arXiv:2603.22305v1 [cs.LG] 18 Mar 2026

CN-Buzz2Portfolio：基于每日热门财经新闻的宏观与行业资产配置大语言模型中国市场数据集与基准

Liyuan Chen[1,2], Shilong Li[2],Jiangpeng Yan[2],Shuoling Liu[2],Qiang Yang[3],Xiu Li[1]

[1]清华大学深圳国际研究生院[2]易方达基金管理有限公司[3]香港理工大学**电子邮箱：**{lishilong, yanjiangpeng}@efunds.com.cn

摘要

大语言模型（LLM）正迅速从情感分析和事件抽取等静态自然语言处理（NLP）任务，转向在复杂金融环境中充当动态决策智能体。然而，大语言模型演变为自主金融智能体，在评估范式方面面临重大困境。直接进行实盘交易既无法复现，也容易产生结果偏差，因为运气和能力会相互混淆；而现有静态基准往往局限于实体级选股，忽视了更广泛的市场关注度。为推动对这些挑战开展严谨分析，我们推出**CN-Buzz2Portfolio**，这是一个以中国市场为基础、可复现的基准，旨在将每日热门新闻映射为宏观与行业资产配置。数据集覆盖从2024年至2025年年中的滚动时域，模拟真实的公众注意力流，要求智能体从高曝光度叙事中提炼投资逻辑，而非依赖预先筛选的实体新闻。我们提出三**阶段CPA 智能体工作流**，包含压缩、感知和配置三个阶段，用于评估大语言模型在交易型开放式指数基金（ETF）等广泛资产类别（而非个股）上的表现，从而降低特质波动率。对九种大语言模型开展的大量实验揭示了不同模型将宏观层面叙事转化为投资组合权重的能力存在显著差异。本研究为一般推理能力与金融决策之间的对齐提供了新的见解，并公开所有数据、代码和实验，以推动金融智能体研究的可持续发展。

1 引言

风险提示：“本工作仅供学术研究。所有实验均在假设简化的模拟市场环境中进行。本文内容不构成投资建议，若将其用于实际交易，风险由您自行承担。”

大语言模型（LLM）融入金融领域，正推动研究前沿从被动文本分析转向主动的金融通用智能体（Chen 等人，2025a；Guo 和 Shum，2025年；Huang 等人，2025）。然而，在高噪声、非平稳市场中评估这些智能体仍是一项尚待解决的挑战。

如图1所示，当前的评估方法呈现两极分化。一方面，**直接实盘交易**平台，例如*AI-Trader*(香港大学数据科学实验室，2025年)、*RockFlow AIArena* (RockFlow 团队，2025年)和 *nof1.ai* (NoF1 团队，2025 年)，虽然真实感十足，但在科学研究上受限于不可复现性。具体而言，盈利结果往往难以区分究竟源于合理推理还是市场运气。另一方面，**静态基准**如*InvestorBench*(Li 等人，2025b) 和*FinMem* (Yu 等人，2025b) 虽然实现了标准化，但通常聚焦于狭窄任务或历史选股，无法反映开放世界信息流的复杂性。

我们认为，稳健的评估框架必须充当**诊断工具**，以弥合**双层评估瓶颈**，该瓶颈阻碍了金融智能体的系统性发展：

1. **推理对齐（语义 → 逻辑）：**在摘要等通用自然语言处理任务中表现出色，并不意味着其投资逻辑必然有效。智能体必须展现出将叙事转化为可执行金融假设的能力，而这一过程在端到端评估中往往难以显现。

2. **归因噪声（逻辑 → 结果）：**在个人股票市场等高方差环境中，价格变动往往受特异性噪声主导。这会造成归因差距：即使逻辑严谨的智能体也可能因运气不佳而遭受损失，而存在缺陷的智能体则可能偶然获利，从而使决策的验证更加复杂。

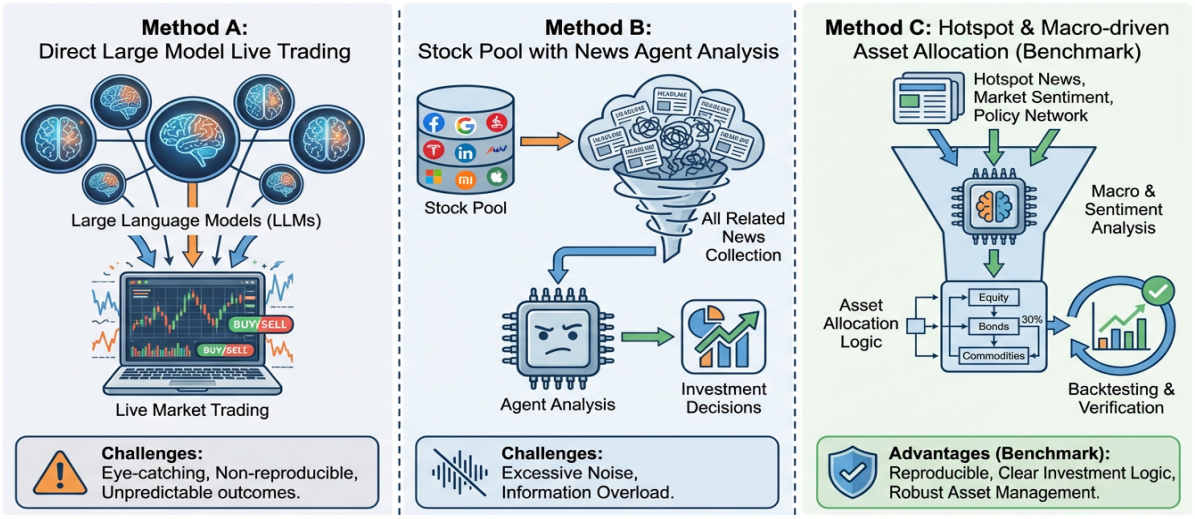

图1：**金融智能体研究范式比较。（a）直接实盘交易：**智能体与实时市场交互。尽管这种方法最贴近现实，但在**可复现性和归因**方面存在科学挑战，因此难以从市场随机性中分离出有效推理。**（b）实体中心基准（例如 StockBench）：**将新闻映射到预定义目标股票的标准范式。这忽略了“公众注意力筛选”过程，而且个股层面往往**特质波动率较高**，使逻辑一致性的验证变得复杂。**（c）CN-Buzz2Portfolio（本文提出）：**一种滚动时域基准，模拟从每日热门新闻（公众注意力）到宏观与行业配置的流程。通过配置多元化资产类别来降低噪声，该框架可作为**诊断工具**，严格评估语义理解与可验证的投资组合逻辑之间的一致性。

现有范式无法解决这一瓶颈，原因在于当前数据集设计存在两个关键局限性：

局限性1：范围错配（实体中心与市场叙事驱动）。主流基准测试（Li 等人，2025年b）通常遵循实体中心范式，由系统检索预定义目标股票池对应的新闻。然而，现实世界的交易在很大程度上受到**公众注意力**和**信息曝光**（Barber 和 Odean，2008）的驱动。交易者身处“热门话题”信息流中，例如政策变化或全球宏观事件，并必须自主识别相关行业。现有研究通过预先筛选特定实体的新闻，绕过了这一“叙事筛选”机制，因而未能检验智能体从原始全市场信息中发现机会的能力。

局限性 2：新兴市场中宏观语义推理的匮乏。当前基准测试主要围绕成熟的美国市场和微观层面的股票预测展开。然而，在高度敏感于政策叙事和行业整体市场情绪的中国市场中，稳定的收益率往往通过**资产配置**和**行业轮动**获得，而不是通过个股选取。当前缺乏高质量的基准，来评估跨层级推理能力：即在源文本未明确提及股票实体的情况下，将市场叙事和公众情绪映射至广泛资产篮子（例如交易型开放式指数基金）的能力。

为应对这些挑战，我们推出**CN-Buzz2Portfolio**，这是一个专为宏观语义金融推理设计的滚动时域基准测试。我们的贡献总结如下：

- **基准测试（2024—2025年滚动时域）：**我们整理了一个基于多平台每日热门新闻的数据集，用以模拟现实世界中的“公众注意力流”。我们将完整数据集、评估代码和实验结果开源。[1]。
- **任务（市场注意力到资产配置）：**我们提出一项新任务，要求智能体依据热门叙事，使用多元化交易型开放式指数基金（宏观与行业）构建投资组合，将重点从噪声较大的股票层面预测转向逻辑驱动的资产配置。
- **评估：**借助标准化的 **三阶段 CPA 智能体工作流**，我们对顶尖大语言模型进行了比较分析，揭示了其在政策敏感型金融环境中的不

[1]正式发布后，链接将更新为官方 GitHub 仓库。

同决策行为模式。

2 相关工作

金融领域的大语言模型（LLM）研究迅速扩展。我们将现有研究分为三个方向：金融评估基准、自治智能体框架和语义对齐研究。

2.1 金融大语言模型基准与预测任务

早期研究（如 FinBERT（Araci，2019））侧重于静态情感分类。随着 FinGPT 系列（Yang 等人，2023年；Zhang 等人，2023a、b；Wang 等人，2023年；Liu 等人，2023年）和 BloombergGPT（Wu 等人，2023年）等基础模型的出现，研究重点转向了生成能力。近期的基准测试已对金融决策的评估进行了标准化。Stock-Bench（Chen 等人，2025b）为股票走势预测提供了全面的测试平台。InvestorBench（Li 等，2025b）评估智能体在多种金融任务上的表现，包括行为分析和交易。AlphaFin（Li 等人，2024b）和 LiveTradeBench（Yu 等人，2025a）整合了现实世界的新闻检索，以寻求交易超额收益，推动模型处理时间序列信息。

这些基准测试中一种主流的方法遵循**以实体为中心的信息搜集** 范式：系统从预先定义的目标股票池出发，检索与这些实体相关的新闻，并将其输入大语言模型进行预测。尽管这种方法有助于验证实体层面的推理能力，却简化了真实市场中固有的**“公众注意力筛选”** 挑战：投资者必须从开放世界的热门话题流中自主识别哪些资产与投资相关，而不能依赖预设目标。

2.2 自主交易智能体与实时系统

从预测迈向执行，研究人员开发了智能体框架。**Fin-Robot**(Yang 等人，2024年）和**TradingAgents**(Xiao 等人，2024年）将交易分解为感知、推理和执行模块。**Deep-Fund** (Li 等人，2025a）将深度强化学习应用于连续投资组合管理，而**FinMem** (Yu 等人，2025b）则通过为智能体配备分层记忆，使其能够适应不断变化的市场状况。

2024–2025年也见证了**实时交易系统**和竞赛的兴起。诸如**AI-Trader** (HKUDS)（HKU 数据科学实验室，2025年）和**RockFlow AI**(RockFlow 团队，2025年）等平台允许多种大语言模型（如 GPT-4、DeepSeek）在美国股票市场进行实时竞赛。在加密货币领域，Truth Terminal**TruthTerminal** (Ayrey，2024年）和**nof1.ai** (NoF1 团队，2025年）等自主智能体以自主投资型去中心化自治组织的形式运作，通过社交叙事推动资产价格变动。我们的工作为这些高自主性系统提供补充，构建了一个可复现的滚动时域测试平台，将推理逻辑与实盘交易环境中常见的随机波动（运气）分离开来。

2.3 跨文化与宏观语义对齐

虽然像**C-Eval**（Huang 等人，2023年）和**CMMLU**（Li 等人，2024a）这样的通用中文基准测试能够评估语言能力，却很少涵盖专业领域的金融逻辑。在A股市场中，政策叙事（例如“逆周期调节”）会导致与西方市场截然不同的资产表现。

目前缺少能够评估**宏观语义推理**的基准测试——即将抽象的政策和市场情绪变化映射为广泛的资产配置策略（例如行业/宏观ETF），而非个股选股的能力。

3 CN-Buzz2Portfolio基准测试

3.1 任务形式化

我们将投资决策过程形式化为不确定性下的序列映射任务。我们不采用强化学习目标，而是关注智能体从不完整观测中推导出可执行决策的能力。在每个时间步 t，智能体根据以下组成部分进行操作：

- **观测（$\mathcal{O}_t$）：** 在一个滚动时间窗口内，智能体观测一个元组$\langle N_t$、P_{hist}、T_{hist}、$H_t\rangle$。其中，N_t 表示非结构化的**Buzz信息流**（开放世界中的热门新闻），P_{hist} 和 T_{hist} 分别表示历史市场价格和交易记录，H_t 则表示当前投资组合状态。“Buzz”内容混杂着金融政策、社会事件以及平台特有的诱导点击信息，且没有预定义的实体映射，因此需要智能体自主识别

相关性。

- **动作空间（$\mathcal{A}_t$）：** 动作定义为一条再平衡指令 w_{t+1}，用于下一期。为确保可复现性并聚焦于推理，动作空间被限制为程序化执行接口，使智能体能够专注于高层战略决策，详情见第3.5节。

3.2 数据构建："热度"流

我们汇总了来自 4 个主要中国金融平台的每日排名前20的热门话题，并将完整热门榜单作为输入。我们严格执行时间戳筛选：仅使用收盘前于第 T 日发布的新闻来决定第 T 日的配置（在收盘时执行），以防止"前视偏差"。详细的数据集统计见附录A。

3.3 资产池：宏观与行业视角

为评估从宏观到行业的推理，我们使用交易型开放式指数基金（ETF）联接基金构建了两个不同的资产池。这些基金可作为精细化画布，将语义逻辑映射到市场板块：

- **任务A（宏观与主题配置）：** 此任务评估对经济周期的解读能力。投资范围包括**11个宽基指数**，涵盖股票、债券和黄金等主要资产类别，以及大盘和小盘指数等不同市场风格。

- **任务B（行业轮动）：** 此任务要求深入理解产业政策。我们选取**14只行业ETF**，代表中国工业链中的关键环节，例如新能源和科技、媒体和电信（TMT）。

有关流动性和管理资产规模的详细筛选标准见**附录B**。

3.4 UnifiedTradingProtocol：三阶段CPA多智能体框架

为提供标准化评估流程，我们建立了三阶段CPA多智能体框架。该框架旨在从热点新闻中提取市场情绪和政策倾向，并生成投资逻辑和投资组合再平衡指令。

阶段 1：压缩（摘要器 $\mathcal{A}_{sum}$）。 原始趋势榜单往往包含噪声，例如标题党内容或非金融社会事件。$\mathcal{A}_{sum}$ 充当信息过滤器，将嘈杂的 N_t 提炼为结构化的金融相关事件列表。此过程可在深入分析前显著提高信噪比。

阶段2：感知（分析器 $\mathcal{A}_{ana}$）。 $\mathcal{A}_{ana}$ 作为分析引擎运行。该模块将提炼后的事件与资产定义结合起来进行处理，评估不同新闻叙事可能如何影响各行业和指数。它通过逻辑推理评估整体市场情绪并识别潜在机会，专注于叙事影响，而不依赖技术价格数据。

第3阶段：配置（交易者 $\mathcal{A}_{trade}$）。 $\mathcal{A}_{trade}$充当执行控制器。它将 $\mathcal{A}_{ana}$ 的定性见解与历史数据（P_{hist}、T_{hist}）及当前持仓（H_t）相结合。最终输出包括投资逻辑和具体的再平衡指令。

3.5 执行层与行动约束

为尽量减少大语言模型中的算术错误，我们将数值计算交由确定性执行引擎处理。我们设计了一种结构化的基于指令的行动空间，将定性意图转化为精确交易，并以散户投资者行为为参考：

- **基于预算的配置（买入）：** 智能体指定购买的金额（例如，"将 5,000 人民币配置到 AssetX"）。这可以避免模型在计算股数与价格的乘积时出错。
- **基于比例的持仓管理（卖出）：** 智能体指定要卖出的当前持仓比例（例如，"卖出 As-set Y 的 50%"）。这种方法可以减少卖空错误，并模拟止盈等合理的风险管理策略。

4 实验设置

4.1 评估期间

我们选取具有挑战性的时间窗口，以测试模型在不同市场状态下的稳健性：

- **阶段1（2024全年）：** 这是一个"由熊转牛"的过渡时期，其特点是波动剧烈、政策变化频繁（例如"新国九条"改革）。这将考验智能体适应市场状态变化的能力。

• **第二阶段（2025年上半年）：** 这是一个“高波动震荡”时期。在此阶段，市场指数（沪深300指数）大幅波动，但**净值变化较低**（横盘走势）。这将考验智能体在缺乏明确方向性贝塔的市场中，通过精准择时和轮动创造阿尔法的能力。

4.2 模拟环境

我们构建了一个散户模拟环境（针对个人散户投资者）。这一选择经过审慎考虑，旨在确保普通投资者也能实现这些结果，而非仅适用于理论上的机构回测。

• **资金限制：** 初始资金为100,000元人民币。

• **资产代理：** 为方便投资者参与，我们使用ETF联接基金。执行价格采用标的ETF的收盘价，以确保较高的数据保真度并实现流动性匹配。

• **交易成本：** 我们采用0.01%（1个基点）的现实费率。这反映了中国交易型开放式指数基金具有竞争力的低佣金结构。虽然费率较低，但仍可对过高的换手率形成惩罚，避免智能体随机频繁买卖和过度交易。

•**频率：** 每日收盘时进行再平衡。这与“Buzz”榜单的日度频率相一致。

4.3 模型库与选择逻辑

我们评估了九种多样化的先进大语言模型，并根据其底层推理范式和架构规模进行分类。选择这些模型旨在比较专用推理模型与通用指令模型在政策敏感型金融决策情境中的表现。

•**推理型模型：**此类模型包括**DeepSeek-R1**（DeepSeek-AI，2025年），**Qwen-3-Max-Think**（Qwen 团队，2025年），以及**Qwen-3-32B-Think**（Qwen 团队，2025年）。这些模型采用集成的思维链（CoT）流程，非常适合用于测试深入推理能力；这种能力对于分析热门叙事的复杂影响并将其转化为连贯的投资逻辑至关重要。

• **通用指令模型：**我们纳入了全球前沿模型，例如**GPT-5**（OpenAI，2025年）和**Gemini-2.5-Pro**（Google Gemini 团队，2025年），以及领先的国内模型，包括**DeepSeek-V3**(DeepSeek-AI，2024年），**GLM-4.6**(人工智能，2025年），**通义千问-3-Max**和**通义千问-3-32B**。这些模型构成了零样本金融语境下指令遵循和语义压缩能力的基准。

4.4 评估指标

为对智能体表现进行多维评估，我们采用以下金融和运营指标：

• **累积收益率：**评估期内投资组合价值的总百分比变化。这是衡量智能体最终盈利能力的主要指标。

• **夏普比率：**衡量风险调整后收益的指标，计算方式为超额收益率与收益率标准差之比。SR 越高，表明智能体的逻辑越能有效地平衡收益与波动。

•**最大回撤（MaxDD）：**投资组合价值从峰值到谷值的最大跌幅。该指标用于评估智能体的风险控制能力，以及其在不利市场变化期间的韧性。

• **波动率：**投资组合日收益率的标准差，反映价格波动的剧烈程度。该指标用于评估智能体面临的市场风险，以及其在高不确定性环境下维持权益曲线稳定的能力。

5 实验结果与分析

我们从多个维度分析智能体行为，涵盖财务表现、决策一致性和定性研究。附录中提供了对成功战略推理的详细案例研究，以及典型失败模式C。

5.1 基线有效性与市场背景

表1 全面展现了模型在整个滚动时域（2024–2025）中的能力。

• **总体有效性：** 在这两个时期内，三智能体流程均成功帮助大多数大规模模型实现了正的绝对收益，验证了“Buzz”信息流中蕴含着可提取的金融信号。

• **2024年任务A（宏观）中的“贝塔陷阱”：**在2024年任务A中，我们观察到若干模

模型	2024年任务A：宏观配置				2024年任务B：行业轮动			
	收益率	夏普比率	最大回撤	波动	收益率	夏普比率	最大回撤	波动
推理模型								
DeepSeek-R1	7.47%	0.51	-10.71%	0.177	28.37%	1.14	**-12.93%**	0.255
Qwen3-32B-Think	5.56%	0.50	**-7.66%**	**0.128**	21.67%	0.94	-16.21%	0.249
Qwen3-Max-Think	8.98%	0.62	-10.95%	0.167	**40.11%**	**1.52**	-15.84%	0.251
标准模型								
DeepSeek-V3	9.28%	0.62	-11.48%	0.174	6.97%	0.42	-19.51%	**0.224**
Gemini-2.5-Pro	**10.67%**	**0.67**	-10.89%	0.181	18.10%	0.80	-17.10%	0.255
GPT-5	7.31%	0.53	-10.79%	0.164	24.42%	1.07	-16.35%	0.239
GLM-4.6	6.25%	0.47	-12.01%	0.159	31.58%	1.25	-14.11%	0.254
Qwen3-32B	5.59%	0.44	-11.80%	0.158	13.26%	0.63	-17.31%	0.257
Qwen3-Max	8.03%	0.60	-10.31%	0.155	26.50%	1.03	-16.02%	0.273
量化基准								
动量	5.46%	0.40	-13.55%	0.143	11.35%	0.48	-17.68%	0.250
MVO	**16.99%**	**1.73**	**-8.64%**	**0.103**	**17.76%**	**0.87**	**-14.39%**	**0.214**
市场基准								
沪深300指数	16.20%	0.84	-14.41%	0.214	16.20%	0.84	-14.41%	0.214
朴素等权投资组合	9.21%	0.53	-15.60%	0.218	9.80%	0.50	-18.05%	0.263

模型	2025年任务A：宏观配置				2025年任务B：行业轮动			
	收益率	夏普比率	最大回撤	波动	收益率	夏普比率	最大回撤	波动
推理模型								
DeepSeek-R1	**11.21%**	1.81	**-4.17%**	0.130	6.08%	0.70	-16.19%	0.215
Qwen3-32B-Think	10.49%	**1.98**	-4.36%	**0.111**	5.78%	0.83	**-10.79%**	**0.160**
Qwen3-Max-Think	6.37%	1.08	-8.04%	0.131	2.84%	0.45	-11.14%	0.165
标准模型								
DeepSeek-V3	7.41%	1.16	-6.20%	0.140	5.58%	0.69	-13.40%	0.198
Gemini-2.5-Pro	7.76%	1.24	-7.69%	0.137	7.38%	0.88	-13.16%	0.195
GPT-5	3.84%	0.69	-8.38%	0.128	5.12%	0.70	-12.13%	0.176
GLM-4.6	11.06%	1.72	-5.40%	0.136	5.03%	0.63	-14.47%	0.198
Qwen3-32B	8.99%	1.46	-5.56%	0.133	1.44%	0.27	-11.61%	0.166
Qwen3-Max	9.48%	1.58	-6.90%	0.129	**8.81%**	**1.04**	-13.01%	0.192
量化基准								
动量	2.08%	0.20	-16.56%	0.230	1.79%	0.16	-17.83%	0.247
MVO	3.78%	0.98	**-3.56%**	**0.086**	4.36%	0.67	**-9.12%**	**0.146**
市场基准								
沪深300指数	3.03%	0.50	-10.49%	0.156	3.03%	0.50	-10.49%	0.156
朴素等权投资组合	5.17%	0.77	-10.51%	0.159	5.14%	0.66	-12.59%	0.193

表1：在两种截然不同的市场状态下的比较表现：**2024年高波动动量环境**和**2025年低收益区间震荡环境**。在高波动市场状态下，推理型模型在复杂的行业轮动任务中展现出更强的阿尔法收益生成能力，而通用指令模型在稳定的宏观配置中则保持了较强的鲁棒性。量化基准指经典算法策略：动量和均值-方差优化。市场基准指更广泛的指数（沪深300指数），而朴素等权投资组合则表示对所有候选资产进行等权配置。2025年的收益率表示期间累计收益率（未年化）。

型（如 DeepSeek-V3、GLM-4.6）的表现落后于市场基准（沪深300指数收益率：16.20%）。这种表现不佳源于2024年的特定市场环境：市场先经历了漫长的熊市阶段，随后在后期出现了**剧烈的政策驱动型上涨行情**（系统性贝塔爆发）。沪深300指数作为100%的股票指数，完整反映了这一波动。相比之下，我们的智能体作为理性的主动管理者——管理者——在熊市阶段通常会维持防御性仓位（黄金/债券/现金）以控制回撤。因此，尽管他们降低了风险，但在突发的流动性驱动飙升期间，自然会落后于原始指数。这反映的是现实中的“主动管理”行为，而非模型失效。

- **任务B（行业）中的结构性阿尔法：** 相反，在任务B（行业轮动）中，模型的表现显著优于基准测试。这

表明，尽管智能体在广泛资产敞口方面可能较为保守（任务A），但它们擅长识别**结构性机遇**——将资本配置到从新闻中识别出的特定领先行业，从而创造出显著超越市场贝塔的阿尔法。

- **滚动更新的必要性：**2024年（趋势）与2025年（振荡）之间的性能差异凸显了滚动更新设计的关键价值。静态基准存在让模型“记住”历史数据的风险（例如，通义千问在2024年的高性能可能涉及隐性知识泄漏）。通过持续引入未见数据，**CN-Buzz2Portfolio**能够主动缓解“前视”问题，确保评估聚焦于**真正的时间泛化**。

5.2 方差分解：模型能力与随机性

为验证观察到的性能差距具有统计显著性，而非随机初始化造成的假象，我们进行了方差分解分析（表2）。

方差统计量（按 10^{-3} 缩放）	2024 宏观	2024 行业	2025 宏观	2025 行业
总方差	1.57	9.14	0.48	0.50
模型间方差（σ^2_{model}）	1.26	6.43	0.25	0.33
模型内方差（$\sigma^2_{stochastic}$）	0.32	2.71	0.23	0.17
效能比（$\sigma^2_{model}/\sigma^2_{stochastic}$）	**3.98**	**2.37**	**1.09**	**1.98**
平均推理稳定性 *($\bar{\sigma}$)*	1.27	1.58	0.83	1.14

表 2：**方差分解分析**。 效能比用于量化**模型叙事处理一致性** 相对于**随机噪声**的主导程度。比值显著高于 1.0 表明，性能是由算法逻辑而非随机方差驱动的。

模型间方差在 2024年 显著高于随机成分（比率$>$ 2.3），这证实了模型表现层级具有稳健的结构性，并反映出模型推理能力的差异。然而，在**2025年任务A**中，该比率趋近于持平（**1.09**）。这一现象表明，在以均值回归动态为特征的低波动市场状态下，当前大语言模型的信噪比会降低，使其决策过程几乎与**随机排列**无法区分。这些结果揭示了一个关键边界：当代**智能体主要充当依赖市场状态的决策者**，在趋势跟随环境中表现出较高效能，但在均值回归或横盘市场状态下，其预测能力则有所减弱。

6 消融研究

我们进一步探究智能体处理信息强度的机制，揭示上下文数量与决策质量之间存在非线性关系。

6.1 信息效用曲线

不同 Top-N （即框架中使用的头条新闻数量）设置下的结果（表3）对“信息越多越好”这一假设提出了质疑。相反，我们观察到信息效用呈现出明显不同的市场状态：

1. “最佳区间”与过滤失效。对于大多数表现优异的模型（例如 Gemini-2.5-Pro、Qwen3-Max），表现通常在**前5名**或**前10名**时达到峰值。这样的信息密度足以识别主导市场主题。然而，当上下文扩展到**Top-20**时，表现往往会下降（例如 DeepSeek-V3 的表现显著下滑）。定性检查表明，这是一种**“过滤失效”**：Top-20 榜单不可避免地包含娱乐八卦和非金融噪声。

2. Top-0 悖论：新闻何以误导。在 2025 年振荡阶段（图 B）中，一个引人注目的观察是，**前0名** 设置（纯价格历史）经常优于加入新闻的设置（例如任务B中的 GLM-4.6）。这表明信息价值依赖于市场状态。在缺乏趋势的市场中，金融新闻往往充斥着相互矛盾的分析师观点或“噪声”。接收这类相互矛盾的信息流的智能体可能会“产生幻觉”，编造出并不存在的叙事，从而导致过度交易。在这种市场状态下，仅依赖价格动量的“盲目”智能体（Top-0），比试图强行进行叙事拟合的智能体更为稳健。

6.2 “缩放定律”悖论

我们的数据集揭示了一个与模型规模有关的显著反常现象。虽然通用自然语言处理任务通常遵循严格的缩放定律（性能 $\propto$参数），但金融决策呈现出一种非平凡且依赖于市场状态的行为。

1. “知识优势”。在2024年阶段（图A），我们观察到了预期的层级差异：较大的模型显著优于较小的模型。例如，在任务B中，**Qwen3-Max-Think**（44.98%）远胜于**Qwen3-32B-Think**（19.27%）。

图A：2024全年（高波动性 / 政策转向）								
模型	任务A：宏观配置				任务B：行业轮动			
	前0名	前5名	前10名	Top-20	前0名	前5名	前10名	Top-20
推理模型								
DeepSeek-R1	5.93	7.38	7.47	8.56	26.32	21.66	28.37	23.90
Qwen3-Max-Think	10.27	11.46	8.98	10.03	29.10	**44.98**	**40.11**	**36.25**
Qwen3-32B-Think	**10.34**	2.77	5.56	6.99	28.82	19.27	21.67	24.02
标准模型								
DeepSeek-V3	5.47	**10.37**	9.28	3.93	13.91	14.16	6.97	15.41
Gemini-2.5-Pro	12.77	**15.35**	**10.67**	11.93	19.25	24.09	18.10	18.34
GPT-5	10.96	11.40	7.31	**15.75**	15.73	19.06	24.42	25.94
GLM-4.6	7.04	3.31	6.25	14.65	19.99	20.69	31.58	15.95
Qwen3-Max	**16.50**	12.77	8.03	14.20	34.61	33.85	26.50	32.03
Qwen3-32B	10.38	11.15	5.59	5.14	**35.88**	21.56	13.26	22.78

图 B：2025年上半年（振荡 / 净值变化较低）								
模型	任务A：宏观配置				任务B：行业轮动			
	前0名	前5名	前10名	Top-20	前0名	前5名	前10名	Top-20
推理模型								
DeepSeek-R1	7.47	**11.94**	**11.21**	6.20	7.51	5.55	6.08	5.32
Qwen3-Max-Think	4.32	8.36	6.37	6.99	10.42	3.90	2.84	6.17
Qwen3-32B-Think	6.16	9.28	10.49	7.07	7.20	6.72	5.78	7.71
标准模型								
DeepSeek-V3	5.34	7.86	7.41	6.00	7.79	4.21	5.58	2.52
Gemini-2.5-Pro	6.50	11.10	7.76	9.53	9.45	8.44	7.38	10.21
GPT-5	5.39	6.24	3.84	4.79	6.74	5.83	5.12	6.00
GLM-4.6	4.16	9.13	11.06	7.88	**16.86**	7.58	5.03	**18.00**
Qwen3-Max	6.47	7.14	9.48	8.13	9.09	**10.00**	**8.81**	10.01
Qwen3-32B	**13.64**	7.43	8.99	10.74	2.33	4.06	1.44	6.09

表 3：**Top-N 新闻上的消融结果（累积收益率 %）**。**粗体**表示同一 Top-K 设置下的最佳模型（列最佳）。下划线表示特定模型的最佳 Top-K 设置（行最佳）。

我们将此归因于**知识密度**以及潜在的**隐式泄漏**。“Max”模型可能在其参数中保留了更高保真度的世界知识。32B模型可能经过蒸馏压缩，因此遭受了**“知识压缩损失”**，遗漏了训练语料中包含的细粒度政策线索（例如改革的具体日期）。在这里，“记忆”有助于“推理”。为严格区分这些影响，我们在附录D中提供了实证记忆探测分析。

2. “能力陷阱”。关键的是，在2025年上半年阶段（图 B），这一层级关系发生逆转或瓦解。在任务A（前0名）中，规模较小的**Qwen3-32B**（13.64%）表现意外地优于规模庞大的**Qwen3-Max**（6.47%）。这表明**投资表现并不严格取决于模型能力**。在此前未见过、噪声强烈的震荡市场中，拥有超长上下文窗口和深度推理能力的大模型可能会陷入**“对噪声过度反应”**的陷阱——从随机波动中臆造复杂叙事。相比之下，小模型可能依赖更简单、鲁棒性更强的启发式方法（例如直接依据动量），因而在不确定市场环境中具有更好的泛化能力。这一发现对金融大语言模型领域“越大越好”的信条提出了挑战。这意味着，金融推理所需的不仅仅是原始模型能力，也凸显了我们的基准测试在检验超越单纯能力的鲁棒性方面的必要性。

7 结论

我们推出**CN-Buzz2Portfolio**，这是一个滚动式基准测试，通过将热门新闻映射到资产配置，评估语义理解与宏观层面金融决策之间的一致性。该开源数据集和框架将作为一种诊断工具，供

研究社区开发更可靠、以逻辑为驱动且可解释的金融智能体。

8 局限性

依赖市场状态的效能与能力对齐。本研究发现的一个主要局限是，在不同市场状况下，通用推理能力与金融稳健性之间存在明显分歧。目前的大语言模型（LLM）表现出显著的依赖市场状态的性能差异：在趋势跟随时期，它们善于识别结构性机遇；但在收益较低的横盘市场状态下，却难以从随机波动中辨别有意义的信号。这表明，实现金融对齐不能仅靠注入领域专属知识，还需要开发用于市场状态识别的自适应机制，避免模型在低信号情境中过度响应噪声。

时间粒度与信息延迟。我们的方法侧重于基于每日公众关注度进行宏观和行业层面的资产配置，其运作频率天然低于高频交易（HFT）系统。尽管我们验证了汇总的“热度”信号是中期资本流动的可靠指标，但这种方法无法捕捉日内价格动态或微观结构阿尔法。因此，如何弥合缓慢推理的语义分析与全面市场套利所需的毫秒级执行之间的延迟差距，仍是一项重大挑战。

市场摩擦与可扩展性限制。
本基准测试采用的模拟环境优先考虑战略推理，而非执行复杂性。由于假设收盘价处具有完美流动性，当前框架忽略了滑点、买卖价差和市场冲击等关键市场摩擦。虽然这些假设通常适用于小规模零售模拟，但它们限制了所观察到的策略直接扩展至机构投资组合的能力。CN-Buzz2Portfolio 旨在作为评估逻辑一致性的诊断工具，而非用于专业流动性管理的高保真引擎。

投资组合约束与市场完备性。本研究的资产范围仅限于仅做多的 ETF 工具，因此无法采用卖空或基于衍生品的对冲策略。尽管这种配置准确反映了中国市场大多数散户投资者面临的监管和实际限制，但这也限制了智能体在持续熊市周期中获取绝对收益的能力。未来版本的这一基准测试可纳入更复杂的工具，以评估智能体在多维且完备的市场环境中的表现。

伦理声明

研究目的与金融风险。 本研究所展示的数据集、基准和基线模型仅供**学术研究之用**。本文报告的模拟结果基于简化假设（例如每日收盘价、无限流动性），未考虑滑点、市场冲击或极端尾部风险等现实市场摩擦。因此，在 *CN-Buzz2Portfolio* 上取得优异表现并不保证实盘交易能够盈利。本研究不构成财务建议，对于任何尝试在真实市场中部署这些方法而遭受的金融损失，作者概不承担责任。

预期用途：将人工智能作为副驾驶。 我们主张将金融智能体部署为人类投资顾问的**辅助工具**（“人工智能副驾驶”），而非完全自主的“黑箱”交易者。我们所提出框架的主要价值在于能够处理海量信息流，并提出合乎逻辑的配置假设；这些假设始终应接受**人工监督** 和专业审查。鉴于我们的分析发现了幻觉和“能力陷阱”的潜在风险，我们提醒不要在缺乏监督的情况下将大语言模型用于资本管理。

数据与监管合规。 本基准测试使用的所有数据均源自公开可用的“热门榜单”和公开市场数据。未涉及任何私人用户信息或专有内幕数据。研究人员和从业者在采用本研究成果时，必须确保严格遵守各自司法管辖区的当地金融法规（例如关于算法交易和投资咨询的证券法）。

参考文献

智谱AI。2025年。GLM-4.6：先进的文本生成基础模型。 https://huggingface.co/zai-org/GLM-4.6。

Dogu Araci。2019年。Finbert：使用预训练语言模型进行金融市场情绪分析。*arXiv* 预印本 *arXiv:1908.10063*。

Andy Ayrey。2024年。Truth Terminal：一个用于模因工程的自主人工智能智能体。X（原 Twitter）个人主页：https://x.com/truth_terminal。访问日期：2025年2月15日。

Brad M Barber 和 Terrance Odean。2008年。浮华背后：注意力和新闻对个人投资者和机构投资者购买行为的影响。《金融研究评论》，21(2)：785–818。

Liyuan Chen、Shuoling Liu、Jiangpeng Yan、Xiaoyu Wang、Henglin Liu、Chuang Li、Kecheng Jiao、Jixuan Ying、Yang Veronica Liu、Qiang Yang 等人。2025a。借助基础模型推进金融工程：进展、应用与挑战。《工程》。

Yanxu Chen, Zijun Yao, Yantao Liu, Jin Ye, Jianing Yu, Lei Hou 和 Juanzi Li。2025b。Stockbench：大语言模型智能体能否在现实市场中盈利性地交易股票?

DeepSeek-AI。2024年。DeepSeek-V3 技术报告。*arXiv* 预印本 *arXiv:2412.19437*。

DeepSeek-AI. 2025年。Deepseek-r1：通过强化学习激励大语言模型的推理能力。*arXiv* 预印本 *arXiv:2501.12948*。Google Gemini 团队。2025年。Gemini 2.5：以先进推理、多模态、长上下文和新一代智能体能力推动前沿。Jian Guo 和 Heung-Yeung Shum。2025年。大型投资模型。《信息技术与电子工程前沿》，26(10):1771–1792。香港大学数据科学实验室。2025年。AI-Trader：多模型实盘交易竞赛平台。平台/竞赛。访问日期：2025年2月15日。Jincai Huang、Yongjun Xu、Qi Wang、Qi Cheems Wang、Xingxing Liang、Fei Wang、Zhao Zhang、Wei Wei、Boxuan Zhang、Libo Huang 等。2025年。基础模型与智能决策：进展、挑战与展望。《创新》。Yuzhen Huang、Yuzhuo Bai、Zhihao Zhu、Junlei Zhang、Jinghan Zhang、Tangjun Su、Junteng Liu、Chuancheng Lv、Yikai Zhang、Yao Fu 等。2023年。C-Eval：面向基础模型的多层次、多学科中文评测套件。《神经信息处理系统进展》，36:62991–63010。Changlun Li、Yao Shi、Yuyu Luo 和 Nan Tang。2025a。DeepFund：大语言模型能否胜任基金投资？来自实盘竞技场的视角。

Haohang Li、Yupeng Cao、Yangyang Yu、Shashidhar Reddy Javaji、Zhiyang Deng、Yueru He、Yuechen Jiang、Zining Zhu、Kp Subbalakshmi、Jimin Huang 等。2025b。InvestorBench：面向基于大语言模型的智能体金融决策任务的基准测试。第2509–2525页。

Haonan Li、Yixuan Zhang、Fajri Koto、Yifei Yang、Hai Zhao、Yeyun Gong、Nan Duan 和 Timothy Baldwin。2024a。CMMLU：衡量中文大规模多任务语言理解。载于 计算语言学协会：*ACL 2024* 论文集，第 11260–11285 页。

Xiang Li、Zhenyu Li、Chen Shi、Yong Xu、Qing Du、Mingkui Tan 和 Jun Huang。2024b。AlphaFin：基于检索增强型股票链框架的金融分析基准测试。载于*2024*年计算语言学、语言资源与评估联合国际会议论文集（*LREC-COLING 2024*），第 773–783 页。

Xiao-Yang Liu、Guoxuan Wang、HongyangYang 和 Daochen Zha。2023年。数据中心型 FinGPT：为金融大语言模型普及互联网规模数据。*NeurIPS* 指令微调与指令遵循研讨会。

NoF1 团队。2025年。nof1.ai：自主对冲基金与人工智能对战。在线平台。访问时间：2025年。

OpenAI。2025年。GPT-5 技术报告：推进通用人工智能。https://openai.com/index/introducing-gpt-5/。

Qwen 团队。2025年。Qwen-3 技术报告：扩展智能与推理。阿里巴巴集团。RockFlow 团队。2025年。RockFlow AI Arena：用于美国股票交易的自主智能体。在线平台。访问日期：2025年2月15日。Neng Wang、Hongyang Yang 和 Christina Dan Wang。2023年。FinGPT：面向金融数据集中的开源大语言模型的指令微调基准。*NeurIPS* 指令微调与指令遵循研讨会。Shijie Wu、Ozan Irsoy、Steven Lu、Vadim Dabravolski、Mark Dredze、Sebastian Gehrmann、Prabhanjan Kambadur、David Rosenberg 和 Gideon Mann。2023年。BloombergGPT：面向金融的大语言模型。*arXiv* 预印本 *arXiv:2303.17564*。Yijia Xiao、Edward Sun、Di Luo 和 Wei Wang。2024年。TradingAgents：多智能体大语言模型金融交易框架。*arXiv* 预印本 *arXiv:2412.20138*。Hongyang Yang、Xiao-Yang Liu 和 Christina Dan Wang。2023年。FinGPT：开源金融大语言模型。*IJCAI 2023 FinLLM* 研讨会。Hongyang Yang、Boyu Zhang、Neng Wang、Cheng Guo、Xiaoli Zhang、Likun Lin、Junlin Wang、Tianyu Zhou、Mao Guan、Runjia Zhang 和其他 1 位作者。2024年。Fin-robot：一个基于大语言模型、用于金融应用的开源人工智能智能体平台。*arXiv* 预印本 *arXiv:2405.14767*。

Haofei Yu、Fenghai Li 和 Jiaxuan You。2025年。Live- tradebench：利用大语言模型探寻现实世界中的阿尔法。

Yangyang Yu、Haohang Li、Zhi Chen、Yuechen Jiang、Yang Li、Jordan W Suchow、Denghui Zhang 和 Khaldoun Khashanah。2025b。Fin Mem：具备分层记忆和角色设计的性能增强型大语言模型交易智能体。*IEEETransactionsonBig Data*。

Boyu Zhang、Hongyang Yang 和 Xiao-Yang Liu。2023年。Instruct-fingpt：通过对通用大语言模型进行指令微调实现的金融情感分析。*IJCAI 2023 FinLLM* 研讨会。

Boyu Zhang、Hongyang Yang、Tianyu Zhou、Ali Babar 和 Xiao-Yang Liu。2023年b。通过检索增强型大型语言模型提升金融情感分析。*ACM人工智能金融国际会议（ICAIF）*。

A 数据集统计

表4 全面细分了 CN-Buzz2Portfolio数据集。为捕捉公众关注度的动态演变，我们的系统每天在各大金融平台上进行多次抓取。我们在以下两种设置下报告统计数据：**日内去重**（反映多次日常抓取中捕获的唯一新闻条目）以及**全局去重**（反映全新叙事进入趋势信息流的情况）。不同渠道中文本长度的差异（每条约900至1,900个字符）体现了信息密度的多样性，从简洁的标题到详细的政策解读，为大型语言模型推理提供了丰富的语义空间。

频道	日内去重		全局去重	
	数量	平均长度	数量	平均长度
财新	13,365	926.71	9,813	967.08
新浪财经	35,595	1,890.82	32,187	1,878.74
腾讯股票	29,284	1,720.21	25,104	1,699.11
天天基金	28,573	1,653.94	25,195	1,643.44

表4：不同金融新闻渠道的数据集统计信息。“Count”指排名前20的新闻条目数量。长度以字符数衡量。

B 资产范围详情

为确保模拟的散户环境具有代表性和充足的流动性，我们采用严格的筛选规则：**对于每个目标指数或行业，我们选择管理资产规模（AUM）最大的单只ETF联接基金**。这最大限度地降低了跟踪误差，也反映了寻求流动性和安全性的理性散户投资者最可能做出的选择。表5和表6 列出了我们实验中使用的具体资产。

C 案例研究：推理轨迹与失败分析

为评估定性叙事与投资逻辑之间的一致性，我们分析了 DeepSeek-R1 的推理轨迹。这些案例展示了该智能体在高噪声趋势环境中的能力与局限性。

C.1 成功案例：战略对冲（2025年4月）

背景： 全球市场对出乎意料的关税公告作出反应，引发了广泛的风险规避情绪。

类别	Code	资产名称	经济代理角色与语义范围
股票	000300.SH 000905.SH 399006.SZ 000688.SH	沪深300指数 中证500指数 创业板 科创 50	蓝筹股：代表中国核心经济（金融业、消费）。可作为“整体市场贝塔”的代理指标。 中盘成长：代表制造业和次级增长驱动力。 创新：聚焦深圳的高科技初创企业（医疗保健、新能源）。高波动性。 硬科技：“国家战略科技”（半导体、生物技术）及研发强度的代理指标。
周期性	000932.SH 000941.SH 399971.SZ 000819.SH 000928.SH	消费 新能源 传媒 有色金属 能源	内需：追踪必需消费和可选消费，是“零售复苏”叙事的代理指标。 绿色转型：涵盖光伏、风电和电动汽车电池，受“碳中和”政策影响。 数字经济：涵盖游戏、人工智能应用和知识产权。对监管和人工智能趋势高度敏感。 工业大宗商品：铜、铝、锂。与全球制造业周期相关。 传统能源：煤炭、石油、天然气。是“能源安全”和通胀交易的代理指标。
Safe	000012.SH 518880.SH	政府债券 黄金ETF	无风险锚：10年期国债。在经济衰退期间属于防御性资产。 通胀对冲：实物黄金。作为“全球不确定性”的代理，并用于货币对冲。

表5：TaskA（宏观/主题）的资产范围。这些资产使智能体能够表达对经济增长、通胀和战略政策方向的看法。

类别	Code	资产名称	经济代理角色与语义范围
金融	512880.SH 512800.SH 512070.SH	证券 银行 保险	市场贝塔：对市场情绪高度敏感。“牛市旗手。” 价值防御：高股息收益率。系统性稳定和国企改革的代理指标。 长期利率：收益率上升和人口趋势的受益者。
Tech	159995.SZ 159819.SZ 515880.SH 159852.SZ	半导体 AI 通信设备 软件	科技自主权：芯片、集成电路。“自主”叙事的关键。 趋势：算力、算法。全球“人工智能热潮”的代表。 基础设施：5G/6G、数据中心。“新型基础设施”支出的代表。 数字化：软件即服务、操作系统。“数据作为生产要素”的代理指标。
健康	512010.SH 512170.SH 159992.SZ	生物医药 医疗保健 创新药	创新：创新药、医药外包服务。受“人口老龄化”政策影响。 服务：医院、消费医疗。 研发重点：专注于创新药物的纯粹投资标的。风险与回报均较高。
周期性	515170.SH 512690.SH 515220.SH 512200.SH 159870.SZ	食品与饮料 白酒 Coal 房地产 化学品	必需消费品：加工食品、乳制品。防御性消费，现金流稳定。 高端：白酒。商业活动和财富效应的代理指标。 股息：现金牛能源。在波动期间具有防御性。 政策转向：开发商。对“宽松/紧缩”信贷政策高度敏感。 上游：原材料价格。与生产者价格指数相关。

表6：任务B（行业轮动）的资产范围。这种细粒度选择用于测试智能体能否根据新闻细微差异，区分相似行业（例如制药与医疗保健）。

推理摘录： “由于关税政策，市场风险厌恶情绪加剧。全球股票正在下跌，尤其是科技和成长板块。沪深*300*指数和科创 *50* 的市场情绪偏负面。黄金 *ETF* 和债券充当避险资产。我们当前的投资组合股票敞口较高（*~¥37,000*），面临系统性风险。决策：减少沪深*300*指数的配置；增加黄金*ETF*以进行对冲。”
结果： 模型在趋势加剧之前，将资本轮动至避险资产，成功降低了回撤。

C.2 失败案例：政策时机偏差（2024年9月至10月）

背景： “924”刺激政策引发了一轮历史性上涨，随后在假期后出现大幅回调。

推理摘录（9月30日）： “创业板注意力高，并出现超买信号。操作：持有。科创 *50* 代表硬科技，获得强有力的政策支持。操作：买入 *¥5,000*。受刺激政策叙事影响，整体市场情绪极度看涨。”

分析： 尽管模型识别出了技术面的“超买”信号，却任由政策驱动的乐观情绪（叙事偏差）压过风险警惕。它未能预见10月8日节后回调的速度，导致了大幅回撤。

关键见解： 大语言模型智能体表现出**“持续性偏差”**，过度依赖当前强势叙事，并低估均值回归风险。未来的设计可以采用混合架构：利用大语言模型进行定性行业选择（阿尔法），并通过统计模块（例如均值—方差优化）来控制风险敞口规模（贝塔）。

D 一种系统性数据泄漏的诊断方法

D.1 动机：普遍预训练时代的基准测试

在大规模预训练时代，数据泄漏（历史污染）已成为所有静态基准测试面临的系统性挑战。与其

将潜在信息泄漏视为致命缺陷，我们认为，稳健的金融基准测试可以作为一种**分析工具**，用于区分历史记忆与主动语义推理。

D.2 记忆探测方法

为量化模型记忆的边界，我们设计了一项记忆探测实验，采用两个极端案例：**(1)沪深300指数**，代表高曝光度的“共识记忆”；以及 **(2) ETF159852.SZ**，代表“未见数据”的小众资产。我们在2024年随机选取的100个日期上评估模型，且不提供新闻背景，并采用两项指标：

- **趋势准确率**：预测相邻两个交易日期之间相对价格变动（上涨/下跌）的二元准确率。
- **价格准确率**（±1%/3%/10%）：模型预测落在实际市场收盘价指定容差区间内的百分比。

D.3 结果：记忆并非万能解法

如表 7所示，模型确实对热门的沪深300指数表现出中等程度的记忆。然而，这种“先验知识”并未使模型在我们的主要任务中取得完美表现。

模型	趋势准确率	价格准确率（±1%/3%/10%）
图*A*：热门资产（沪深*300*指数）		
GPT-5	0.46	0.24 / 0.57 / 0.76
DeepSeek-R1	0.74	0.20 / 0.44 / 0.73
Qwen3-Max	0.32	0.06 / 0.13 / 0.67
Qwen3-32B	0.60	0.05 / 0.13 / 0.37
图 *B*：冷门资产（*ETF 159852.SZ*）		
GPT-5	0.20	– / – / 0.10
DeepSeek-R1	0.64	– / – / 0.00
Qwen3-Max	0.36	– / – / 0.00
Qwen3-32B	0.54	– / – / 0.00

表7：记忆探测结果。ETF159852代表不太可能出现在预训练数据中的资产。

D.4 分析：语义推理的首要性

我们的分析揭示了一个关键的**“逻辑—结果错配”**：即使是对沪深300指数具有较强趋势记忆的模型，也经常无法完成我们的任务B（行业轮动）。这表明，知道“结果”（价格上涨）并不能帮助模型解决“过程”（为什么这条新闻能够证明应进行这样的行业配置）。

我们框架的战略价值：

1. **以滚动更新进行缓解：**我们的滚动时域从2024年至2025年，确保模型接触到各种“已记忆”和“未见过”的市场状态，促使它们依赖通过我们的三阶段*CPA*工作流程提炼出的可泛化逻辑。
2. **基准测试作为诊断工具：** 通过分析存在潜在信息泄漏的模型仍在哪些方面失败，研究人员可以识别出记忆无法解决的具体推理瓶颈。

总之，*CN-Buzz2Portfolio*实现了从“黑箱测试”到“诊断分析”的方法论转变，为在数据污染普遍存在的世界中评估金融智能体提供了一条可行路径。