LLM 多智能体投资系统的架构选择与校准优先评估

——证据综述、实验记录与研究计划

Sizhuo Long

openInvest 项目（github.com/longsizhuo/openInvest）

2026 年 6 月

摘要：针对“多智能体大语言模型（LLM）投资系统的智能体数量与角色划分是否存在最优配置”这一问题，本文完成三项工作。其一，系统梳理 2024–2026 年间的受控实验证据，表明：主流多智能体金融框架均未对角色配置做受控消融，其角色划分源于对真实机构的类比而非实证优化；等算力条件下，归因于角色扮演的增益主要来自多次采样与聚合；多智能体辩论使智能体信念轨迹构成鞅（martingale），期望正确率不因辩论提升，且存在损害概率校准的机制。其二，报告 openInvest 系统已完成的实验：基于 2005 年至今 1579 个零前视重算点的全历史诊断发现条件概率分布存在结构性覆盖不足；经预注册拟合/样本外验收的校准层（小样本收缩 k=80、带宽扩张 γ=1.1）使样本外 30/60/90 天窗口覆盖率升至 81%/80%/76%，Brier 分数全窗改善。其三，提出分层目标框架——优化层以 Brier 与覆盖率拟合、决策层以确定性规则映射、监控层以配对损益归因——并给出五阶段预注册架构实验与黄金单资产 MVP 计划。结论指出：不存在被实证确认的最优角色配置；可检验的设计原则为信息源分解、独立样本聚合与确定性护栏；以 LLM 担任最终概率聚合者是现行架构中证据支持最弱的环节。

关键词：多智能体系统；大语言模型；概率校准；预测聚合；量化投资；消融研究

Abstract: This paper asks whether an optimal agent count or role division exists for LLM-based multi-agent investment systems. We (i) survey controlled evidence from 2024–2026, showing that no published financial multi-agent framework has ablated its role configuration, that gains attributed to role-play largely reduce to sampling plus aggregation under equal compute, and that multi-agent debate forms a martingale over beliefs while risking calibration harm; (ii) report experiments on openInvest, where a pre-registered fit/out-of-sample calibration layer (shrinkage k=80, band expansion γ=1.1) raised OOS coverage to 81%/80%/76% over 30/60/90-day windows with Brier improvements, across 1,579 zero-lookahead recomputation points since 2005; and (iii) propose a layered objective framework, a five-stage pre-registered experimental plan, and a single-asset (gold) MVP. No empirically optimal role configuration exists; the defensible design principles are information-source decomposition, independent-sample aggregation, and deterministic guardrails.

Keywords: multi-agent systems; large language models; probability calibration; forecast aggregation; quantitative investing

【增订版说明】本版在原文基础上新增图 *1*–图 *7* 及对应论述，并对 *5.1*、*5.2*、*5.4*、*5.5* 节作了数据化扩写。其中图 *3*–图 *7* 由复现脚本 *make_figures.py* 直接重算自归档数据 *path_review.jsonl*（*n=1579*）、委员会决策记录与 *TA* 复现包，未引入任何新的建模假设；图 *1*–图 *2* 分别为三机制示意与可验证性解析关系图。所有图内数字与正文及表 *1*–表 *3* 互校一致。

1 引言

1.1 问题提出

基于大语言模型的多智能体金融决策系统近两年快速增长。TradingAgents [1] 以七类角色（基本面、情绪、新闻、技术四类分析师，多空研究员，交易员，风控团队）模拟真实交易公司的组织结构，是该方向的代表性工作；openInvest 采用四角色纵向投资委员会（Macro Strategist / Quant / Risk Officer / CIO），单轮汇总、无辩论，并在 LLM 层之外附加确定性后处理与独立防御组件。两种配置共享同一个

未经检验的前提：角色划分本身有效。前者的七角色来自对交易公司部门设置的类比，后者的四角色来自设计者的经验直觉，均未经受控实验筛选。

1.2 研究问题

RQ1：智能体数量与角色划分方式是否存在可被实证支持的最优解？

RQ2：若不存在唯一最优解，架构设计应依据何种可检验的原则？

RQ3：在校准优先目标下，openInvest 各组件分别处于何种证据地位，应如何修订？

RQ4：校准目标与盈利目标之间的关系应如何界定，使二者既不混淆也不割裂？

1.3 评估口径

本文的主要目标为预测校准质量，即系统输出的概率本身是否可信，以 Brier 分数与概率带覆盖率（coverage）度量；回测收益与 Sharpe 比率为次要目标。两类口径存在本质差异：收益类指标评估整条决策链的最终结果，包含仓位管理与风控规则等与预测无关的成分；校准类指标只评估系统所述概率与实际频率的一致性，两者可严重背离。该口径与已发表框架不同——TradingAgents、FinCon 等均以收益类指标评估，未报告任何校准指标。约束条件为：无 LLM 调用成本约束；辩论机制的接受标准为无减益即可，不要求显著增益。

1.4 本文贡献

（1）以校准优先口径系统整理多智能体架构的受控证据；（2）报告一组经预注册样本外验收的校准实验，附可复现的数据归档与验收流程；（3）提出三机制分解与分层目标框架，回答 RQ4；（4）给出五阶段架构实验与黄金单资产 MVP 的预注册计划。

2 相关工作与证据综述

2.1 多智能体金融框架的实证基础

TradingAgents（arXiv 2412.20138，v7）的协作结构为：四类分析师并行产出报告，多空研究员辩论，交易员综合决策，风控团队复核。其实证主张建立在如下回测之上：窗口为 2024-01-01 至 2024-03-29（约三个月），标的为 AAPL、GOOGL、AMZN 三只股票，对照组仅为五个规则策略（Buy&Hold、MACD、KDJ+RSI、ZMR、SMA），指标限于累计收益、年化收益、Sharpe 与最大回撤；论文报告 Sharpe 达 8.21/6.39/5.60，并在脚注中自行说明该数值超出经验合理范围。与本文直接相关的事实是：该论文不含任何关于角色数量或划分方式的消融实验，不含单 LLM 基线对照，不含任何校准指标；辩论轮数仅作为代码参数存在。FinCon [16]、AlphaAgents [15]、FinMem、FinAgent 等框架同样以“模拟真实机构分工”作类比论证其角色配置，未见角色层面的受控消融。各框架实证基础对比见表 1。

表 **1** 主要多智能体金融框架的实证基础对比

框架	角色数	协作机制	评估指标	角色消融	校准指标
TradingAgents [1]	7	并行分析＋多空辩论＋流水线	收益类（CR/AR/Sharpe/MDD）	无	无
FinCon [16]	多角色（经理—分析师层级）	层级化语言强化	收益类	无	无

AlphaAgents [15]	3	协作评估	收益类	无	无
openInvest（本文）	4	纵向单轮汇总	校准类为主（Brier/覆盖率）	计划中（7.1 节）	有（OOS 验证）

2.2 等算力对照下的单智能体与多智能体

四项独立研究在控制计算预算条件下的结论方向一致。Wang 等 [2]（ACL 2024）发现，配备强结构化提示的单智能体可匹配最优的多智能体讨论方法，多智能体仅在提示中不含示例时占优。Smit 等 [3]（ICML 2024）发现多智能体辩论（MAD）不能可靠地优于 self-consistency 与集成等更简单的提示策略。Zhang 等 [4] 在九个基准、四个模型上复核，结论为 MAD 即使消耗更多推理算力，也无法可靠超越单智能体基线。Tran 与 Kiela [5] 基于数据处理不等式论证：在固定推理 token 预算且上下文被充分利用时，单智能体在信息效率上占优；多智能体仅在单智能体的有效上下文利用退化时才具备竞争力。小结：受控条件下，归因于角色扮演的增益主要还原为“多次采样＋聚合”的集成效应，该效应不依赖角色划分。

2.3 辩论机制的鞅性质与校准风险

Choi 等 [6]（NeurIPS 2025）在七个基准上分解 MAD 的增益来源，发现多数投票单独即可解释通常归因于辩论的大部分性能增益，并在理论上证明辩论使智能体的信念轨迹构成鞅：设第 t 轮信念为 B_t，则 E[B_{t+1} | F_t] = B_t，期望正确率不随辩论轮次提升；只有向纠正方向施加偏置的干预（如 oracle-locking、置信度加权）能打破该中性。与此同时，多项工作记录了辩论的系统性失效模式：谄媚、置信度模仿、从众、过早收敛，以及被论断性而非证据说服 [10–12]，其共同方向是推动群体走向自信的共识。由于 Brier 与覆盖率恰好惩罚过度自信，辩论与校准优先目标存在方向性冲突；该风险在收益口径下可能被掩盖，在校准口径下直接计入损失。

2.4 言语化概率与角色人设

Xiong 等 [7]（ICLR 2024）对 LLM 言语化置信度的实证结果为：模型输出的置信度高度聚集于 80%–100% 区间且呈 5 的倍数分布；GPT-3/GPT-3.5/Vicuna 的平均期望校准误差（ECE）超过 0.377；GPT-4 区分自身对错的 AUROC 仅 62.7%，接近随机水平。该工作的人设对照实验（“自信的”与“谨慎的”系统提示）显示两组在准确率与校准上的差异极小。该结果直接削弱了“通过角色人设改善概率质量”这一路径的依据，并佐证在 LLM 下游设置确定性校准层的必要性——原生言语化概率不具备可直接使用的校准质量。

2.5 预测聚合文献的可迁移结论

Schoenegger 等 [8]（Science Advances, 2024）令 12 个 LLM 组成集成，在 31 个二元问题、为期三个月的预测锦标赛中与 925 名人类预测者对照，集成预测与人类群体在统计上不可区分。Halawi 等 [9]（NeurIPS 2024）对同一问题多次提示，以截尾均值（trimmed mean）聚合概率，以 Brier 分数评估，平均接近、部分场景超越人类竞争性预测群体，提供了“多次采样＋确定性聚合”的已验证模板。经典超级预测文献 [21] 关于极化聚合（extremizing）的条件性结论为：仅当成员持有相互独立、不重叠的信息时有效；成员共享信息或经过讨论（即辩论的作用方式）时失效甚至有害。该结论为聚合方式与成员信息结构的匹配给出了精确规则。

2.6 团队规模的弱先验

两项邻域辩论消融提供了关于规模的方向性信号:MADIAVE [13] 在固定调用预算下发现智能体数从 2 增至 3–5 反而降低准确率, 2 智能体、2–3 轮为最优;PD3 [14] 发现 3 智能体、3 轮之后因误差传播与上下文过载出现收益递减。两者均非金融场景, 仅作弱先验, 但方向一致:反对向七角色扩张, 温和支持小型委员会。

3 openInvest 系统与实验设施

3.1 架构与输出

系统采用四角色纵向分工:Macro Strategist(宏观环境)、Quant(量化与技术信号)、Risk Officer(风险约束)、CIO(综合裁决);协作机制为单轮、无辩论的报告汇总。核心输出为带概率的路径预测:30/60/90 天窗口的涨跌概率、P10/P90 概率带、路径形状分类(先跌后涨、直接涨、冲高回落、一路收跌), 并配有事后对账闭环。

3.2 确定性组件

LLM 层之外包含四类确定性组件:Sanity Check 后处理, 对 LLM 裁决施加规则覆盖;独立快崩防御层(INDEP_DEFENSE_FLAG), 基于 VIX 分位等哨兵信号对买入侧实施确定性拦截;偿付能力兜底规则(SOLVENCY), 限制特定条件下的减仓;摩擦成本模块。

3.3 数据与实验设施

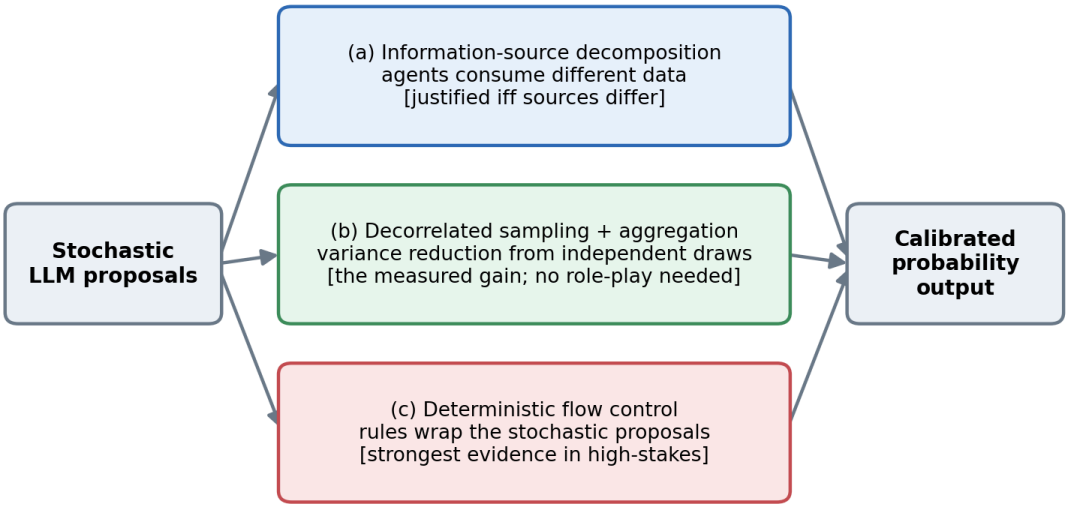
walk-forward 重算框架覆盖 2005 年至今, 产生 n=1579 个零前视重算点;配套预注册拟合/样本外(fit/OOS)验收流程。研究归档包含 path_review 1579 条、verdict_review 276 条、事件记录 1676 条、委员会决策记录 517 份(2026-04-28 至 06-12)与每日报告 82 份;单元测试 657 项。防自污染设计:重算快照永久保存未经校准的原始分布, 校准仅在生产出口应用, 保证未来重新拟合不会摄入系统自身的输出。

4 理论框架

4.1 多智能体价值的三机制分解

综合第 2 节证据, 任意多智能体架构的潜在价值可分解为三种可独立检验的机制, “几个角色”的设问混淆了三者。机制(a)信息源分解:不同智能体消费不同数据, 避免单一上下文过载。由文献 [5] 的条件可知, 这是多智能体相对单智能体的信息论正当性来源:当单智能体的有效上下文利用退化时, 分解才产生增益;对同时摄入宏观、基本面、技术面与新闻数据的系统, 这是设立多个智能体的最强理由。机制(b)去相关采样与聚合:集成降方差效应。2.2 与 2.5 节表明这是受控实验中测得增益的主要来源;该效应是采样与聚合的性质, 单模型多次采样即可获得, 不依赖角色扮演。机制(c)确定性流程控制:以确定性规则包裹随机的 LLM 提案, 对应护栏与神经符号方向的文献 [17–19], 是高风险场景下证据支持最强的模式。角色人设本身不属于上述任何机制:当且仅当角色映射到不同信息源时, 它才落入机制(a);若多个角色是在同一上下文上叠加不同人设, 依据 2.4 节证据, 其增益有限且引入置信度扭曲风险。

Fig. 1 Three independently testable mechanisms behind 'multi-agent value'



Role persona is NOT a fourth mechanism: it counts only when a role maps to a distinct information source (a).

图 1 多智能体价值的三机制分解：信息源分解(a)、去相关采样＋聚合(b)、确定性流程控制(c)。角色人设本身不构成第四种机制，仅当映射到独立信息源时才计入(a)。

4.2 分层目标框架

为回答 RQ4，本文将系统目标按层分置，各层使用不同口径且互不越权。

优化层（参数拟合）：仅使用 Brier 与覆盖率。理由有三。其一，估计方差：Sharpe 将整段历史压缩为单一统计量，其标准误近似 $\sqrt{((1+SR^2/2)/T)}$ [23]；以十年数据估计 SR=0.5 的策略，标准误约 0.34，与估计值同量级；重叠窗口下有效样本进一步缩水（本系统概率表对此显式标注，如某资产 uptrend 桶 n=1423、独立样本约 47），对此类统计量直接调参即构成回测过拟合 [24]。其二，可诊断性：Brier 可按 regime、窗口、资产分桶定位失效来源；Sharpe 恶化无法区分预测错误、仓位错误与运气。其三，不可操纵性：严格 proper 评分规则下报告真实信念是期望得分的唯一最优策略 [25]，而 Sharpe 可被卖尾部风险类行为在样本内推高。

决策层（概率到动作）：由校准概率经确定性规则映射为仓位与拦截动作。该层参数极少（阈值与 Kelly 分数量级），自由度小，可用钱口径（回撤约束、反事实账）做粗验收而不构成过拟合。拟合容量留在样本充足的概率层，钱口径留给近乎无参数的规则层。

监控归因层：损益为输出而非损失函数。三臂对照中的关键读数为配对差值（完整系统减纯确定性策略），共同市场噪声相消，统计功效高于绝对收益检验。应予声明：LLM 对历史时段的记忆构成训练数据前视污染，使 LLM 臂的历史读数系统性偏乐观；由此得到不对称推断逻辑——若在该有利偏置下 LLM 臂仍不优于确定性臂，则“委员会对收益无增量”的结论成立强度更高；反之则需进一步实验确认。

4.3 从校准到收益：可验证性分析

收益声明的统计可验证性受 $t \approx SR\times\sqrt{T}$ 约束：SR=0.5 的策略需约 16 年方能达到 $t\approx 2$，SR=1 亦需约 4 年。故在任何 MVP 时间尺度上，“系统盈利”与“系统运气好”在数据上不可区分，而校准声明（覆盖率、Brier）当下即可验证。主动管理基本定律 $IR \approx IC\times\sqrt{breadth}$ [22] 表明：系统可控变量为预测质量（IC，对应校准）与避免负 IC 行为（对应纪律层），下注广度（breadth）为第三变量。就价值来源而言，按可信度排

序为:防大回撤(复利不对称, 下跌 50% 需上涨 100% 回本)、行为纪律(投资者行为差约每年 1 个百分点量级 [27])、成本控制、预测 alpha(已发表框架均未在长窗口、含成本条件下证明净 alpha)。校准概率同时是仓位公式(Kelly 类 [26])的合法输入;以未校准的过度自信概率确定仓位, 将系统性放大仓位误差。

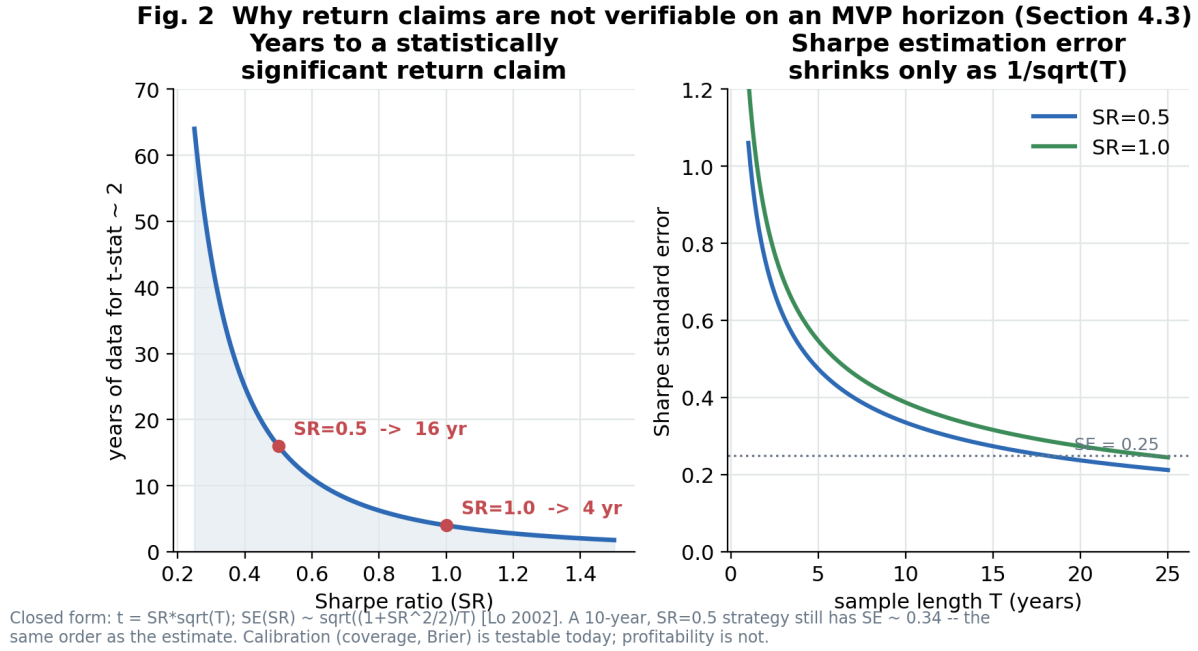


图 2 收益类声明在 MVP 时间尺度上的统计不可验证性(第 4.3 节)。左:达到 t≈2 所需年数随 Sharpe 变化(t=SR·√T);右:Sharpe 标准误随样本年数仅以 1/√T 收缩 [Lo 2002]。

5 实验记录

5.1 Regime 锁定层消融

对 regime-lock 层的消融显示, 该层在绝大多数决策点为恒等操作(no-op), 对最终裁决的边际贡献接近零。该结果将后续工作重心从规则层转向条件概率表本身的质量。

该消融基于 2026-06-04 的成对重放:在 GC=F 与 NDQ.AX 的周频决策日上, 分别运行开启 uptrend 方向锁的 A 链(n=912)与生产用无锁配置的 B 链(n=912)。两链裁决分布的差异集中于仓位档位而非方向——A 链将更多决策推向 ACCUMULATE(348 比 234, 约 +12.5 个百分点), HOLD 由 648 降至 506;而 30 天命中率在两链间为 57.2% 对 55.7%, 其差小于重复 draws 之间 35%–45% 的裁决翻转噪声。据此该层被定性为"杠杆旋钮":它平移仓位激进度, 却不改变方向判别的期望正确率, 对校准目标无可测增量。这一发现将后续工作重心由规则层转向条件概率表本身的质量。

5.2 全历史扩窗校准诊断

将诊断窗口扩展至全历史(n=1579;时代分桶覆盖 2007–2026, 含 2008 危机、2013–2015 金熊、2020 疫情、2022 加息等七个时代)后:整体概率带覆盖率为 69%–76%(此前单窗诊断为 56%–66%), 中位偏差约为零, 表明早先观察到的"系统性偏保守"属窗口效应。但时代分桶显示七个时代中六个覆盖率低于 80%;downtrend regime 整体覆盖率仅 54%, 最差的时代×regime 桶为 GC=F 2007–2009 downtrend, 覆盖率仅 7%(独立样本为个位数), 即条件分布在样本不足时呈现虚假精确。p_below 的 Brier 在多数时代劣

于无条件基率(2013–2015 金熊时段为 0.356 对 0.228)。结论:覆盖不足为结构性问题而非单窗效应,据此进入校准层开发。

上述诊断可由三组重算图直接读出(均重算自 path_review.jsonl, n=1579, 零前视)。图 3 以时代×regime 网格呈现 90 天覆盖率:downtrend 列系统性偏低(整体约 54%),最差桶 GC=F 2007–2009 downtrend 仅 7%(n=14)。图 4 将该现象归因于样本量——条件带宽随有效样本数几乎不变(中位约 9%),而覆盖率在 effective_n=1 时跌至 42%,即模型在稀薄数据上保持同等自信,构成伪精确。图 5 比较条件概率表与同资产无条件分布的 p_below Brier:在所考察的七个时代中,条件表 Brier 均不优于同资产无条件分布(金熊时段条件表 0.356、无条件 0.333;此处基线取逐点的同资产无条件分布,即收缩目标本身,与本节上文相对全局基率的 0.228 口径不同),直接为"向同资产无条件分布收缩"这一校准动机提供了证据。

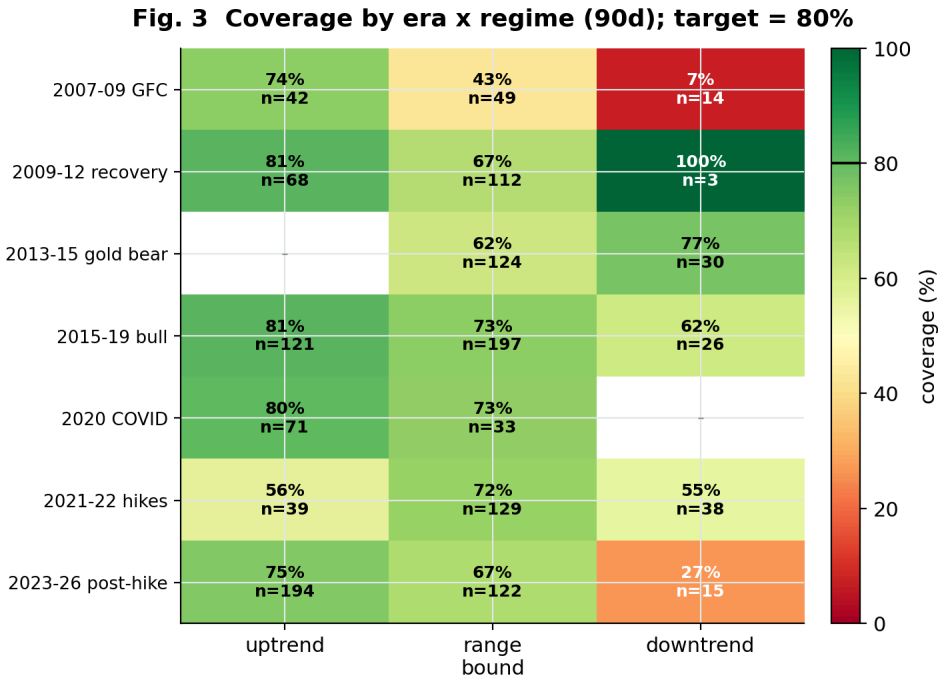
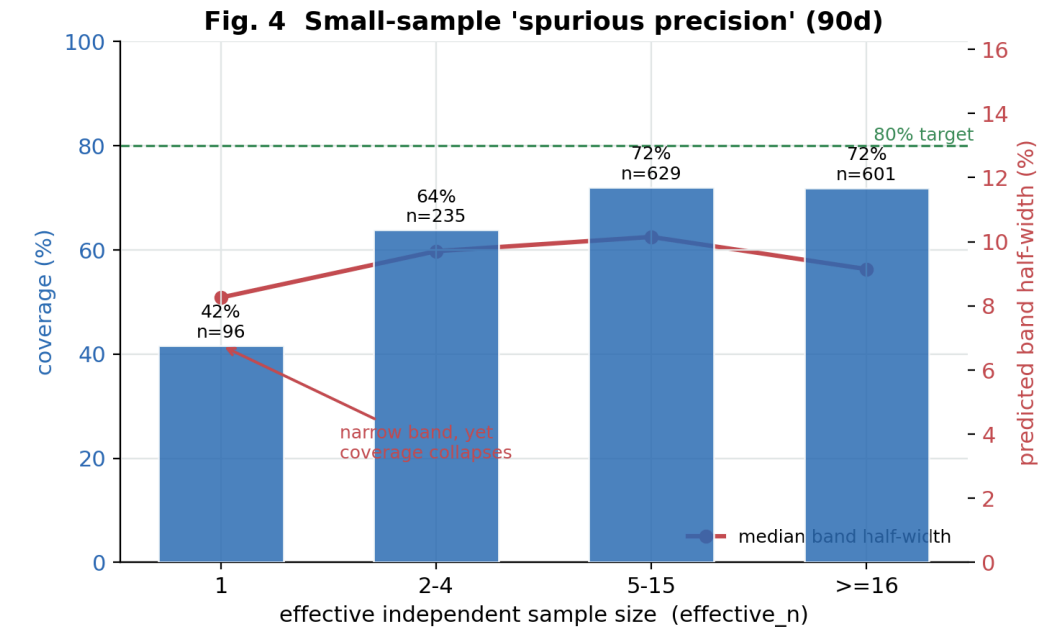
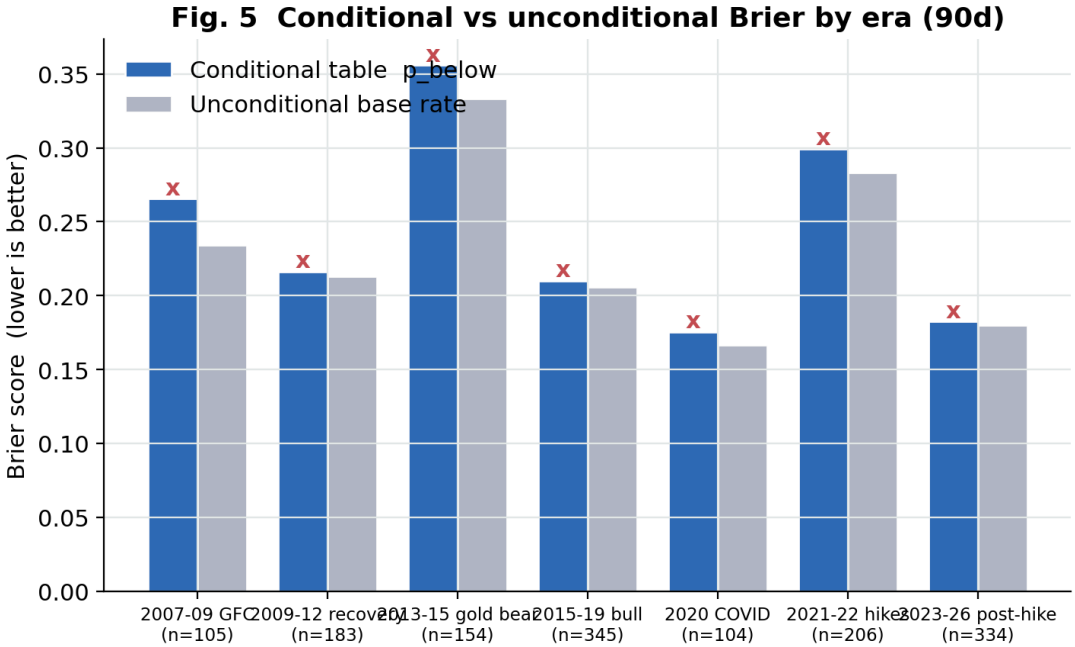


图 3 按时代×regime 的 90 天概率带覆盖率(重算自 path_review.jsonl)。downtrend 列系统性偏低,最差桶 GC=F 2007–2009 downtrend 仅 7%(n=14),系小样本下的伪精确。



The conditional band stays ~9% wide regardless of sample size, but coverage falls to 42% at effective_n=1: the distribution is equally confident on thin data. Reproduced from path_review.jsonl.

图 4 “伪精确”现象(90 天窗口):条件带宽随有效样本数几乎不变(中位约 9%),但覆盖率在 effective_n=1 时跌至 42%——模型在稀薄数据上保持同等自信。



Red x marks eras where the conditional table is WORSE than the naive same-asset base rate -- the motivation for shrinkage toward the unconditional prior. Reproduced from path_review.jsonl.

图 5 各时代条件概率表 p_below 的 Brier 与同资产无条件分布对比(90 天)。红叉标记条件表劣于无条件分布的时代;本图所考察七个时代全部如此,构成“向无条件分布收缩”的校准动机。

5.3 校准层与预注册样本外验收

校准机制包含两项统计修正：小样本收缩，λ = eff_n/(eff_n+k)，将条件分布向同资产无条件分布收缩；带宽扩张，对 P10/P90 与下行分位乘以系数 γ（不作用于概率字段）。以 2007–2017 拟合、2018–2026 纯样本外验证，结果见表 2。三窗覆盖率均进入预注册目标带 [75%, 85%] 且 Brier 全窗改善，验收通过，k=80、γ=1.1 写入默认配置。生产端的即时效应为：GC=F downtrend 最差桶 p_below 由 35% 升至 41%，悲观分位由 −2.3% 修正至 −3.0%，方向与诊断一致。应予声明：校准后 p_below 与事后基率的差距收窄至 0.5–1.5 个百分点但未反超；事后基率不可预知，仅作参考线。

表 **2** 校准层预注册拟合/样本外验收结果（拟合 **2007–2017**，样本外 **2018–2026**）

窗口	**OOS** 覆盖率（未校准 → 校准）	**OOS Brier**（未校准 → 校准）	目标带 **[75,85]**
30 天	76% → 81%	0.236 → 0.233	达成
60 天	73% → 80%	0.237 → 0.230	达成
90 天	70% → 76%	0.215 → 0.209	达成

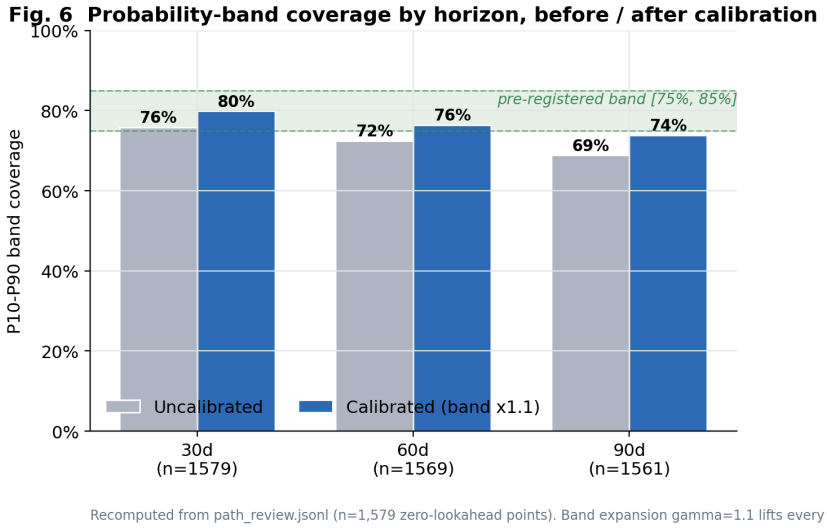


图 6 按预测窗口的概率带覆盖率：未校准 vs 仅施加带宽扩张 γ=1.1（全历史，n=1579）。（口径说明见下文；表 2 为更严格的样本外验收口径。）

图 6 以全历史口径直观呈现带宽扩张的方向性效应：仅施加 γ=1.1 即把 30/60/90 天覆盖率由 76%/72%/69% 抬升至 80%/76%/74%。需说明该图为全历史诊断口径，与表 2 报告的拟合（2007–2017）/样本外（2018–2026）验收口径不同——后者叠加了小样本收缩 k=80，故样本外数值（81%/80%/76%）更高；图 6 仅用于隔离展示带宽扩张单一因子对覆盖率的影响。

5.4 技术面信号实验复现包

另一独立实验线（TA 实验）已固化为标准复现包：2540 条 Phase A 报告按“窗口_模型”规范命名，数据、输入缓存与评估脚本三层分置，评估门可在不重新调用 LLM 的条件下直接复算。复现包同时声明已知局限：仅 2 资产 × 2 窗口、存在前视风险、规则基线未对标经典因子。含真实财务上下文的委员会原始记录已与公开复现数据分离，归档于私有仓库。

该实验的核心读数为预注册命中率门：对每个分析师，要求其 30 天方向命中率的 95% 置信区间下界超过“同子集方向基率”与“确定性机械映射基线”二者的较大值，方判 PASS。在 732 条 STANCE 报告（覆盖 GC=F 与 NDQ.AX、2022 与 2024 两窗）上，三个分析师的 30 天命中率分别为 fundamental 50.3%（CI [42.4%, 58.3%]）、news 57.4%（CI [49.4%, 65.1%]）、sentiment 64.7%（CI [57.0%, 71.8%]），而对应方向基率约 71%–73%。三者 CI 下界均显著低于基率，全部判定为 FAIL，据预注册流程不再进入 Phase B。该结果与本文主张方向一致：在本系统的信息结构下，单独的分析师人设未能在方向判别上超越朴素基率，增益（若有）须来自独立样本的聚合而非人设本身。

5.5 生产运行案例观察

一次决策日运行（2026-06-12）提供了确定性层行为的案例：VIX 为 20.7，处于近两年 85% 分位，独立防御标志触发；委员会对两项资产给出 ACCUMULATE ¥3,000 的裁决，均被防御层确定性降级为 HOLD；对黄金的集中度减仓提议被偿付能力兜底规则拦截。该案例表明确定性层按设计覆盖 LLM 裁决；其长期净效应需依赖反事实记账积累样本后方可评估（7.2 节）。该次运行同时暴露了报告层问题：经样本外验证的概率表在报告中处于次要位置，而 LLM 自报置信度（CONFIDENCE 字段）作为主要数字呈现——依据 2.4 节证据，后者缺乏校准依据，已列入修订计划。

图 7 给出 CONFIDENCE 字段在 2100 次委员会决策上的分布：取值高度聚集于 0.55/0.60/0.65 等 0.05 的整数倍，中位数 0.60，复现了 Xiong 等 [7] 记录的整数化伪精度。该形态本身即说明此原生数值缺乏校准依据，不宜作为对外呈现的主数字；以经样本外验收的概率表替代之，已列入修订计划。

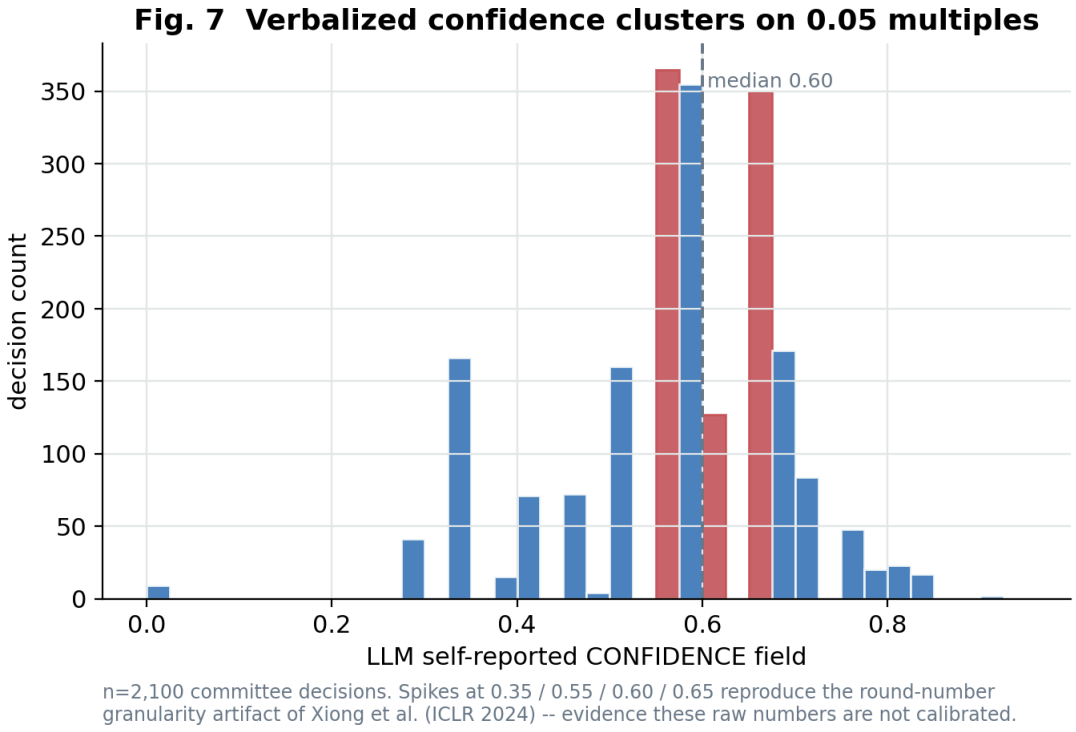


图 7 LLM 自报 CONFIDENCE 字段的分布（n=2100 次委员会决策）。取值聚集于 0.05 的整数倍（0.35/0.55/0.60/0.65），复现了言语化置信度的整数化伪精度（Xiong 等, ICLR 2024）。

6 架构评估

6.1 Macro 与 Quant：条件性成立

该划分受机制(a)支持的充要条件是两者消费不同的信息源(宏观/regime　数据与价格/因子数据)。若满足，划分有据；若两者实际读取高度重叠的上下文、仅以人设区分，则依据 2.4 节证据，应预期增益有限。该条件可由两者概率输出的相关性直接检验(7.1 节阶段 3)。

6.2 CIO 概率裁决：证据支持最弱的环节

以 LLM 担任最终聚合者存在双重脆弱性：其一，依据 2.3 节，其裁决易受输入报告的论断性与说服力影响，而非仅受证据影响；其二，依据 2.4 节，其自身言语化的概率与置信度存在系统性过度自信。当前报告中 CONFIDENCE: 0.75 一类数值即属原生言语化置信度，校准质量缺乏依据。替代方案有已验证模板：对各智能体的概率输出做确定性聚合(均值、中位数或截尾均值　[9])，以聚合值作为系统对外概率。在现有证据下，这是预期校准收益最大的单项架构改动。其边界为：该论点仅针对最终概率的产生方式，CIO 输出的执行计划、风险预案等叙事性内容不在替换范围内。

6.3 确定性层：证据支持最强的部分

Risk Officer 的价值在于约束与护栏(机制 c)而非辩论参与。Sanity Check 覆盖与独立快崩防御为机制(c)的标准实现，与护栏及神经符号文献的方向一致；已上线的校准层同属此类——它承担了原生 LLM 概率无法自行完成的校准工作。

6.4 无辩论设计与证据一致

单轮、无辩论的纵向汇总避开了鞅中性(2.3 节）与谄媚、从众等失效模式。在校准优先目标下，引入多空对抗辩论是所有可选改动中正当性最低的一项；如需引入，应按无害性门槛处理(7.1 节阶段 4)。

6.5 对 RQ1 的回答

基于现有证据，不存在被实证确认的最优角色数量或划分。TradingAgents 的七角色与本系统的四角色处于同等实证地位(均为未消融的设计选择），且 2.6 节弱先验反对扩张。决策相关变量不是角色计数，而是可供分解的独立信息源数量(机制 a)与可聚合的独立概率样本数量(机制 b)。角色数应是信息源结构的因变量，而非自变量。

7 研究计划

7.1 五阶段预注册架构实验

以下五阶段实验全部可在现有 walk-forward 框架上执行，按预期信息增益排序，均要求预注册假设与判定阈值，见表 3。

表 **3**　五阶段预注册架构实验设计

阶段	实验	设计	预注册假设	判定阈值
1	委员会折叠对照(决定性实验）	等采样预算三臂：A＝现行四角色＋LLM CIO；B＝单智能体(结构化提示含四类视角) × N 次采样＋确定性聚合；C＝单智能体单采样(下界)	B ≈ A(Brier 与覆盖率)	若 A 的 OOS Brier 优于 B 的幅度不超出框架噪声带，则以采样＋聚合替代角色扮演

2	聚合器消融	固定四智能体, 以均值/中位数/截尾均值/极化均值替换 LLM CIO 的概率裁决, 按 regime 分桶比较	确定性聚合在校准上不劣于 LLM CIO	取 OOS Brier 最小且覆盖率不劣者;极化仅样本内有效则弃用(提示成员信息相关)
3	信息源去相关检验	计算四智能体概率输出在全部重算点上的两两相关;按信息源(宏观、价格因子、跨资产流动性、新闻事件)重划角色	重划提升去相关度并降低集成 Brier	输出相关大于 0.9 的角色对合并;重划后集成 Brier 下降则保留
4	辩论无害性检验	可选 1–2 轮多空或互评通道	中性(鞅预期)	Brier 与覆盖率双重非劣, 且极端概率(大于 0.8 或小于 0.2)占比不上升时方启用;若仅概率变尖锐而命中率未改善则拒绝
5	确定性层强化	将全部 LLM 概率视为提案, 确定性规则可钳制、加宽或否决	—	钳制仅针对可判定的病态输出;避免对正常尾部预测过度收紧而侵蚀 Brier 的分辨率成分(对齐税 [20])

三点补充。其一, 阶段 1 直接回答 RQ1 的域内版本:若 B 臂与 A 臂在校准上不可区分, 四角色委员会的实质为一种高成本的采样方式, 应予简化;若 A 臂显著占优, 则角色分解在本系统的信息源结构下确有增益, 获得首个域内实证依据。其二, 阶段 2 先于阶段 4, 因其不依赖辩论结论且单项预期收益更高。其三, 阶段 5 受对齐税约束 [20]:确定性覆盖的触发条件应限于可判定的病态输出, 避免对正常尾部预测过度收紧而损失分辨率。

7.2 黄金单资产 MVP

定位声明:MVP 验证管线闭环(概率、仓位、防御、记账四环节的贯通), 不验证盈利能力。单资产将下注广度压至 1, 使盈利验证在统计上更难而非更易(4.3 节), 该声明需预先写入文档, 以防短窗结果被误读为成功或失败。选择黄金的依据:组合内仓位占比最大(49.7%)、其余权益资产处于冻结状态、概率表历史最深、TA 复现包内 GC=F 数据与评估管线现成。工作项按预期信息增益排序如下。

(1)防御层黄金语义重检。现行快崩防御(高 VIX 分位拦截买入)为权益类语义;黄金在危机中呈双相行为——流动性挤兑期与风险资产同跌(2020 年 3 月两周回撤约 12%–14%), 随后危机买盘接力。以自有概率表按 VIX 分位分桶, 检验高分位桶 60/90 天条件分布是否右偏;若是, 规则应修订为拦截 30 天追买、放行分批建仓, 而非一刀切。该项零 LLM 成本, 且可能直接改变现行规则方向。

(2)确定性 sizing 层。由校准概率与分位确定性映射为仓位增量(带上限的分数 Kelly 或阈值带)。该层必须确定性:将校准概率输入随机化的 sizing, 等价于在净化后的信号下游重新引入噪声, 使校准层的贡献被抵消。

(3)反事实记账。每次拦截与裁决落快照, 并记录未执行路径的损益;积累的是未来可下结论的样本, 而非即时结论。

(4)特征审计。剔除权益原生特征(Fear&Greed、trailing PE), 保留实际利率代理与美元指数;显式声明黄金特征集, 而非继承权益特征集。

(5) 三臂归因。买入持有、纯确定性策略(校准概率表＋防御＋sizing, 零 LLM 成本)、完整系统三臂对照, 并附 200 日均线趋势规则作为经典外部基线——同时回应复现包已声明的“基线未对标经典因子”局限。

路线图定位:多资产扩展是盈利能力验证的前置条件(下注广度的唯一来源), 而非可选项。

8 结论

对 RQ1:现有文献不支持任何特定角色数量或划分方式为最优;七角色与四角色处于同等实证地位。对 RQ2:可检验的设计原则为按信息源分解智能体、以确定性方法聚合独立概率样本、以确定性规则层包裹 LLM 提案, 并避免辩论与人设对概率输出的扭曲;这是一套设计纪律, 而非固定蓝图。对 RQ3:openInvest 的无辩论汇总与确定性防御及校准层与证据方向一致;以 LLM 叙事性推导最终置信度是证据支持最弱的环节, 应优先以确定性聚合替换其概率裁决职能;Macro 与 Quant 划分的正当性可由输出相关性检验。对 RQ4:校准目标与盈利目标经分层框架联接——优化层以校准拟合, 决策层以确定性规则翻译, 监控层以配对损益归因;在主动管理基本定律下, 校准与纪律即系统可控变量的名称, “经样本外验证的概率, 加上强制执行的纪律”由此构成该系统可验证的能力边界。

9 局限性

其一, 证据迁移缺口:第 2 节受控研究多在通用推理基准上进行, 而非金融路径概率预测;第 7 节实验即为弥合该缺口的手段。其二, 校准导向的多智能体金融研究缺乏先例, 本文评估口径无直接对照。其三, 文献权重:核心结论依托同行评审场所(NeurIPS、ICML、ACL、ICLR、Science Advances)的工作[2,3,6,7,8,9];[5]、[10–14]、[17–20] 等为预印本, 权重应相应下调。其四, 本文实验记录中, 5.1 的消融与 5.5 的案例属工程记录与描述性观察, 未达受控发表级实验标准;5.4 复现包的局限已随包声明。其五, 作者同时为系统开发者, 存在确认偏误风险, 预注册流程为主要缓解手段;单系统案例研究的外推性有限, 可移植的部分为第 4 节框架与第 7 节实验设计, 而非具体参数取值。

参考文献

[1] TradingAgents: Multi-Agents LLM Financial Trading Framework. arXiv:2412.20138 (v7, 2025).

[2] Wang Q., et al. Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key? ACL 2024. arXiv:2402.18272.

[3] Smit A. P., et al. Should We Be Going MAD? A Look at Multi-Agent Debate Strategies for LLMs. ICML 2024. arXiv:2311.17371.

[4] Zhang H., et al. If Multi-Agent Debate Is the Answer, What Is the Question? arXiv:2502.08788 (2025).

[5] Tran H., Kiela D. Single-Agent LLMs Outperform Multi-Agent Systems on Multi-Hop Reasoning Under Equal Thinking Token Budgets. arXiv:2604.02460 (2026).

[6] Choi et al. Debate or Vote: Which Yields Better Decisions in Multi-Agent Large Language Models? NeurIPS 2025 (spotlight). arXiv:2508.17536.

[7] Xiong M., et al. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. ICLR 2024. arXiv:2306.13063.

[8] Schoenegger P., Tuminauskaite I., Park P. S., Tetlock P. E. Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy. Science Advances, 10(45), 2024. DOI: 10.1126/sciadv.adp1528.

[9] Halawi D., Zhang F., Yueh-Han C., Steinhardt J. Approaching Human-Level Forecasting with Language Models. NeurIPS 2024. arXiv:2402.18563.

[10] Talk Isn't Always Cheap. arXiv:2509.05396 (2025).

[11] Peacemaker or Troublemaker. arXiv:2509.23055 (2025).

[12] Herd Behavior（LLM 多智能体从众行为研究）. arXiv:2505.21588 (2025).

[13] MADIAVE. arXiv:2510.05611 (2025).

[14] PD3. arXiv:2505.17492 (2025).

[15] AlphaAgents (BlackRock). arXiv:2508.11152 (2025).

[16] FinCon: A Synthesized LLM Multi-Agent System with Conceptual Verbal Reinforcement. NeurIPS 2024. arXiv:2407.06567.

[17] R2-Guard. arXiv:2407.05557 (2024).

[18] LLMSymGuard. arXiv:2508.16325 (2025).

[19] 基于 Lean 4 的金融决策形式化验证. arXiv:2604.01483 (2026).

[20] Strat-LLM（含对齐税测量）. arXiv:2605.06024 (2026).

[21] Tetlock P. E., Gardner D. Superforecasting: The Art and Science of Prediction. Crown, 2015; Satopää V. A., et al. 关于极化聚合适用条件的系列工作.

[22] Grinold R. C. The Fundamental Law of Active Management. Journal of Portfolio Management, 15(3), 1989.

[23] Lo A. W. The Statistics of Sharpe Ratios. Financial Analysts Journal, 58(4), 2002.

[24] Bailey D. H., Lopez de Prado M. The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality. Journal of Portfolio Management, 40(5), 2014.

[25] Gneiting T., Raftery A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association, 102(477), 2007.

[26] Kelly J. L. A New Interpretation of Information Rate. Bell System Technical Journal, 35, 1956.

[27] Morningstar. Mind the Gap（年度投资者回报差研究）.