arXiv:2602.10886v1 [cs.CL] 11 Feb 2026

CLEF-2026 FinMMEval 实验室：金融人工智能系统的多语言多模态评测

谢卓涵[1[0009−0008−2650−2857]]、Rania Elbadry[1[0009−0002−9337−9680]]、张帆[2[0009−0008−7515−3542]]、Georgi Georgiev[3[0009−0006−4418−7622]]、彭雪晴[4[0009−0004−1484−0622]]、钱灵飞[4[0009−0007−8342−1837]]、黄继敏[4[0000−0002−3501−3907]]、Dimitar Dimitrov[3[0000−0003−1308−180X]]、Vanshikaa Jani[5[0009−0001−1292−7386]]、戴宇扬[6[0009−0005−4077−0493]]、耿佳慧[1[0000−0002−4205−8230]]、YuxiaWang[6[0000−0002−4474−1826]]、Ivan Koychev[3[0000−0003−3919−030X]]、Veselin Stoyanov[1[0009−0003−1250−4820]]和Preslav Nakov[1[0000−0002−3600−1510]]

[1] 穆罕默德·本·扎耶德人工智能大学，阿布扎比，阿联酋[2] 东京大学，东京，日本 [3] 索非亚大学“圣克莱门特·奥赫里德斯基”，索非亚，保加利亚[4] The Fin AI，美国[5] 亚利桑那大学，图森，美国 [6] INSAIT，索非亚大学“圣克莱门特·奥赫里德斯基”，索非亚，保加利亚 {zhuohan.xie, preslav.nakov}@mbzuai.ac.ae

摘要。 我们介绍了 CLEF 2026 FinMMEval Lab 的设置和任务。该实验室推出了首个面向金融大语言模型（LLM）的多语言、多模态评估框架。近年来，金融自然语言处理取得进展，使市场报告、监管文件和投资者沟通的自动化分析成为可能，但现有基准测试仍大多局限于单一语言和纯文本，并且只涵盖范围狭窄的子任务。FinMMEval 2026 通过提供三个相互关联的任务来弥补这一不足，涵盖金融理解、推理和决策：金融考试问答、多语言金融问答（PolyFiQA）和金融决策。这些任务共同构成一套全面的评测方案，用于衡量模型在多种语言和模态中进行推理、泛化和行动的能力。本实验室旨在推动稳健、透明且具有全球包容性的金融人工智能系统的发展，并公开发布数据集和评测资源，以支持可复现研究。

关键词： 金融自然语言处理 · 多语言评估 · 多模态推理 · 金融问答 · 金融决策 · 金融人工智能 · 跨语言理解

1 引言

大语言模型（LLMs）的快速发展推动了金融自然语言处理（金融自然语言处理，FinNLP）的显著进步，使得对市场报告、监管文件和投资者沟通内容进行自动化分析成为可能 [30, 20, 32]。尽管取得了这些进展，现有的大多数基准测试仍然是单语言、纯文本的，并且狭隘地聚焦于情感分类或事实问答等子任务 [12, 1]。然而，现代金融沟通日益呈现全球化和多模态特征，涵盖多语言新闻、监管申报文件和实时市场数据。这一演变要求建立全面的评估框架，以评估模型跨语言、跨模态进行推理、泛化和行动的能力。为弥合这一差距，CLEF 2026 的 FinMMEval Lab 通过推出首个面向金融大语言模型的多语言和多模态共享任务，拓展了评估的边界。在 Regulations Challenge [25], Fin-DBQA [17],和 Earnings2Insights [23], 等早期项目奠定的基础上，FinMMEval 将金融推理、多语言理解和决策整合到一个统一的评估套件中，旨在促进稳健、透明且具备全球能力的金融人工智能。该项目涵盖五种语言的三个相互关联的任务，概述见表 1。每项任务都针对金融推理的一个不同层面，从概念理解到现实世界中的决策，同时应对多语言和多模态金融自然语言处理中的关键挑战：

– 任务 *1*：金融考试问答： 作为 FinMMEval 的基础组成部分，重点关注概念理解和领域推理（参见第 3 节）。该任务评估模型能否用多种语言应对专业考试类问题，为后续任务奠定知识基础。– 任务 *2*：多语言金融问答（*PolyFiQA*）： 在多语言和多模态语境中探讨金融推理（参见第 4 节），以此拓展任务 1 奠定的概念基础。该任务要求系统整合不同语言的金融报告和新闻，以完成复杂的分析型问答，检验其在语言和信息模态方面的泛化能力。– 任务 *3*：金融决策： 将任务 1 和任务 2 培养的知识与推理能力综合起来，转化为可执行情报（参见第 5 节）。该任务通过要求模型解读市场动态、情感和风险，并据此生成有证据支撑且具备风险意识的交易决策，评估其从推理到行动的能力。

这些任务共同构建了一个多层次评估框架，将理解、推理与决策贯通起来：这是推进可信、具备全球能力的金融人工智能系统的三大核心支柱。

2 相关工作

金融人工智能 金融自然语言处理已从特定任务建模发展到大规模领域适应。早期研究如FinBERT [12] 建立了面向情感与分类的领域特定变换器，随后扩展到个人理财 [7], 、信用评分 [6], 以及风险感知评估 [33]。包括REFinD、FinARG、FiNER-ORD和ECTSum [8, 16, 21, 27]在内的领域数据集，支持了实体识别、关系抽取、论证挖掘和摘要生成等细粒度任务。与此同时，英语金融自然语言处理基准测试，如FiQA [13], 、TAT-QA [37], 、FinQA [1],、ConvFinQA [2], 、FinanceMATH [35],和FinChain [31]，推动了情感分析、文本-表格推理、数值计算和可验证推理方面的进展。大型金融语言模型进一步推动了该领域的发展。BloombergGPT [26]展示了对金融自然语言处理任务的广泛覆盖，FinGPT [10]强调开源适应性，FinMA [28] 则以紧凑架构取得了有竞争力的性能。它们的出现催生了FLANG [22], 、FinBen [27], 和FinMTEB [24], 等基准测试，拓展了任务多样性与标准化程度。近期研究还指出，检索与证据选择是金融问答中的关键瓶颈：FINCARDS [36] 等方法将文档级问答重构为约束感知证据重排序，强调在实体、指标、财政期间和数值上的显式对齐。BizBench [9] 和PIXIU [28]等更新的评估进一步揭示了定量推理和多模态理解方面持续存在的不足，凸显了对能够稳健整合文本、数值与结构信号的模型的需求。在这些基础上，多语言金融自然语言处理已开始应对全球及跨文化语境。CFinBench [18], 、FLARE-ES [34], 、Plutus [19], 、SAHM [5],和MultiFin-Ben [20]等工作将金融基准扩展到英语之外，支持在多种语言和监管环境中进行评估。这些举措揭示了各语言在数据可用性、标注质量和任务覆盖范围方面的结构性差异，凸显了具有文化根基且语言包容的基准测试对于开发具备全球稳健性的金融大语言模型的重要性。

相关共享任务 随着人们对金融自然语言处理的兴趣日益增长，已有众多共享任务旨在应对虚假信息检测 [11], 金融分类、摘要生成、股票交易 [29],和因果关系抽取 [15]等挑战。虽然现有研究大多聚焦于英语，但也有新兴工作开始探索阿拉伯语 [14]等低资源语言。近期举措包括 Earnings2Insights [23], ，该任务评估系统能否根据财报电话会议文字稿生成有说服力的投资指导；Regulations Challenge [25], ，该任务通过九项任务对大语言模型理解和应用金融法规的能力进行基准测试；以及 Fin-DBQA [17],，该任务评估需要数据库查询、多跳推理和表格数据处理能力的多轮问答。总的来说，这些共享任务凸显了金融自然语言处理基准测试范围的不断扩大和复杂度的日益提升，但迄今尚无任务专门聚焦于多语言或多模态金融理解。

表 1. CLEF 2026 FinMMEval 实验室三个任务所涉及的语言。白色方块表示仅有测试数据的语言。

	英语	中文	阿拉伯语	印地语	希腊语	日语 西班牙语	
T1	■	■	■	■	□		■
T2	■	■			■	■	■
T3	■						

3 任务 1：金融考试题问答

动机。 专业金融资格考试（例如 CFA、EFPA）要求考生将理论与监管知识和应用推理结合起来。现有大语言模型往往依赖事实回忆，却未能展现人类考生所应具备的分析严谨性。本任务旨在评估模型能否在多语言金融语境中实现领域级理解和推理一致性。

任务定义。 给定一道独立的选择题 Q 及四个候选答案 $\{A_1, A_2, A_3, A_4\}$，系统必须选出正确答案 A^*。题目涵盖估值、会计、职业道德、公司金融和监管知识。重点在于概念理解和精准的金融推理，而非表层模式识别。

数据。 我们将现有的多语言金融考试数据集与新收集的材料结合起来：

– **EFPA** [20]（西班牙语）：230道源自EFPA认证考试备考材料的考试风格选择题，涵盖投资产品、投资组合管理、金融监管和职业标准。该数据集评估西班牙语环境下的应用金融知识与概念推理。– **GRFinQA** [19]（希腊语）：268道源自希腊大学层次金融、商业和经济学课程的考试风格选择题。这些问题由希腊语母语者人工整理并核验，以确保语言保真度与领域正确性。– **CFA** （英语） [3]: 600道考试风格选择题，涵盖九个核心领域，包括职业道德、定量方法、财务报告、公司金融和投资组合管理。这些问题强调概念理解以及专业级金融推理。– **CPA** （中文） [3]: 300道考试风格选择题，聚焦会计、审计、财务管理、税务、经济法和战略。该数据集面向中文职业认证情境中的领域特定金融知识。– **BBF** [4]（印地语）：500–1000道考试风格选择题，取自25项以上印度金融与机构考试，涵盖问题求解、金融数学、治理和监管主题。该数据集反映了印度金融考试格局的多样性。

–**SAHM** [5] （阿拉伯语）：873道来自SAHM 基准的考试风格选择题，包括源自真实阿拉伯语考试的会计与商业考试题目。该数据集评估阿拉伯语职业与学术情境中的金融与商业推理。

所有问题均由金融专业人士进行审核，以确保其正确性和概念平衡性。

评估。 模型必须输出正确答案，性能以准确率衡量，即测试集中正确识别选项所占的比例。参与者可以选择在可用语言的任意子集上评估其系统，包括单语言或多语言设置。提交结果将在相同的任务框架内进行评估，语言覆盖范围被视为可配置维度，而非用于定义独立的子任务。

4 任务 2：多语言金融问答

动机。 全球金融日益需要针对英语金融申报文件和外语新闻等文档进行多语言推理。然而，现有大多数基准测试都是单语言的。本任务引入首个跨语言金融推理基准，测试模型整合多语言文本信号，并基于真实金融数据进行分析推理的能力。

任务定义。 给定一份金融报告 R （英语 10-K 年报/10-Q 季报摘录）以及与同一公司申报文件 $N = \{N_{en}, N_{zh}, N_{ja}, N_{es}, N_{el}\}$ 相关的多语言新闻文章，模型会收到一个问题 q，并必须生成一个由多语言文本证据支持的答案 a。其中包含两个难度层级：

– **PolyFiQA-Easy：** 事实类或数值趋势类问题（例如营收增长、现金流异常）；– **PolyFiQA-Expert：** 需要进行多文档推理的复杂分析类问题（例如投资策略、资本配置）。

数据。 我们使用 PolyFiQA-Easy 和 PolyFiQA-Expert 数据集 [20], ，这些数据集结合了美国证券交易委员会申报文件与多语言新闻（英语、中文、日语、西班牙语、希腊语）。每个层级包含 172 个问答实例（总计 344 个）。所有问题和答案均由金融专家手动编写并验证，标注者间一致性超过 89%。该数据集依据 MIT 许可证发布。

评估。 参赛系统需生成简洁、基于证据的文本答案（不超过100词）。我们使用ROUGE-1作为主要评估指标来评估性能，并进一步计算BLEURT和事实一致性作为次要指标。

5 任务3：金融决策

动机。 现实世界投资决策需要整合异构信息源，以形成可付诸行动的洞见，例如文本新闻、市场价格动态以及当前投资组合持仓。与静态问答任务不同，本任务聚焦于从推理到行动：评估模型如何综合复杂且随时间变化的市场上下文，以生成既基于证据又具风险感知的交易策略。

任务定义。 给定市场上下文 $C = P_t, N_t$，其中 P_t 表示历史价格序列，N_t 表示同期文本信息（例如新闻、报告），模型需要预测以下三种离散操作之一：买入、持有或卖出。每项预测都必须附有简洁的文本理由（$\leq$50 字），明确引用支持性证据或推理过程。

为防止数据泄漏，该任务将在实时环境中的预先指定时间窗口内开展。在此期间，我们将按固定频率收集各团队的每日决策（例如！在截止时间前每日提交），且这些决策必须严格基于当天可获取的信息。

数据。 我们使用两个经过整理的 JSON 数据集，为 BTC 和 TSLA 提供每日市场上下文。我们的目标是对大语言模型交易代理进行压力测试，考察其在两种存在实质差异的市场微观结构中的表现：(*i*) 加密货币市场，全天候 24/7 交易，交易场所分散、市场状态转变更快且尾部风险更高；以及 (*ii*) 股票市场，按基于交易时段的交易时间运行，设有集合竞价开盘与收盘，存在公司事件驱动的跳涨跳跌，且市场环境监管更加严格。每个上下文均以 ISO 日期（YYYY-MM-DD）为索引，并汇总该交易日的异构信号：

– 资产与覆盖范围：

- BTC：从 2025年-08-01 开始，每日更新，语境也随之逐日扩展（尽管市场全天候交易，数据仍按日汇总快照）；
- TSLA：从 2025年-08-01 开始，每日更新，语境也随之逐日扩展。

– 每个日期 d 对应的字段：

- 价格（浮点数）：每天一个具有代表性的美元价值（非开高低收，也非日内序列）；
- 新闻（数组[字符串]）：一篇或多篇长篇文本综述，总结当天的市场及新闻流；
- 动量（类别）：{看涨，中性，看跌}；基于配套新闻手动标注的每日市场动量标签（并非计算得出的技术指标）；
- future_price_diff（浮点数 | null）：提前一天的价格变化$P_{d+1} - P_d$；在最后一个可用日期为 null；

- 10k、10q（数组）：与该资产相关的基本面申报文件。此组件专用于股票（针对 TSLA）；10-K 年报/10-Q 季报是高信号的基本面披露，会更新市场对收益能力、业绩指引和风险的预期。此类信息通常会被迅速纳入定价，尤其是在收益发布前后；这些申报文件可能触发大幅重新定价，对于某些公司而言，单日波动达到 10% 左右并不罕见。

– **格式：** UTF-8 JSON；键为日历日期（不含日内时间戳）。

该数据集提供用于决策的每日市场上下文 $C = \{P_t, N_t\}$ ，模型生成操作（买入/持有/卖出），而未来_价格_差值则作为监督信号，用于优化模型。

评估。 我们使用成熟的定量指标评估模型在盈利能力、稳定性和风险控制方面的表现：

– **主要指标：**累计收益率（CR）；– **次要指标：**夏普比率（SR）、最大回撤（MD）、日波动率（DV）和年化波动率（AV）。

这些指标共同反映模型在动态交易环境中平衡收益与风险、保持行为一致性并适应不同市场状态的能力。

6 结论与未来工作

我们介绍了CLEF 2026 FinMMEval实验室的设计与设置，该实验室首次提出面向金融大语言模型的多语言多模态评测框架。通过三项相互关联的任务，包括金融考试问答、多语言金融问答（PolyFiQA）和金融决策，该实验室从概念理解推进到现实世界中的推理与行动。这些任务共同构成连贯的评测层级，反映金融专业能力在实践中如何构建与应用：从知识获取，到跨语言与跨模态的分析整合，最终到基于证据的决策。

FinMMEval旨在推动透明、稳健且具有全球包容性的金融人工智能系统。通过提供多语言覆盖和多模态语境，它弥合了当前金融自然语言处理评测中的一项重要空白，超越以英语为中心的纯文本范式。数据集、评测代码和基线模型将公开发布，以鼓励金融自然语言处理社区的可复现性与协作。

未来的版本将进一步扩展覆盖范围，纳入更多语言，以及图表和监管文件等模态，并加入动态实时评测场景，从而为打造可信、适应性强且具备全球竞争力的金融大语言模型铺平道路。

致谢

我们感谢所有任务组织者和标注人员为 CLEF 2026 FinMMEval 实验室作出的宝贵贡献。我们也感谢 CLEF 社区提供协作平台和基础设施，助力开展这一多语言多模态评测计划。数据集和评测脚本将公开发布，以促进多语言多模态金融人工智能领域的进一步研究。Dimitar Dimitrov 和 Ivan Koychev 的工作部分由欧盟 NextGenerationEU 项目资助，该项目通过保加利亚共和国国家复苏与韧性计划提供资金，项目名称为 SUMMIT，编号 BG-RRP-2.004-0008。

利益披露

作者声明，他们不存在与本文内容相关的竞争性利益。

参考文献

1. Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R., Beane, M., Huang, T.H., Routledge, B., Wang, W.Y.：FinQA：一个基于金融数据进行数值推理的数据集。载于：2021年自然语言处理实证方法会议论文集。第3697–3711页（2021年） 2. Chen, Z., Li, S., Smiley, C., Ma, Z., Shah, S., Wang, W.Y.：ConvFinQA：探索对话式金融问答中的数值推理链。载于：2022年自然语言处理实证方法会议论文集。第6279–6292页。计算语言学协会（2022年） 3. Dai, Y., Lin, Y., Xie, Z., Wang, Y.：RealFin：用户省略信息时，大语言模型的金融推理能力有多强？预印本（2026年） 4. Devane, V., Nauman, M., Patel, B., Wakchoure, A.M., Sant, Y., Pawar, S., Thakur, V., Godse, A., Patra, S., Maurya, N., Racha, S., Singh, N.K., Nagpal, A., Sawarkar, P., Pundalik, K.V., Saluja, R., Ramakrishnan, G.：BhashaBench-finance：基于印度金融知识对人工智能进行基准测试。https://huggingface.co/datasets/ bharatgenai/bhashabench-finance（2025年），访问日期：2025-10-14 5. Elbadry, R., Ahmad, S., Bouch, D., Ahsan, M., Peng, X., Huang, J., Anwar, M., AlMahri, M., Khalil, M.E., Wang, Y., Lahlou, S., Stoyanov, V., Ananiadou, S., Nakov, P., Xie, Z.：SAHM：阿拉伯金融与符合伊斯兰教法推理的基准。预印本（2026年） 6. Feng, D., Dai, Y., Huang, J., Zhang, Y., Xie, Q., Han, W., Chen, Z., Lopez-Lira, A., Wang, H.：赋能多数人，偏袒少数人：借助大语言模型实现通用信用评分。arXiv 预印本 arXiv:2310.00566（2023年） 7. Hean, O., Saha, U., Saha, B.：人工智能能帮你打理个人财务吗？《应用经济学》**57**，1–9（2025年） 8. Kaur, S., Smiley, C., Gupta, A., Sain, J., Wang, D., Siddagangappa, S., Aguda, T., Shah, S.：REFinD：金融关系抽取数据集。载于：第46届国际 ACM SIGIR 信息检索研究与发展会议论文集。第3054–3063页（2023年）

9. Koncel-Kedziorski, R., Krumdick, M., Lai, V., Reddy, V., Lovering, C., Tanner, C.：BizBench：面向商业和金融的定量推理基准。载于：第62届计算语言学协会年会论文集（第1卷：长篇论文）。第8309–8332页（2024年） 10. Liu, X.Y., Wang, G., Yang, H., Zha, D.：FinGPT：为金融大语言模型普及互联网规模数据。arXiv 预印本 arXiv:2307.10485（2023年） 11. Liu, Z., Wang, K., Bao, Z., Zhang, X., Dong, J., Yang, K., Kabir, M., Giannouris, P., Xing, R., Park, S., Kim, J., Li, D., Xie, Q., Ananiadou, S.：FinNLP-FNP-LLMFinLegal-2025 共享任务：金融错误信息检测挑战任务。载于：第9届金融科技与自然语言处理研讨会、第6届金融叙事处理研讨会和首届金融与法律大语言模型研讨会联合论文集。第271–276页（2025年） 12. Liu, Z., Huang, D., Huang, K., Li, Z., Zhao, J.：FinBERT：用于金融文本挖掘的预训练金融语言表示模型。载于：第二十九届国际人工智能联合会议论文集。IJCAI'20（2021年） 13. Maia, M., Handschuh, S., Freitas, A., Davis, B., McDermott, R., Zarrouk, M., Balahur, A.：WWW'18 开放挑战赛：金融观点挖掘与问答。载于：2018年万维网大会伴随论文集。第1941–1942页（2018年） 14. Malaysha, S., El-Haj, M., Ezzini, S., Khalilia, M., Jarrar, M., Almujaiwel, S., Berrada, I., Bouamor, H.：AraFinNLP 2024：首个阿拉伯语金融自然语言处理共享任务。载于：第二届阿拉伯语自然语言处理会议论文集。第393–402页（2024年） 15. Mariko, D., Abi-Akl, H., Labidurie, E., Durfort, S., De Mazancourt, H., El-Haj, M.：金融文档因果关系检测共享任务（FinCausal 2020）。载于：首届金融叙事处理与多语言金融摘要联合研讨会论文集。第23–32页（2020年） 16. Mukherjee, R., Bohra, A., Banerjee, A., Sharma, S., Hegde, M., Shaikh, A., Shrivastava, S., Dasgupta, K., Ganguly, N., Ghosh, S., Goyal, P.：ECTSum：用于长篇财报电话会议文字稿要点摘要的新基准数据集。载于：2022年自然语言处理实证方法会议论文集。第10893–10906页。计算语言学协会（2022年） 17. Nararatwong, R., Kertkeidkachorn, N., Takamura, H., Ichise, R.：Fin-DBQA 共享任务：数据库查询与推理。载于：第9届金融科技与自然语言处理研讨会、第6届金融叙事处理研讨会和首届金融与法律大语言模型研讨会联合论文集。第385–391页（2025年） 18. Nie, Y., Yan, B., Guo, T., Liu, H., Wang, H., He, W., Zheng, B., Wang, W., Li, Q., Sun, W., Wang, Y., Tao, D.：CFinBench：面向大语言模型的综合中文金融基准。载于：2025年美洲计算语言学协会分会会议：人类语言技术论文集（第1卷：长篇论文）。第876–891页（2025年） 19. Peng, X., Papadopoulos, T., Soufleri, E., Giannouris, P., Xiang, R., Wang, Y., Qian, L., Huang, J., Xie, Q., Ananiadou, S.：Plutus：低资源希腊语金融领域大语言模型基准测试。载于：2025年自然语言处理实证方法会议论文集。第30176–30202页（2025年） 20. Peng, X., Qian, L., Wang, Y., Xiang, R., He, Y., Ren, Y., Jiang, M., Zhao, J., He, H., Han, Y., Feng, Y., Jiang, Y., Cao, Y., Li, H., Yu, Y., Wang, X., Gao, P., Lin, S., Wang, K., Yang, S., Zhao, Y., Liu, Z., Lu, P., Huang, J., Wang, S., Papadopoulos,

T., Giannouris, P., Soufleri, E., Chen, N., Xiong, G., Deng, Z., Zhao, Y., Lin, M., Qiu, M., Smith, K.E., Cohan, A., Liu, X.Y., Huang, J., Lopez-Lira, A., Chen, X., Tsujii, J., Nie, J.Y., Ananiadou, S., Xie, Q.：MultiFinBen：用于金融大语言模型评测的多语言、多模态且具备难度感知能力的基准测试。arXiv预印本 arXiv:2506.14028（2025年）21. Shah, A., Gullapalli, A., Vithani, R., Galarnyk, M., Chava, S.：FiNER- ORD：金融命名实体识别开放研究数据集。arXiv预印本 arXiv:2302.11157（2023年）22. Shah, R.S., Chawla, K., Eidnani, D., Shah, A., Du, W., Chava, S., Raman, N., Smiley, C., Chen, J., Yang, D.：当 FLUE 遇上 FLANG：金融领域的基准测试与大型预训练语言模型。载于：2022年自然语言处理实证方法会议论文集。第2322–2335页（2022年）23. Takayanagi, T., Goldsack, T., Izumi, K., Lin, C., Takamura, H., Chen, C.C.：Earnings2Insights：面向投资指导的分析师报告生成。载于：第10届金融科技与自然语言处理研讨会论文集。第246–251页（2025年）24. Tang, Y., Yang, Y.：FinMTEB：金融大规模文本嵌入基准。载于：2025年自然语言处理实证方法会议论文集。第3620–3638页（2025年）25. Wang, K., Patel, J., Shen, C., Kim, D., Zhu, A., Lin, A., Borella, L., Osborne, C., White, M., Yang, S., Xiao, K., Liu, X.Y.：FinNLP-FNP-LLMFinLegal-2025 共享任务：法规挑战。载于：第9届金融科技与自然语言处理研讨会、第6届金融叙事处理研讨会和首届金融与法律大语言模型研讨会联合工作坊论文集。第363–370页（2025年）26. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G.：BloombergGPT：面向金融的大语言模型。arXiv预印本 arXiv:2303.17564（2023年）27. Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., Xiao, M., Li, D., Dai, Y., Feng, D., Xu, Y., Kang, H., Kuang, Z., Yuan, C., Yang, K., Luo, Z., Zhang, T., Liu, Z., Xiong, G., Deng, Z., Jiang, Y., Yao, Z., Li, H., Yu, Y., Hu, G., Huang, J., Liu, X.Y., Lopez-Lira, A., Wang, B., Lai, Y., Wang, H., Peng, M., Ananiadou, S., Huang, J.：FinBen：面向大语言模型的综合金融基准测试。载于：神经信息处理系统进展。第37卷（2024年）28. Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., Huang, J.：PIXIU：面向金融的大语言模型、指令数据与评测基准。arXiv预印本 arXiv:2306.05443（2023年）29. Xie, Q., Huang, J., Li, D., Chen, Z., Xiang, R., Xiao, M., Yu, Y., Somasundaram, V., Yang, K., Yuan, C., Luo, Z., Liu, Z., He, Y., Jiang, Y., Li, H., Feng, D., Liu, X.Y., Wang, B., Wang, H., Lai, Y., Suchow, J., Lopez-Lira, A., Peng, M., Ananiadou, S.：FinNLP-AgentScen-2024 共享任务：大语言模型面临的金融挑战——FinLLMs。载于：第八届金融科技与自然语言处理研讨会暨首届用于情景规划的智能体人工智能研讨会论文集。第119–126页（2024年）30. Xie, Z., Cohn, T., Lau, J.H.：下一篇章：大语言模型故事讲述研究。载于：第16届国际自然语言生成会议论文集。第323–351页（2023年）31. Xie, Z., Orel, D., Thareja, R., Sahnan, D., Madmoun, H., Zhang, F., Banerjee, D., Georgiev, G., Peng, X., Qian, L., Huang, J., Su, J., Singh, A., Xing, R., Elbadry,

R., Xu, C., Li, H., Koto, F., Koychev, I., Chakraborty, T., Wang, Y., Lahlou, S., Stoyanov, V., Ananiadou, S., Nakov, P.：FinChain：用于可验证金融思维链推理的符号基准。arXiv预印本 arXiv:2506.02515（2026年）32. Xie, Z., Xing, R., Wang, Y., Geng, J., Iqbal, H., Sahnan, D., Gurevych, I., Nakov, P.：FIRE：通过迭代检索与验证进行事实核查。载于：计算语言学协会论文集：NAACL 2025 会议论文集。第2901–2914页（2025年）33. Yuan, T., He, Z., Dong, L., Wang, Y., Zhao, R., Xia, T., Xu, L., Zhou, B., Li, F., Zhang, Z., Wang, R., Liu, G.：R-Judge：评测大语言模型智能体的安全风险意识。载于：计算语言学协会论文集：EMNLP 2024 会议论文集。第1467–1490页（2024年）34. Zhang, X., Xiang, R., Yuan, C., Feng, D., Han, W., Lopez-Lira, A., Liu, X.Y., Qiu, M., Ananiadou, S., Peng, M., Huang, J., Xie, Q.：Dólares or dollars? 揭示西班牙语与英语金融大语言模型的双语能力。载于：第30届ACM SIGKDD知识发现与数据挖掘大会论文集。第6236–6246页（2024年）35. Zhao, Y., Liu, H., Long, Y., Zhang, R., Zhao, C., Cohan, A.：FinanceMATH：金融领域的知识密集型数学推理。载于：第62届计算语言学协会年会论文集（第1卷：长篇论文）。第12841–12858页（2024年）36. Zhou, Y., Zhang, F., Chen, Y., Zhang, H., Nakov, P., Xie, Z.：FinCARDS：基于卡片的金融文档问答分析师重排序。arXiv预印本 arXiv:2601.06992（2026年）37. Zhu, F., Lei, W., Huang, Y., Wang, C., Zhang, S., Lv, J., Feng, F., Chua, T.S.：TAT-QA：针对金融领域表格与文本混合内容的问答基准。载于：计算语言学协会年会暨第11届自然语言处理国际联合会议论文集（第1卷：长篇论文）。第3277–3287页（2021年）