业省 / 丰田	主要汽车制造商因零部件短缺而停产；或半导体供应链中断（例如熊本晶圆厂）。
	E. 实体经济活动	
E1	日本银行（短观）	**短观大型制造业景气判断指数** 下降 ≥ 5 点；或工业生产收缩 ≥ 2%（环比）。
E2	内阁府	国内生产总值（年化实际增长率）连续2个季度收缩（技术性衰退-信心）；或消费者信心指数下降。
E3	MLIT	土地价格公告（地价公示）显示主要大都市地区同比下跌;或东京公寓价格进入调整。
E4	METI	宣布向战略领域提供补贴（例如 Rapidus 半导体项目）;或发布人工智能战略指南。
	F. 金融稳定与信贷周期	
F1	BoJ	银行贷款同比增速显著偏离趋势；或企业破产负债激增（帝国数据银行）。
F2	日本金融厅 / 日本银行	金融厅促成地方银行（地银）合并或资本重组；或出现证券投资组合巨额亏损（例如担保贷款凭证）。
F3	东京证券交易所 / 日本交易	日经225指数或东证股价指数从峰值下跌 ≥ 20%；或波动率指数（JNIV）飙升至 > 30。
	G. 结构与监管政策	
G1	经济产业省 / 日本环境省	《绿色转型（GX）促进法》实施；或引入碳定价（GX联盟）。
G2	个人信息保护委员会 / 经济产业省	执行更严格的个人数据保护规则；或针对生成式人工智能版权的新规。
G3	内阁府	推出“新资本主义”政策举措；或大幅修订《劳动基准法》Act.
	H. 劳动力市场与家庭部门	
H1	日本劳动组合总联合会 / 厚生劳动省	**“春斗”（春季工资攻势）**商定的工资涨幅超过 3%（或日本银行目标层级）；或者有效求人倍率下降。
H2	厚生劳动省 / 总务省	实际现金收入同比收缩（工资—价格螺旋失效）；或者家庭支出（家计调查）同比下降。

下页续...

表10——接上一页

代码	主管机构/来源	可量化触发条件/定义
	I. 地缘政治与系统性冲击	
I1	外交部/国防部	钓鱼岛附近发生重大安全事件；或朝鲜导弹发射触发J-Alert系统，对市场造成影响。
I2	内阁府	发布南海海槽地震预警；或自然灾害损失估算 ＞￥1万亿。
I3	BoJ	“日本溢价”在离岸融资市场重新出现；或大规模平仓日元套利交易。

表11. 英国（UK）市场的事件判定标准。

代码	主管机构/来源	可量化触发条件/定义
	A. 货币与金融状况	
A1	英格兰银行（货币政策委员会）	银行利率变动 ≥ 25 基点；或货币政策委员会投票分歧显著变化（例如，从 6-3 变为 5-4），表明政策转向；或主动出售英国国债（量化紧缩）。
A2	英格兰银行 / 英镑隔夜指数平均利率	SONIA与银行利率利差扩大 ＞ 20 基点；或英国国债回购市场流动性出现问题（回购利率失衡）。
A3	英格兰银行（金融政策委员会）	调整逆周期资本缓冲的比率；或干预负债驱动型投资（负债驱动型投资）基金杠杆规则。
	B. 财政政策与公共财政	
B1	英国财政部 / 英国预算责任办公室	“秋季预算案”或“春季声明”宣布酌情措施 ＞ £15B；或英国预算责任办公室就财政可持续性发出警告。
B2	英国债务管理办公室 / 市场	10年期英国国债收益率在一周内飙升 ≥ 30 基点（财政恐慌）；或英国国债拍卖投标认购比降至 1.5 以下。
	C. 贸易与外部部门	
C1	英国商业部	实施英国脱欧后新的边境检查（例如英国边境目标运营模式），导致延误；或温莎框架规则的变更。
C2	英格兰银行 / 市场	英镑/美元（英镑/美元）一周内变动 ≥ 3%；或英镑贸易加权指数下跌幅度显著（通胀性贬值）。
C3	ONS	经常账户赤字超过国内生产总值的 5%（结构性脆弱性警示）。
	D. 大宗商品、能源与供应链	
D1	英国天然气和电力市场办公室	英国天然气和电力市场办公室能源价格上限调整幅度超过 ±10%（直接影响消费者价格指数）；或政府启动能源价格保障计划。
D2	英国环境、食品与乡村事务部 / 英国国家统计局	食品消费者价格指数通胀超过10%同比（生活成本危机指标）。

下页继续.....

表11——接上一页

代码	主管机构/来源	可量化触发条件/定义
D3	英国工业联合会/英国国家统计局	英国工业联合会工业趋势调查“限制产出的因素”（材料/劳动力）激增至历史平均值。
	E. 实体经济活动	
E1	ONS	月度国内生产总值（3M/3M）增速转为负值；或服务业采购经理指数跌破50.0（服务业占英国经济体的 ≈ 80%）。
E2	英国国家统计局 / GfK	GfK消费者信心指数跌破-30；或零售销售量收缩 YoY.
E3	哈利法克斯 / Nationwide	哈利法克斯或 Nationwide 房价指数同比下降；或抵押贷款批准数下降低于每月 5 万笔。
E4	DSIT	宣布人工智能安全研究所相关倡议；或对英国生命科学科技中心（例如黄金三角）进行重大投资。
	F. 金融稳定与信贷周期	
F1	BoE	抵押贷款净流量转为负值；或消费信贷增长（信用卡）激增（困境借贷）。
F2	英格兰银行 / 审慎监管局	挑战者银行承压；或企业破产数量上升（英国公司注册处数据）超过历史平均水平。
F3	伦敦证券交易所 / 富时	富时250指数（国内代理指标）下跌 ≥ 15%；或公司债相对于英国国债的利差扩大。
	G. 结构与监管政策	
G1	能源安全与净零部	调整2050净零排放时间表（例如，推迟内燃机汽车禁令）；或调整暴利税（能源利润征税），适用于能源企业。
G2	CMA	竞争与市场管理局（CMA）阻止重大科技行业并购；或出台新的《数字市场、竞争与消费者法案》执法措施。
G3	英国议会	通过租户改革或移民（签证薪资门槛）方面的重要立法。
	H. 劳动力市场与家庭部门	
H1	ONS	私营部门平均周薪（AWE）常规薪酬增长 > 6%（工资—价格螺旋式风险）；或申领人数上升。
H2	ONS	实际家庭可支配收入（RHDI）人均连续2个季度下降。
	I. 地缘政治与系统性冲击	
I1	FCDO	英国参与海外行动的军事介入；或影响贸易与合作协定（TCA）。
I2	内阁府	国家风险登记册事件启动（例如：电网停电预警）；或大流行病——层级的卫生限制措施。

下页续.....

表11——接上页

代码	主管机构/来源	可量化触发条件/定义
I3	BoE	英镑资产“闪电崩盘”；或养老基金负债驱动型投资策略中的系统性追加保证金要求。策略。

表12. 德国（DE）市场的事件判定标准。

代码	主管机构/来源	可量化触发条件/定义
	A. 货币与金融状况	
A1	欧洲中央银行（管理委员会）	欧洲央行存款便利利率变动 ≥ 25 基点；或欧洲央行宣布新的资产购买计划旨在限制利差的计划（例如 TPI）。
A2	德国联邦银行 / 欧洲中央银行	德国的 Target2 失衡显著扩大；或欧元银行间同业拆借利率—隔夜指数掉期利差扩大 > 20基点（银行间市场压力）。
A3	德国联邦金融监管局 / 德国联邦银行	为德国银行启用逆周期资本缓冲（CCyB）；或严格的贷款价值比（LTV）限制住房抵押贷款上限。
	B. 财政政策与公共财政	
B1	联邦财政部 / 德国联邦议院	德国联邦议院确认暂停“债务刹车”；或宣布——推出“特别基金” > €50B。
B2	德国金融代理机构	10年期德国国债收益率飙升 ≥ 30 基点；或德国国债与意大利国债利差扩大 > 250 基点（欧元区碎片化风险）。
	C. 贸易与外部部门	
C1	德国联邦经济和气候保护部 / 欧盟委员会	欧盟对中国电动汽车征收新关税（影响德国汽车行业）；或对主要贸易伙伴实施两用物项出口管制。
C2	欧洲中央银行 / 市场	欧元/美元汇率一周内变动 ≥ 3%；或欧元名义有效汇率（NEER）大幅下跌。
C3	德国联邦统计局 / 德国联邦银行	经常账户盈余降至国内生产总值的 2% 以下（竞争力出现结构性损失）。
	D. 大宗商品、能源与供应链	
D1	德国联邦网络局	TTF天然气价格（荷兰基准）飙升 ≥ 30%；或宣布“天然气应急计划”（天然气应急计划）第二级/第三级。
D2	德国联邦统计局	生产者价格指数（生产者价格指数）能源分项同比上涨 ≥ 20%。
D3	ifo研究所	ifo“材料短缺”调查（Materialknappheit）指标上升至超过50%的企业。
	E. 实体经济活动	
E1	ifo研究所 / 德国联邦统计局	ifo商业景气指数连续3个月下降；或工业生产（汽车行业）同比收缩 ≥ 5%。

续见下页.....

表12——续前页

代码	主管机构/来源	可量化触发条件/定义
E2	GfK	GfK消费者信心指数跌至 -20 点以下；或零售销售额（实际值）收缩 YoY.
E3	德国联邦统计局 / Bulwiengesa	住宅房地产价格指数同比收缩；或建筑许可（Baugenehmigungen）同比下降 ≥ 20%。
E4	BMWK	宣布为芯片晶圆厂（例如马格德堡英特尔工厂）或水电核心网络提供补贴（> €10B）。
	F. 金融稳定与信贷周期	
F1	德国联邦银行	面向非金融企业（NFC）的贷款同比收缩；或银行贷款调查（美国劳工统计局）显示信贷标准大幅收紧。
F2	德国联邦金融监管局	州立银行业陷入困境；或商业房地产（CRE）不良贷款率上升显著。
F3	德意志交易所	DAX 40指数下跌 ≥ 20%（熊市）；或波动率（VDAX-NEW）飙升 > 35。
	G. 结构与监管政策	
G1	BMWK	实施《供暖法》（建筑能源法）；或碳价（二氧化碳价格）上涨 > €10/吨。
G2	德国联邦卡特尔局	联邦卡特尔局阻止重大合并；或执行《数字服务法》（《数字服务法》）对平台的处罚。
G3	德国联邦议院	联合政府垮台（“交通灯联盟政府垮台”）；或通过《增长机会法》（“Wachstumschancengesetz”）。
	H. 劳动力市场与家庭部门	
H1	联邦就业局	“短时工作”（Short-time work）通知量超过每月 10 万；或失业率上升 ≥ 0.5%。
H2	德国联邦统计局	实际工资（Reallöhne）同比连续两个季度下降。
	I. 地缘政治与系统性冲击	
I1	德国联邦外交部	北溪或关键能源基础设施遭遇重大中断；或德国增加国防基金（> €100B）。
I2	BBK	因关键基础设施故障启动全国预警日；或莱茵河水位跌破“考布”临界水位（导致航运停摆）。
I3	欧洲中央银行／德国联邦银行	核心国家（德国国债）与外围国家（意大利国债）之间的利差扩大 > 250 基点（市场分化风险），从而触发欧洲中央银行干预。

表13. 法国（FR）市场的事件判定标准。

代码	主管机构/来源	可量化触发条件/定义
A. 货币与金融状况		
A1	欧洲中央银行 / 法国央行	欧洲央行存款便利利率变动 ≥ 25 基点；或法国银行行长发表讲话表明其立场偏离市场共识。
A2	法国央行 / 泛欧交易所	3个月欧元银行间同业拆借利率利差相对于隔夜指数掉期扩大 > 20基点；或法国抵押品的。
A3	法国金融稳定高级委员会 / 法国央行	法国金融稳定高级委员会调整逆周期缓冲；或强制执行严格的抵押贷款偿债收入比上限（35%）。
B. 财政政策与公共财政		
B1	经济部 / 议会	通过《财政法案》（PLF），援引第49.3条（强行通过法案，投票）；或赤字超过3%的马斯特里赫特标准，引发欧盟过度赤字程序。
B2	法国国债局 / 市场	10年期OAT收益率飙升 ≥ 30个基点；或法国国债与德国国债利差扩大 > 50 个基点（表明主权风险溢价）。
C. 贸易与外部部门		
C1	海关 / 欧盟	欧盟碳边境调节机制（CBAM）新实施措施影响法国工业；或奢侈品行业的贸易争端。
C2	法国央行 / 市场	欧元/美元波动率年化为 > 15%；或实际有效汇率（REER）升值损害出口竞争力。
C3	BdF	经常账户赤字单季扩大 > €10B；或贸易差额恶化原因是能源进口。
D. 大宗商品、能源与供应链		
D1	CRE / 法国电力公司	法国电力公司核电产量降至 280 太瓦时/年以下（历史低点）；或政府调整电价“价格盾牌”（价格盾牌）上限。
D2	法国国家统计与经济研究所	食品消费者价格指数通胀同比超过 10%（反通胀购物篮监测）。
D3	法国央行 / 法国国家统计与经济研究所	商业景气（商业景气）“供应困难”子指数显著上升——显著。
E. 实体经济活动		
E1	法国国家统计与经济研究所	制造业产出环比收缩 ≥ 1%；或商业景气指数（商业景气）跌至长期平均水平 100 以下。
E2	法国国家统计与经济研究所	消费者信心（家庭信心）跌至 85 以下；或家庭商品消-费同比收缩。
E3	法国公证人协会/法国国家统计与经济研究所	二手房价格指数同比下跌；或住房开工（住房开工）减少 ≥ 15% 同比。

下页继续.....

表13——接上页

代码	主管机构/来源	可量化触发条件/定义
E4	经济部	“法国2030”投资计划拨款加速；或针对……的重大补贴 法国北部的“超级工厂”（电池）。
	F. 金融稳定与信贷周期	
F1	BdF	对非金融企业的信贷增速放缓至 $< 2\%$ 同比；或“国家担保贷款”的（PGE）违约增加。
F2	审慎监管与处置管理局 / 法国央行	大型银行保险集团的偿付能力比率下降；或人寿保险（Assur-）增加 ance Vie）提款。
F3	巴黎泛欧交易所	法国CAC 40指数下跌 $\geq 20\%$（熊市）；或奢侈品行业子索引（LVMH, 爱马仕、开云集团）回调 $\geq 15\%$。
	G. 结构与监管政策	
G1	法国生态部	新的“能源性能诊断”（DPE）规定禁止出租G级住房；或 提高碳税。
G2	CNIL	法国国家信息与自由委员会因违反《通用数据保护条例》对大型科技公司处以罚款；或出台新的《网红法》规定的执行。
G3	议会 / 总统	通过养老金改革（Réforme des retraites），提高退休年龄；或失业-就业保险改革法令。
	H. 劳动力市场与家庭部门	
H1	法国研究、评估与统计局 / 工会	总罢工（总罢工）导致交通运输/炼油厂中断 > 3 天；或私营部门 工资单（受薪就业）收缩。
H2	法国国家统计与经济研究所	每单位合同购买力（Pouvoir d’achat）同比变化；或最低工资（SMIC）自动通胀调整 $> 2\%$。
	I. 地缘政治与系统性冲击	
I1	法国外交部	法国军事干预（例如萨赫勒地区、东欧）；或恐怖袭击警戒级别 提升至“恐怖袭击紧急状态”。
I2	法国内政部	社会动荡（例如“黄背心运动”或2023年骚乱）造成全国范围内的损失，达 $> €1B$；或干旱限制措施影响农业。
I3	法国央行 / 欧洲中央银行	法国国债与德国国债利差扩大 > 80 基点，触发欧洲央行传导保护工具启动。

表14. 新加坡（SG）市场的事件判定标准。

代码	主管机构/来源	可量化触发条件/定义
	A. 货币与金融状况	

下页续...

表14——接续上一页

代码	主管机构/来源	可量化触发条件/定义
A1	MAS	新加坡金融管理局收紧新元名义有效汇率政策区间（例如加大斜率、扩大宽度，或重新-向上居中）；或非例行货币政策声明。
A2	新加坡金融管理局 / 新加坡银行公会	新加坡隔夜利率平均值一周内飙升 > 50 基点；或新加坡元流动性便利使用量激增。
A3	MAS	收紧总偿债率（TDSR）门槛（例如，从 60% 降至 55%）；或下调贷款价值比上限。
	B. 财政政策与公共财政	
B1	MOF	“预算声明”宣布推出重大“生活成本”援助计划（> 50亿新元）；或发放社区发展理事会／消费税券。
B2	财政部／收益率曲线	**新加坡政府证券**（新加坡政府证券）收益率曲线大幅倒挂；或净投资回报贡献（净投资回报贡献）预计将下降。
	C. 贸易与外部部门	
C1	新加坡企业发展局	非石油国内出口（非石油国内出口）收缩 ≥ 10% 同比（科技周期下行）；或影响新加坡港的重大供应链中断。
C2	MAS	新元名义有效汇率触及政策区间顶部或底部（隐性干预区间）；或 新加坡银行同业拆借利率/新加坡隔夜利率平均值与美元利率之间的利差大幅扩大。
C3	MTI	资本外流导致国际收支赤字；或官方大幅下降外汇储备。
	D. 大宗商品、能源与供应链	
D1	新加坡能源集团 / 能源市场管理局	电价环比调整 ≥ 10%；或能源市场管理局（EMA）启动燃料储备。
D2	新加坡贸易与工业部 / 新加坡统计局	新加坡金融管理局核心通胀率超过 5% 同比（触发生活成本政策应对）。
D3	新加坡港务集团 / 海事及港务管理局	新加坡港口集装箱吞吐量同比收缩；或船用燃油价格飙升扰乱航运枢纽地位。
	E. 实体经济活动	
E1	EDB	工业生产（电子集群）同比收缩 ≥ 10%；或采购经理指数（新加坡采购与材料管理学院）跌至 50.0 以下。
E2	DOS	零售销售额（不包括机动车）同比收缩；或餐饮服务索引显著下降。
E3	新加坡市区重建局 / 新加坡建屋发展局	市区重建局私人住宅价格指数环比上涨 > 2%（触发降温措施）；或建屋发展局转售价格指数大幅上涨。
E4	新加坡经济发展局 / 新加坡金融管理局	宣布在半导体/生物医学领域进行重大外国直接投资（例如，GlobalFoundries 扩大投资—张）；或批准新的金融科技监管沙盒。

下页继续.....

表14——续自上一页

代码	主管机构/来源	可量化触发条件/定义
		F. 金融稳定与信贷周期
F1	MAS	信贷与GDP缺口超过10%；或国内银行业务部门（国内银行业务部门）的非银行贷款增长转为负值。
F2	MAS	本地银行（星展银行/华侨银行/大华银行）的不良贷款率上升 ≥ 0.5%；或对区域性违约（例如中国房地产）的风险敞口显著增加。
F3	SGX	海峡时报指数（STI）下跌 ≥ 15%；或每日证券市场成交额降至低于10年平均水平。
		G. 结构与监管政策
G1	MSE / 新加坡国家气候变化秘书处	提高碳税（例如，达到每吨 25 新元）；或实施强制性气候报告撰写，适用于上市公司。
G2	MAS	发布更新后的科技风险管理（TRM）准则；或对数字支-付代币服务提供商实施重大执法。
G3	议会	调整中央公积金（CPF）缴款率或退休年龄；或收紧就业准证（EP）标准。
		H. 劳动力市场与家庭部门
H1	MOM	裁员人数超过每季度 5,000 人；或失业率上升 ≥ 0.5%。
H2	人力部（MOM）／全国工资理事会（NWC）	实际收入中位数增长转为负值；或全国工资理事会（NWC）建议冻结工资。
		I. 地缘政治与系统性冲击
I1	MFA	马六甲海峡发生重大地缘政治紧张局势，影响贸易流动；或制裁产生影响，进而影响新加坡作为转运枢纽的作用。
I2	卫生部 / 国家环境局	疾病暴发应对系统级别升至橙色/红色（大流行病）；或“烟霾”污染物标准指数超过 300（危险的）扰乱经济。
I3	MAS	尽管政策收紧，新元名义有效汇率仍跌破政策区间中点（信心丧失）；或新元融资成本飙升。

B.2 非经常性企业事件类型

为确保我们的分类体系以成熟的金融行业标准为依据，而非主观定义，我们采用了标普全球市场情报 [24]的关键进展框架。该数据库因追踪重要企业事件而广受专业分析师认可，并被视为事件分类的“黄金标准”。

标普模式原本包含 100 多种细分事件类型，我们精选出其中 70 种事件类型（表15），形成精简子集，专为大语言模型预测任务优化。我们的筛选过程遵循两项原则：**(1) 重要性和信号价值：**剔除影响较小的行政事件（例如例行会议演示、网站细微更新），这类事件通常不会引发显著的市场反应，也不需要复杂推理。**(2) 语义区分度：**合并高度重叠的类别，构建适用于分类任务的相互独立、完全穷尽框架。通过这一精选过程，基准测试能够评估模型预测“影响市场”事件的能力，而不会被行政杂音分散注意力。

表15. 非重复性任务的企业事件类型（8个类别，共70种类型）。

No.	事件类型	描述	No.	事件类型	描述
企业结构（8）			**财务——业绩与指引（9）**		
1	高管变动——首席执行官	首席执行官职位变动	9	盈利公告	定期盈利公告
2	高管变更——首席财务官	首席财务官职位变更	10	经营业绩公告	定期经营业绩果
3	高管/董事会——其他	其他领导层变更	11	销售公告	非盈利业绩
4	名称变更	公司名称变更	12	下调指引	预期业绩降低
5	地址变更	总部迁址	13	上调指引	预期业绩提高
6	法律结构变更	法律结构变更	14	指引 - 新建/已确认的公司	新的或已确认的指引ance
7	业务重组	单位合并/拆分-ing	15	指引 - 异常事件	异常事件影响
8	财年结束日期变更	报告期变更	16	业绩重述	业绩更正
			17	减值／核销	资产价值减损
财务－股息 (6)			**财务－资本与融资 (10)**		
18	股息确认	股息重申	24	回购公告	回购获授权
19	股息增加	更高股息	25	回购完成	回购结束
20	股息减少	更低股息	26	回购取消	回购取消
21	股息取消	股息暂停	27	回购计划变更	回购条款修改
22	股息启动	首次股息	28	股票拆分与股息	股份增加
23	特别股息	非经常性分配-tion	29	债务融资相关	募集的债务工具
			30	私募配售	机构销售
			31	IPOs	首次公开募股
			32	后续股票发行	发行-ings 首次公开募股后股权
			33	固定收益发行	债券/票据发行
战略行动 (10)			**并购交易 (5)**		
34	业务扩张	内生增长	43	并购传闻与讨论	协议前谈判内容
35	终止经营	生产线／设施关闭	44	并购公告	交易宣布
36	战略联盟	合作协议	45	并购交割	交易完成
37	客户公告	客户关系变更	46	并购取消	交易取消
38	产品公告	产品介绍/变更	47	分拆/分立	分拆交易
39	战略选项	评估选项	**市场事件 (5)**		
40	寻求出售/剥离	出售机会	48	退市	交易所移除
41	寻求收购	收购权益	49	交易所变更	交易所转移

续见下页.....

表15——接上一页

No.	事件类型	描述	No.	事件类型	描述
42	寻求融资	股权／风险投资融资	50	股票代码变更	股票代码变更
			51	指数成分股剔除	索引移除
			52	指数成分股新增	指数新增
监管合规（9）			**困境指标（7）**		
53	监管问询	美国证券交易委员会/监管问询	62	债务违约	义务违约
54	监管法规	新规发布	63	破产申请	重组/清算 filing
55	监管—合规	批准/许可证	64	破产—结案	申请被驳回
56	监管—执法	罚款/处罚	65	破产—脱离	脱离破产
57	诉讼与法律问题	法律行动	66	破产—其他	其他破产 mo-情况
58	SEC文件延迟提交	文件逾期提交	67	破产——资产出售	清算／资产出售
59	财报延迟发布	推迟发布	68	暂停／恢复运营	异常事件停牌
60	审计师变更	审计师变更	**运营变更 (2)**		
61	审计师持续经营	生存能力问题	69	劳工公告	劳动力新闻
			70	章程／规则变更	治理规则变更

C 非周期性任务生成流程

非周期性任务生成代理采用三阶段的基于大语言模型的流程。在信号检测阶段，搜索增强型大语言模型扫描实时新闻流，以识别结果为二元且存在真实不确定性的候选事件，优先关注政策公告、监管裁决和人事变动，同时排除常规定期发布。在相关性评估阶段，子代理通过分析每个候选事件的具体未知因素（例如有争议的阈值、有争议的措辞）、不同来源之间相互矛盾的信号，以及潜在延迟等风险因素来进行调查，并根据来源分歧分配不确定性评分。在问题生成阶段，代理综合研究报告，围绕边际内容而非单纯的事件是否发生来拟定问题，并通过自动验证确保问题具有二元可回答性和真实的不确定性。

D 非重复性任务的真实值获取

非重复性任务的真实值结合了自动化证据汇总与专家验证。搜索增强型大语言模型代理执行三阶段流程：(1) 搜索官方来源和主要财经新闻；(2) 核实数值阈值和事件定义；(3) 检查是否存在否认或矛盾证据。大语言模型裁定器综合调查结果，作出二元判定：只有得到权威来源确认的事件才归类为 YES；信号含糊或缺乏证据则归类为 NO。为确保达到金标准质量，**所有（100%）**自动生成的提案都必须经过专家审查。专家会对照官方来源核验证据，作出最终裁定，标注者间一致性达到 95%。

表16. FINDEEPFORECASTBENCH 的基本统计数据。W =每周，M = 每月，Q = 每季度。

	递归		非递归	
规格	**宏观**	**企业**	**宏观**	**企业**
# 企业 / 经济体	8	1,314	8	1,314
# 指标 / 事件类型	96	121	208	70
预测期	W / M / Q	Q	W / M / Q	W / M / Q
平均每周任务数	30	72	16	82
任务总数	296	723	128	247

E 统计与分析

除了构建流程和质量控制机制外，我们还提供详细统计数据，以说明该基准测试的规模和分布。表16列出了双轨分类体系下的基本统计数据。

全面任务覆盖。 该基准测试涵盖来自北美、欧洲和亚太地区主要金融中心的1,314家公司，这些公司选自8个经济体和9个指数。重复性任务通过数值预测涵盖96项宏观指标和121项企业指标；非重复性任务则通过二元预测处理208项特定经济体宏观事件说明和70种企业事件类型。

生成状态。 自2025年10月27日起，该基准测试已在10个每周批次中生成总计1,394项任务，平均每周新增139项任务。如表16所示，重复性任务共计1,019项，平均每周生成30项宏观任务和72项企业任务。非重复性任务共计375项，平均每周生成16项宏观任务和82项企业任务。所有1,394项任务均已获得真实值，确保模型评估所需的数据完整可用。

多频率预测期。 三种预测期可实现差异化的时间评估：周预测期评估短期响应能力，月预测期评估中期趋势整合，季度预测期则检验长周期预测能力。重复性企业任务仅关注与财报周期相匹配的季度预测；重复性宏观任务和所有非重复性任务则支持周、月和季度预测期。

跨市场任务分布 表17展示了涵盖8个经济体和四类任务的详细任务分布。

表17. 8个经济体的任务分布。

任务类别	**US**	**CN**	**HK**	**JP**	**UK**	**DE**	**FR**	**SG**	**总计**
宏观重复任务	30	40	39	39	40	38	40	30	296
企业重复任务	413	68	37	141	25	1	24	14	723
非周期性宏观任务	16	16	16	16	16	16	16	16	128
非周期性企业任务	57	25	26	30	30	29	30	20	247
总计	**516**	**149**	**118**	**226**	**111**	**84**	**110**	**80**	**1,394**

任务分布揭示了不同任务类别之间的鲜明模式。通过系统性覆盖预定指标发布，各经济体的周期性宏观任务分布相对均匀。相比之下，周期性企业任务的跨市场差异最大，这反映了在我们的评估期内，各市场的指数成分数量和财报发布时间存在差异。非周期性宏观任务经过设计，在每个经济体中均恰好包含 16 项任务，从而实现均匀分布；而企业任务则因新闻报道密度不同而呈现中等程度的差异。按地理区域划分，亚太地区占任务总数的 41%，美洲占 37%，欧洲占 22%。这种均衡的地理构成对于确保模型的

性能能够在多样化的市场微观结构、监管环境和文化背景下接受评估，而非针对美国等单一主导经济体过度拟合。

F 模型设置

表18列出了我们在基准测试中评估的方法配置。所有未特别说明的设置均保持默认值。

G 错误案例研究

表19按情境（非递归与递归）和层级（企业与宏观）细分了**Perplexity Sonar 深度研究** 最常见的错误模式。对于每个情境—层级分组，我们报告任务总数、预测错误的数量以及相应的错误率。为说明典型的失败情况，我们还提供了每组中最常见错误子类型的一个代表性示例。

H 模型预测示例

图5展示了一个企业非重复性任务示例的提示配置。后续各图分别呈现了受评模型针对该任务生成的推理和答案：OpenAI o3-deep-research（图6）、Perplexity Sonar 深度研究（图7）、通义深度研究（图8）、OpenAI GPT-5（T+S）（图9）以及 Claude-Sonnete-4.5 (T)（图10）。

表18. 基准别名、API标识符与API设置之间的对应关系

基准别名	API标识符	API设置
大语言模型（思考）		
Gemini 2.5 Pro（T）	gemini-2.5-pro-preview-05-06	推理_预算=-1
Deepseek-v3.2（T）	deepseek-v3.2-exp	推理.enabled=True
Claude-Sonnet-4.5（T）	claude-sonnet-4-5-20250929	思考.type=已启用，思考.budget_令牌秒=10000
Grok 4（T）	grok-4-0709	所有默认设置
OpenAI GPT-5（T）	gpt-5-2025-08-07	推理.effort=high
大语言模型（思考 + 搜索）		
Gemini 2.5 Pro（T+S）	gemini-2.5-pro-preview-05-06	思考_预算=-1,工具=[谷歌_搜索]
Deepseek-v3.2（T+S）	Deepseek-v3.2-Exp	推理.enabled=True,插件=[exa(max_results=8)]
Claude-Sonnet-4.5（T+S）	claude-sonnet-4-5-20250929	thinking.type=已启用，thinking.budget_令牌s=10000,工具=[网页_搜索_20250305]
Grok 4（T+S）	grok-4-0709	搜索_参数.mode=开启
OpenAI GPT-5（T+S）	gpt-5-2025-08-07	推理.effort=中等，工具=[网页_搜索]
深度研究		
Perplexity Sonar 深度研究	sonar-deep-research	推理.effort=high
通义深度研究	tongyi-deepresearch-30b-a3b	温度=0.6，top_p=0.95，存在惩罚_penalty=1.1
OpenAI o3-deep-research	o3-deep-research-2025-06-26	工具=[网页_搜索_预览，代码_解释器]

表19. **Perplexity Sonar 深度研究**的错误分析。在代表性错误案例中，我们展示了最常见的错误子类型（类型/指标/指示器）及其占该组错误的比例（括号内为百分比），并给出一个示例问题（Q）、标准答案和模型预测。

情境	层级	# 任务	# 错误	% 错误	代表性错误案例
非周期性企业任务	企业	247	50	20.24	**类型：** 企业事件（32.0%） **问题：** 液化空气集团是否有可能宣布一项新的到2025年12月19日，北美的生产设施? **黄金：** 是 **预测：** 否
非周期性宏观任务	宏观	128	34	26.56	**类型：** – **问：** 英国财政大臣蕾切尔·里夫斯是否会在秋季预算声明中，于2025年11月26日宣布至少一项具体的新税收增加措施（不包括11月17日前已宣布或泄露的措施），而且此前未在政府官方通报中明确提及? **黄金：** 否 **预测：** 是
递归	企业	724	537	74.17	**指标：** 经营活动现金流（9.9%） **问题：** 使用可获得的财务数据，估算Booking Holdings Inc (BKNG) 2025年第三季度的经营活动现金流。 **黄金：** 1435.0 **预测：** 2900.0
递归	宏观	277	213	76.90	**指标：** 3个月期政府债券收益率（34.3%） **问：** 利用现有经济数据，估算中国的中国国库券10月31日 ，2025年。 **黄金：** 1.2885 **预测：** 1.35

提示输入

液化空气集团在2025年12月19日之前宣布将在北美新建生产设施的可能性有多大?

事件详情

- 公司：液化空气集团
- 股票代码：AI
- 索引：CAC 40指数
- 事件类型：业务扩张

要求：
请将您的预测以**是**或**否**的形式提供。请根据
以下内容完成评估：

- 近期新闻和公司公告
- 该公司类似事件的历史模式
- 当前市场和行业背景
- 公司的财务状况和战略重点
- 其他相关信息

提交格式
请在此 Markdown 表格中填写您的预测：
| 字段 | 值 |------|------| 预测（是/否） | [您的预测] |

补充指南

(1) **独立性：** 请基于公司申报文件、新闻来源以及您自己的分析开展研究。不要仅仅引用预测市场或第三方预测结果。(2) **时效性：** 优先采用最新的可用数据和公告（请注明数据的截止日期）。(3) **输出格式：** 您的回复必须仅包含按照提交格式要求填写完整的 Markdown 表格。不得在表格之外添加任何文字。

答案：是

图5. 企业非经常性任务的提示输入。该任务的真实答案是“是”。

OpenAI o3-deep-research

...
考虑谅解备忘录与合资企业的潜力：鉴于今天是12月14日，这个时间相当近！我应该查查是否有任何近期声明或新闻提到伊藤忠商事对这类合作的兴趣。有消息称，ACWA Power 与伊藤忠商事签署了谅解备忘录，以推动全球清洁能源发展。谅解备忘录是迈向更大合作的一步，或许会促成合资企业（JV）到2025年，不过这并非板上钉钉。
分析绿色氨项目：搜索结果显示，L&T Energy 正在与伊藤忠商事合作开展一个绿色氨项目，该项目计划于2025年8月13日启动。不过，我不确定这属于正式的合资企业，还是仅仅一种合作关系。
澄清联合开发协议：我看到用户正在讨论一份“联合开发协议”，而这并不完全算是正式的合资企业，因为没有涉及独立的实体。
评估合资企业：过去，他们曾为合资企业成立独立公司，例如与 NF 合作开展的储能项目。他们还与中信集团建立过合作关系，这可能并非直接聚焦于可再生能源。2024年，他们已与东京世纪成立合资企业，开展太阳能开发。目前看来，他们不太可能很快成立新的合资企业，尤其是在接下来的几天内。
考虑合资企业和合作：考虑到时间安排，如果预计会有重大消息，那么截至目前，它很可能已经宣布或传出风声。
评估成立合资企业的可能性：我在判断这件事是否会在指定日期当天或之前发生，也就是下周。这个问题可能预期答案为“否”，因为目前没有已知迹象表明此事会发生。历史趋势表明，这类形成过程每2至3年发生一次。鉴于2025年目前还没有发生，因此在此之前似乎不太可能会有一次第19次。
...

答案：否

图6。OpenAI o3-deep-research 对企业非经常性任务的模型推理。仅展示重要的思考步骤。模型的答案是否定的。

Perplexity Sonar Deep Research

用户要求我预测伊藤忠商事株式会社是否会在2025年12月19日前在可再生能源行业成立合资企业。不过，当前日期是2025年12月14日，距离截止日期只剩5天。让我查看搜索结果，找出近期有关伊藤忠商事株式会社成立可再生能源行业合资企业的公告：1. **来源 [1] - 2025年7月28日：** 伊藤忠商事宣布与九电工株式会社合作，在佐贺大内太阳能发电站开展FIP转换项目——这是合作关系，不一定是合资企业。2. **来源 [2] - 2024年6月11日：** 伊藤忠商事投资Grandview风电场，并与东京世纪株式会社在美国特拉华州成立合资企业，开展太阳能项目——这是2024年已经成立的合资企业。3. **来源 [3]** - 伊藤忠商事清洁技术业务概览，提及多个可再生能源项目。4. **来源 [4] - 2025年12月11日：** 伊藤忠商事宣布投资科罗拉多州的Black Hollow Sun发电厂。5. **来源 [5] - 2024年5月24日：** 伊藤忠商事宣布与谷歌开展虚拟购电协议服务。6. **来源 [6] - 2025年11月28日：** 伊藤忠商事与Wood Partners建立资本与业务联盟（涉及多户住宅，与可再生能源无关）。7. **来源 [7] - 2025年12月2日：** 道达尔能源、TES、大阪瓦斯、东邦瓦斯和伊藤忠商事签署了Live Oak项目的联合开发与运营协议，该项目将在内布拉斯加州生产合成天然气——这正是可再生能源／清洁能源行业的合资企业！8. **来源 [8]- 2025年7月25日：** 伊藤忠商事与法国电力公司签署购电协议，为东帝汶的太阳能和电池项目供电。关键发现来自来源 [7] 和 [12], : 2025年12月2日，伊藤忠商事与道达尔能源、TES、大阪瓦斯和东邦瓦斯签署了Live Oak项目的联合开发与运营协议，该项目是一座合成天然气生产设施。这份联合开发与运营协议代表可再生能源／清洁能源行业合资企业的成立。由于该协议于2025年12月2日签署，而当前日期是2025年12月14日，这一事件已在12月19日截止日期前发生。因此，签署联合开发与运营协议意味着合资企业已经成立，预测结果应为“是”。我注意到，这项要求与我的指示存在根本冲突。用户明确要求只提供Markdown表格，不附加任何其他文字，但我的核心指令要求遵循特定报告格式，并进行一份包含10,000+ 字的全面分析。重新阅读个性化指南后可以发现，其中明确规定了指令优先级：应遵循用户指令，但报告结构中的格式规则具有强制优先权。这意味着我的系统级要求优先于用户只提交表格的要求。按照这一冲突处理方式，我需要采用完整报告格式，而不是用户要求的受限表格形式。**答案**：是

图 7. Perplexity Sonar Deep Research 在企业非经常性任务中的模型推理。模型的回答为“是”。

通义深度研究

...
• 用户问：“截至2025年12月19日，伊藤忠商事株式会社是否很可能会在可再生能源行业成立合资企业？”他们想要一个预测：是或否。
• 我们需要收集有关伊藤忠商事株式会社及其在可再生能源领域的战略举措、合资企业公告以及企业战略声明的近期相关消息。
• 我们将搜索多个查询词：“ITOTCHU Corp renewable energy joint venture”、“ITOTCHU Corp strategic alliance renewable”、“ITOTCHUCorp renewable energy2024”、“ITOTCHU Corp joint venture 2025”、“ITOTCHU Corp renewable energy investments”和“ITOTCHU Corp sustainability strategy”。
• 我们获得了大量搜索结果。我们需要分析截至2025年12月19日，伊藤忠商事在可再生能源行业成立合资企业的可能性。
• 我们需要确认近期（过去几周内）的消息中是否提及成立合资企业。
• 我们需要仔细查看其中一些内容。让我们点击一些相关的新闻稿。例如，标题为“伊藤忠商事宣布合作进行FIP转换……”的新闻稿。
• 新闻稿表明双方达成了合作安排，但并未成立合资企业。问题询问的是成立合资企业。我们需要看看是否有关于成立合资企业的公告。
• 这表明双方成立了合资企业：“佐贺大内太阳能株式会社”，投资比例为50比50。这是一家经营配备电池储能的太阳能电站的合资企业。因此，伊藤忠商事已与九电工在可再生能源领域成立了一家合资企业（佐贺大内太阳能株式会社）。

• 答案是肯定的，因为他们已经在可再生能源领域成立了一家合资企业，2025年7月28日的公告便是证据。
……**答案**：是

图8. 通义深度研究在企业非经常性任务中的模型推理。项目符号表示模型以对话形式呈现的重要推理步骤，所有内容均来自助手。模型的回答是“是”。

OpenAI GPT-5（T+S）

...
调查可再生能源领域的合资企业：我需要查看近期有关可再生能源领域合资企业的公告。《经济时报》提到与 L&T GreenTech 在古吉拉特邦开展绿色氨项目的合作关系，但这意味着双方达成的可能是谅解备忘录（MoU），而非正式的合资企业。我们要找的是作为“战略联盟”一部分创立合资企业的具体信息。除非已有传闻或备案，否则这件事在五天内发生的可能性似乎很低。我应该进一步搜索，以弄清楚情况。
评估合资企业： 我需要确认伊藤忠商事是否已宣布成立或拟议成立任何合资企业（JV）以开展可再生能源业务。例如，他们可能与KDDI或软银等公司建立了合作关系。此外，他们还于2025年12月2日与道达尔能源签署了一份联合开发与运营协议（JDOA），但我想知道这是否算是真正的合资企业。由于要求是在12月19日前成立合资企业，我需要评估考虑到此前的联合开发与运营协议，其发生的可能性。
评估合资企业的成立：我看到，2024年6月已确立了与东京世纪成立合资企业的证据，涉及太阳能电站项目。然而，关于截至12月19日、2025年成立另一家合资企业的问题，更属于前瞻性预测，因为当前日期是2025年12月14日。虽然他们已经成立了一家合资企业，但这并不保证会在最近五天内再成立一家。他们近期的活动显示有投资行为，但尚未成立新的合资企业实体。
评估合资企业的时间线：我发现，2024年6月，伊藤忠商事与东京世纪明确成立了一家合资企业用于太阳能项目。用户的问题是，伊藤忠商事株式会社是否会在可再生能源领域成立合资企业截至2025年12月19日。现在是2025年12月14日，我认为“截至”包含接下来的五天。不过，通常在预测中，我也会考虑过去发生的事件，但这里我们关注的是未来成立合资企业的机会。这个时间范围有些棘手！
分析合资企业的定义：联合开发与运营协议为项目分配股权，但不一定会成立新的合资合资企业。Live Oak可能是一个项目公司，参与方持有其股票，运作方式类似于合资企业特殊目的载体。不过，他们并未将其标记为“合资企业”，而只是称其为对一个项目的持股，并与合作伙伴共同参与，这可以被视为以类似方式运作。关键问题仍然是：对该项目的股权参与是否构成新的合资企业？与伊藤忠商事等公司共同参与的Live Oak项目是否构成新的合资企业？
合资企业预测最终确定：新闻稿表明，该项目正在开发中，并且他们已经签署了联合开发与运营协议（JDOA），但尚未达到最终投资决策（FID）阶段阶段。也就是说，我不能将此视为成立合资企业。因此，截至12月19日成立新合资企业的可能性很低，所以我会选择“否”。我会将最终结果整理为一个 Markdown 表格，只包含这两行。
...

答案：否

图9. OpenAI GPT-5（T+S）对企业非经常性任务的推理。这里只呈现重要的思考步骤。模型的答案是否。

Claude-Sonnet-4.5 (T)

我需要搜索有关伊藤忠商事株式会社近期信息，了解其在可再生能源行业的活动，尤其关注截至2025年12月19日可能正在形成的任何合资企业或战略联盟。由于今天是2025年12月14日，时间范围非常短（仅5天）。这意味着我需要查明：

1. 最近是否有关于可再生能源合资企业的公告。2. 是否有可能在2025年12月19日前完成的交易。3. 伊藤忠商事近期在可再生能源领域的战略方向。让我搜索有关伊藤忠商事和可再生能源合资企业的近期新闻。

答案：否

图10。Claude-Sonnet-4.5 (T) 对企业非经常性任务的模型推理。模型的答案是否。