Finance 金融, 2026, 16(5), 588-601
Published Online September 2026 in Hans. https://www.hanspub.org/journal/fin
https://doi.org/10.12677/fin.2026.165058

Hans汉斯

大语言模型交易决策趋同特征：A股情境实验

聂佳伦

中央财经大学金融学院，北京

收稿日期：2026年8月5日；录用日期：2026年8月25日；发布日期：2026年9月9日

摘　要

目的：考察统一政策与账户设定下，通用大语言模型产品的A股调仓输出是否趋同。方法：选取2026年沪深交易规则修订中的两项规定，设置实施首日和标记为实施21天后的两个情境。2026年7月30日对ChatGPT、Claude和DeepSeek在每个条件下各开启5次全新对话，取得30份首次JSON回答，比较方向、目标仓位和配置距离。结果：29份回答通过权重与资金平衡校验。条件A、B分别有14/15、10/15份回答减持提示词中的ST及*ST资产类别，其他三类风险资产均未出现卖出。有效向量的产品内与跨产品平均距离分别为6.0和7.7个百分点，分组表现并不一致。结论：样本呈现风险警示资产类别上的局部趋同，配置幅度和条件反应仍有产品差异。JSON示例可能造成选择性锚定，结果只适用于给定提示框架。

关键词

大语言模型，交易决策，决策趋同，策略拥挤，A股市场

Convergence Patterns in LLM Trading Decisions: An A-Share Scenario Experiment

Jialun Nie

School of Finance, Central University of Finance and Economics, Beijing

Received: August 5, 2026; accepted: August 25, 2026; published: September 9, 2026

Abstract

Objective: To examine whether general-purpose LLM products converge in A-share reallocation outputs under a common policy and account setting. Methods: Two prompt conditions represented the implementation day and a date labelled as 21 days after implementation of two selected provisions from the 2026 revision of the Shanghai and Shenzhen trading rules. On 30 July 2026, ChatGPT, Claude, and DeepSeek were queried in five fresh conversations per condition, producing 30 first-pass JSON responses. Direction, target weights, and allocation distances were compared. Results:

文章引用: 聂佳伦. 大语言模型交易决策趋同特征: A 股情境实验[J]. 金融, 2026, 16(5): 588-601.
DOI: 10.12677/fin.2026.165058

Twenty-nine responses passed the weight and cash-balance checks. In Conditions A and B, 14/15 and 10/15 responses, respectively, reduced the prompt-defined ST and *ST category. None sold the other three risky-asset categories. Mean within-product and cross-product distances were 6.0 and 7.7 percentage points, but subgroup patterns varied. Conclusion: The sample shows local convergence in the risk-warning category, alongside product differences in allocation magnitude and responses to scenario framing. Because the JSON example may have induced selective anchoring, the findings apply only to the specified prompt framework.

Keywords

Large Language Model, Trading Decision, Decision Convergence, Crowded Trade, A-Share Market

1. 引言

大语言模型在金融场景中的用途正由信息提取和文本分析延伸到资产配置与交易建议。FinMem 等研究把记忆、角色设定和投资动作纳入同一交易智能体框架，表明语言模型可以直接生成可执行的投资动作[1]。当多个投资者依赖数量有限的通用模型或应用产品时，问题不再只是单个建议是否准确，还包括这些建议是否会在相同信息下趋于一致。

国内研究已将新闻、股价和宏观信息接入金融大模型，并据此生成和检验买入、持有、卖出信号[2]。金融大模型的评测也在从单一任务走向多场景决策。InvestorBench 以 13 种大语言模型为基础模型，覆盖股票、加密资产和 ETF 等任务[3]；CryptoTrade 则通过链上、链下信息和反思机制评估加密资产交易智能体[4]。Fin-Bias 进一步发现，在含有明确人类意见的金融文本中，模型生成的投资判断会向这些意见靠拢[5]。这些研究说明，金融决策能力和语境敏感性已经可以系统评估，但尚未直接回答：不同公众模型应用在同一 A 股政策提示下，是否会反复给出相近的目标组合。

策略趋同并非语言模型出现后才有。Stein 指出，使用相近且缺少基本面锚的策略时，专业投资者难以判断同一交易中已经聚集了多少同业，拥挤本身会成为风险来源[6]。Khandani 和 Lo 对 2007 年量化基金震荡的研究显示，相似组合的集中去杠杆能够通过价格压力相互传导[7]。专业机构和量化组合的证据不能直接移植到语言模型，但可据此区分文本相似与交易同步：只有方向、幅度和执行窗口同时接近，才可能形成市场层面的集中交易。

通用模型可能增加决策相关性，也可能保留产品差异。基础模型、常见叙事与数据供应链的集中会压低差异；系统提示、风险规则、检索工具和产品路由又可能产生分化。金融稳定理事会已将服务商集中、市场相关性、模型风险与数据治理列为人工智能在金融体系中的潜在脆弱性[8]。共同的技术标签与决策一致性并非同一概念，本文因此把方向和幅度分开考察。

本文以一项 A 股交易规则调整为信息材料，在统一初始组合、风险偏好、持有期目标和输出格式的基础上，比较 ChatGPT、Claude 与 DeepSeek 在两种情境日期条件下的 30 份首次回答。分析只使用可复核的方向、仓位与理由文本，不把界面显示的“思考”过程等同于模型内部推理。已有研究指出，语言模型生成的解释可能具有可读性，却未必忠实反映实际决策原因[9]。研究问题有三项：提示词中的风险警示资产类别是否出现稳定的同向订单；同一产品的重复回答是否比跨产品回答更接近；两个情境日期条件之间的差异是否因产品而异。

2. 概念界定与度量

2.1. 方向、幅度与执行窗口

交易决策同质化至少包含叙述、方向、幅度和执行窗口四个层次。叙述相似指回答使用类似的风险或收益理由；方向相似指对同一资产同时增配、减配或保持；幅度相似指目标仓位接近；执行窗口相似指订单在相近时间提交。本文量化比较方向与幅度，叙述层面仅作定性记录；由于没有真实订单，执行窗口不进入检验。为进一步整理叙述层面的差异，本文另对模型陈述的理由文本进行人工主题编码。

设资产 k 的初始权重为 $w_{0,k}$，某次回答给出的目标权重为 $w_{1,k}$，则带符号订单权重定义为：

$$\Delta w_k = w_{1,k} - w_{0,k}$$

订单权重大于 0 编码为买入，小于 0 编码为卖出，等于 0 编码为持有。统一初始仓位后，“目标 0%”只有在原组合持有该资产时才构成卖出，由此区分“回避”与实际卖单。

2.2. 配置距离与重复稳定性

对两份有效目标权重向量 i 和 j (五类目标权重合计均为 100%)，本文采用下式计算配置距离：

$$D_{ij} = 0.5 \times \sum_{k=1}^{5} \left| w_{i,k} - w_{j,k} \right|$$

五类资产权重以百分数计，配置距离的单位为百分点；距离越小，两个目标组合越接近。产品内距离由同一产品、同一条件下的所有两两组合计算，跨产品距离则在同一条件下配对。

重复稳定性还使用两个描述性指标：一是同一产品与条件下最常见完整方向组合所占比例，二是最常见目标仓位组合所占比例；判断两个组合相同时，要求五项目标仓位全部一致。每个单元只有 5 次，这些比例仅用于描述本样本的重复稳定性。

3. 实验设计

3.1. 政策情境与账户约束

实验材料来自沪深交易所 2026 年交易规则修订。上交所公告显示，盘后固定价格交易范围扩展至全部 A 股和交易型开放式基金，主板风险警示股票涨跌幅限制由 5%调整为 10%，相关规则自 2026 年 7 月 6 日起实施[10]；深交所发布的 2026 年修订交易规则亦自同日起施行[11]。条件 A 在提示词中把决策日期设为 2026 年 7 月 6 日 14:50，即实施首日收盘前；条件 B 设为 2026 年 7 月 27 日 14:50，并增加规则于 4 月 24 日发布、已实施 21 个自然日且不提供期间行情的说明。两组回答均在 7 月 30 日实际采集，故 A、B 是情境日期条件，而非真实的纵向观测。

实验提示词只选取了上述两项规定。上交所公告还包括基金收盘阶段由连续竞价调整为收盘集合竞价，该项没有进入提示词。因此，本文的情境材料是完整规则修订的一个信息子集，ETF 相关输出仅反映模型对所给两项规定的反应。

两个条件均使用 100 万元模拟账户，风险偏好为中等，目标为未来 1~3 个月的风险调整后收益，不允许杠杆或卖空。提示词还给出最大可容忍回撤 10%和单边交易成本 0.1%。由于没有收益路径、协方差或成本扣减字段，两项作为决策背景，不进入数值合规指标。初始组合为提示词中的 ST 及*ST 资产类别 10%、宽基 ETF 20%、券商股 10%、其他 A 股 20%和现金 40%，并要求五项目标仓位合计 100%，只输出规定结构的 JSON 对象。两个条件的完整设定见表 1。

Table 1. Experimental conditions and common constraints
表 1. 实验条件与共同约束

维度	条件 A：实施首日情境	条件 B：实施 21 天后情境
提示词日期	2026-07-06 14:50	2026-07-27 14:50
附加时间信息	实施首日；无期间行情	4 月 24 日发布；已实施 21 天；无期间行情
初始组合	提示词中的 ST 及*ST 资产类别 10%；ETF 20%；券商 10%；其他 A 股 20%；现金 40%	与条件 A 相同
决策背景	中等风险；回撤目标 10%；成本 0.1%；不加杠杆或卖空	与条件 A 相同
输出	首次 JSON 回答；目标仓位合计 100%	与条件 A 相同
重复	每个产品 5 次全新对话	每个产品 5 次全新对话

3.2. 产品设置与采集程序

数据于 2026 年 7 月 30 日北京时间约 16:30 集中采集。每次均开启全新对话，只保留第一次完整回答，不追问、不纠错，也不向产品提供此前回答。记忆功能没有关闭，因此全新对话并不等同于完全没有跨会话状态。三项产品的采集顺序及每次精确时间未单独记录，30 份记录按重复产品输出处理，不视为独立随机样本。三个产品的界面与实验设置见表 2。

Table 2. Product interfaces and experimental settings
表 2. 产品界面与实验设置

产品	使用界面	可见模型/模式	推理设置	联网与界面记录
ChatGPT	网页端	GPT-5.6 Sol	High	无手动联网开关；显示“思考”；回答未见搜索或引用标记
Claude	电脑客户端	Sonnet 5	Thought Effort High	无手动联网开关；显示“思考、分析、权衡”；回答未见搜索或引用标记
DeepSeek	网页端	未单独记录	开启深度思考	联网搜索关闭；回答未见网址或来源标记

注：表中“思考、分析、权衡”仅记录界面显示状态，不作为模型内部思维链证据。ChatGPT 与 Claude 界面未提供可操作的联网开关，统一提示词明确禁止搜索和外部工具；“未见搜索或引用标记”不能证明后台绝对未联网。三个账户均关闭了用于改进模型的数据选项；该项属于数据使用设置，并非生成条件。

每个产品在条件 A 和 B 下各取得 5 份回答，形成 $3 \times 2 \times 5 = 30$ 份观察。可见产品名称不必然对应固定模型快照，三个产品的推理档位也不能视为等量处理；本文只在产品及其当时可见设置层面报告差异，不把结果归因于某个固定的底层模型版本。

附录 A 逐字列出条件 A 提示词，并逐项列明条件 B 与 A 的差异；条件 B 与条件 A 相同的内容不再重复列出。其 JSON 示例把四类风险资产的 target_pct 写为 0、order_pct 写为负值，可能把回答锚定在减仓方向；条件 B 示例又预填 information_staleness 为“高”。条件 B 还同时改变日期、发布与实施说明，现有设计据此识别的是两套提示条件的总体差异。

3.3. 校验、编码与异常处理

程序依次校验 JSON 语法、字段完整性、五项目标权重之和、order_pct 与目标减初始仓位的一致性、action 与订单符号的一致性、现金变化与风险资产净订单的平衡，以及持有期和理由条数。最大回撤与交易成本因缺少可计算字段而不进入数值合规校验。格式或数值错误均保留原貌，不要求模型重写。

方向频数和合规率使用全部 30 份回答。完整组合均值、重复目标组合比例和配置距离只使用权重合计为 100%的 29 份有效向量。唯一被排除的权重向量来自 DeepSeek 条件 A 第 3 次回答；该回答仍进入 JSON 合规、方向和异常率统计。本文以描述性比较为主，区间估计与检验仅作探索性补充；不使用实施后的市场表现评价建议收益。

30 份原始回答及其产品、条件和重复序号均由作者留存。DeepSeek 条件 A 第 3 次回答的权重合计存在异常，正文仍将其纳入 JSON 合规、方向和异常率统计，但不纳入完整组合均值和配置距离计算。

3.4. 统计方法与文本编码

每个实验单元只有 5 次回答，因此本文仍以描述性结果为主，置信区间和检验结果只作为补充。方向比例采用 Wilson 95%置信区间；条件 A、B 之间的方向差异采用双侧 Fisher 精确检验；同一份回答中不同资产的方向差异采用 McNemar 精确检验。配置距离来自同一批回答，彼此并不独立，所以不使用常规两样本检验，而是在每个条件内置换产品标签，并保持各产品有效样本数不变。置换检验重复 200,000 次，随机种子为 20,260,819。

上述比较有一个工作前提，即把 30 次首次回答近似看作彼此独立的重复生成结果。但实际采集中，同一产品连续测试 5 次，账户设置相同，记忆功能也没有关闭；采集顺序和每次的准确时间又没有单独记录，因此回答之间仍可能相互关联。基于这一限制，文中的置信区间和 p 值只用于本批样本的探索性比较，不能外推到模型更新后的长期表现。

本文对 decision_basis 和 main_risks 两个字段进行主题编码。通读 30 份回答后，把文本归为五个可以重叠的主题：ST 波动与下行风险(T1)、信息不足与政策反应不确定(T2)、交易机制与流动性(T3)、组合约束、回撤与机会成本(T4)、ETF 或券商间接受益及其不确定性(T5)。编码单位是一份回答中的一个字段；某主题在该字段中明确出现一次即记为命中，同一字段可以命中多个主题。编码只记录文字中明确表达的内容，不推断模型内部推理。全部编码由作者独立完成，因此不报告编码者间一致性。样本共 180 条短文本(30 份回答 × 6 条理由)，每条不超过 60 个汉字，因而采用可以逐条核对的人工内容编码。

4. 实验结果

4.1. 输出合规性与外部信息使用痕迹

30 份回答均能解析为合法 JSON，且日期、行动标签、订单符号、持有期及理由条数均符合要求。29 份回答的五项目标仓位合计 100%，严格格式与数值合规率为 96.7%。DeepSeek 条件 A 第 3 次回答的目标仓位合计为 90%，风险资产订单与现金变化不平衡；其风险说明称“权益总仓位升至 80%”，列示的四类权益仓位实际合计 70%。本文保留该回答原貌，不进行人工补足。

30 份回答均将 used_external_information 填为 false，原文未出现网址、来源或引用标记，也没有给出实施后的行情、成交量或收益率数据。部分理由把交易范围扩展解释为“提升流动性”或“改善券商收入”，这些是模型作出的政策机制推断，而非已经观察到的市场事实。

4.2. 方向趋同集中在提示词的风险警示资产类别

条件 A 的 15 份回答中，14 份卖出提示词中的 ST 及*ST 资产类别，1 份维持原 10%仓位，卖出比例为 93.3%；条件 B 有 10 份卖出、5 份持有，卖出比例为 66.7%。两个条件合计 24/30 份回答减持该类别，占 80%，没有任何回答增持。固定初始仓位后，这些“卖出”均对应带符号的负订单，而不是由“回避”

推断出来的方向。各产品与条件的方向频数见表 3。

Table 3. Direction frequencies by product and condition
表 3. 各产品与条件的方向频数

产品	条件	ST (买/卖/持)	ETF (买/卖/持)	券商(买/卖/持)	其他 A 股(买/卖/持)
ChatGPT	A	0/4/1	3/0/2	0/0/5	0/0/5
ChatGPT	B	0/4/1	0/0/5	0/0/5	0/0/5
Claude	A	0/5/0	3/0/2	3/0/2	0/0/5
Claude	B	0/5/0	4/0/1	2/0/3	0/0/5
DeepSeek	A	0/5/0	3/0/2	1/0/4	2/0/3
DeepSeek	B	0/1/4	1/0/4	1/0/4	0/0/5

注：“买/卖/持”依次表示买入、卖出和持有的回答次数，每个单元合计 5 次。

间接受益资产没有形成同样集中的方向。条件 A 中，ETF 买入 9/15、持有 6/15；条件 B 中，ETF 买入 5/15、持有 10/15。券商买入比例分别为 4/15 和 3/15，其余均为持有；其他 A 股只在条件 A 的 DeepSeek 回答中出现 2 次买入。方向的集中主要出现在提示词的风险警示资产类别，并未扩展为整个政策相关组合的一致扩张或收缩。

方向指标的区间估计与精确检验结果见表 4。条件 A 的卖出比例为 93.3%，Wilson 95%置信区间为 70.2%~98.8%；条件 B 为 66.7%，区间为 41.7%~84.8%；两个条件合计为 80.0%，区间为 62.7%~90.5%。两个条件之间的差异未达到统计显著(双侧 Fisher 精确检验，$p = 0.169$)。由于每个单元只有 5 次回答，条件 B 卖出比例较低只作为本批样本的描述性差异。

Table 4. Interval estimates and exact tests for direction measures
表 4. 方向指标的区间估计与精确检验

指标	比例	Wilson 95%置信区间
条件 A：ST 及*ST 卖出	14/15 (93.3%)	70.2%~98.8%
条件 B：ST 及*ST 卖出	10/15 (66.7%)	41.7%~84.8%
两条件合计：ST 及*ST 卖出	24/30 (80.0%)	62.7%~90.5%
两条件合计：卖出其他三类资产之一的回答	0/30 (0.0%)	0.0%~11.4%

注：末行以回答为单位计数，即 30 份回答中有多少份至少卖出宽基 ETF、券商或其他 A 股之一。条件 A 与条件 B 的 ST 及*ST 卖出比例差异经双侧 Fisher 精确检验为 $p = 0.169$；ST 及*ST 与其他三类资产来自同一批回答，属配对观测，以 McNemar 精确检验比较，不一致对为 24 比 0，$p < 0.001$。ETF 买入比例在两个条件之间的差异未达到显著(9/15 与 5/15，$p = 0.272$)。

4.3. 目标仓位的产品与条件差异

在 29 份有效权重向量中，Claude 对提示词中 ST 及*ST 资产类别的平均目标仓位在两个条件下分别为 2.8%和 3.0%，ChatGPT 分别为 5.0%和 6.0%，变化均较小。DeepSeek 在条件 A 的 4 份有效向量中将 ST 目标仓位全部设为 0%，条件 B 则有 4 次维持 10%、1 次降至 0%，平均升至 8.0%。条件 B 的卖出比例低于条件 A 并非三个产品同步变化，差异主要来自 DeepSeek。各单元有效向量的平均目标仓位见表 5。

Table 5. Mean target weights of valid allocation vectors
表 5. 有效权重向量的平均目标仓位

产品	条件	有效 n	ST	ETF	券商	其他 A 股	现金
ChatGPT	A	5	5.0%	23.6%	10.0%	20.0%	41.4%
ChatGPT	B	5	6.0%	20.0%	10.0%	20.0%	44.0%
Claude	A	5	2.8%	22.6%	12.0%	20.0%	42.6%
Claude	B	5	3.0%	24.0%	12.0%	20.0%	41.0%
DeepSeek	A	4	0.0%	25.0%	11.25%	21.25%	42.5%
DeepSeek	B	5	8.0%	21.0%	11.0%	20.0%	40.0%

注：DeepSeek 条件 A 第 3 次回答因目标仓位合计 90%而不进入本表；其他 29 份回答均按原始目标权重计算，未作标准化或人工修正。

从单次取值看，条件 A 共有 6 份回答将提示词中的 ST 及*ST 资产类别目标仓位设为 0%，条件 B 为 3 份。条件 B 的组间差异主要来自 DeepSeek 的 4 次 10%目标仓位。全部单次取值由作者留存。

ETF 方面，ChatGPT 的平均目标仓位由 23.6%降至 20.0%，DeepSeek 由 25.0%降至 21.0%，而 Claude 由 22.6%升至 24.0%。券商仓位在 ChatGPT 中始终为 10%，Claude 两个条件均为 12%，DeepSeek 约为 11%。同一政策材料没有对应唯一的“政策受益组合”：风险警示资产类别的方向较集中，ETF 和券商等间接受益链条的仓位判断则随产品和提示条件而异。

4.4. 产品内重复稳定性差异

ChatGPT 在条件 B 的 5 次回答中有 4 次给出完全相同的目标组合，DeepSeek 条件 B 亦为 4/5；Claude 条件 B 只有 2/5 完全相同。最常见完整方向组合的占比分别为 ChatGPT 条件 A 60%、条件 B 80%，Claude 为 60%和 40%，DeepSeek 为 40%和 80%。“同一产品重复回答稳定”只在部分产品和条件中成立。产品内重复稳定性指标见表 6。

Table 6. Within-product repetition stability
表 6. 产品内重复稳定性

产品	条件	有效向量 n	最常见方向占比	最常见目标组合占比	平均组内距离	中位组内距离
ChatGPT	A	5	60%	20%	5.1	5.0
ChatGPT	B	5	80%	80%	2.0	0
Claude	A	5	60%	20%	7.6	7.5
Claude	B	5	40%	40%	8.0	7.5
DeepSeek	A	4	40%	50%	11.7	10.0
DeepSeek	B	5	80%	80%	4.0	0

注：方向占比使用全部回答；目标组合占比和配置距离仅使用有效权重向量。最常见目标组合占比按五项目标仓位完全一致识别；5 份有效回答各不相同时，该值为 20%。配置距离单位为百分点。

汇总所有有效配对后，产品内平均配置距离为 6.0 个百分点，跨产品平均距离为 7.7 个百分点。条件 A 的产品内与跨产品距离分别为 7.6 和 8.5，条件 B 分别为 4.7 和 7.0。总体上，同一产品的目标组合略为接近，但 Claude 条件 B 的组内平均距离仍达 8.0，高于部分跨产品配对。这表现为有限的产品内接近，而非固定模板。两类平均距离相差 1.70 个百分点。由于这些距离共享同一批回答，不能按独立样本处理，

因此在条件内置换产品标签并保持各产品有效样本数不变。200,000 次重复检验得到双侧 $p = 0.005$。该结果说明本批样本中的产品内距离较小，但差距不足 2 个百分点，仍不能据此认定产品使用固定模板。

4.5. 决策依据与主要风险的主题分布

180 条理由文本(30 份回答 ×6 条)的主题编码频数见表 7。决策依据字段中，ST 波动与下行风险(T1)与组合约束、回撤与机会成本(T4)各命中 25/30，信息不足与政策反应不确定(T2)和交易机制与流动性(T3)各为 21/30，ETF 或券商间接受益及其不确定性(T5)为 15/30；主要风险字段中，T2 最高，为 27/30，T1 与 T4 各为 22/30，T3 为 13/30，T5 为 6/30。24/30 份回答减持风险警示资产类别，25/30 份在决策依据中明确把该类别与波动或下行风险联系起来。

T5 在不同产品之间的差别较大。该主题在决策依据中的命中数分别为 ChatGPT 2 次、Claude 9 次、DeepSeek 4 次，在主要风险中分别为 1 次、3 次和 2 次。ChatGPT 在决策依据中提及 T1 共 9 次，但在主要风险中只有 3 次；Claude 和 DeepSeek 在主要风险中则分别提及 10 次和 9 次。由此可见，不同产品会把同类信息放在不同字段中。这里的主题编码只记录显性文字，不能代表模型内部推理。

Table 7. Theme frequencies in decision bases and main risks
表 7. 决策依据与主要风险的主题命中频数

主题	决策依据 ChatGPT/Claude/DeepSeek (合计)	主要风险 ChatGPT/Claude/DeepSeek (合计)
T1 ST 波动与下行风险	9/10/6 (25)	3/10/9 (22)
T2 信息不足与政策反应不确定	10/6/5 (21)	9/10/8 (27)
T3 交易机制与流动性	5/10/6 (21)	5/4/4 (13)
T4 组合约束、回撤与机会成本	10/9/6 (25)	9/6/7 (22)
T5 ETF 或券商间接受益及其不确定性	2/9/4 (15)	1/3/2 (6)

注：编码单位为单份回答的单个字段，主题不互斥，各列合计可超过 30；主题定义和编码方法见 3.4 节。

5. 讨论

5.1. 风险警示资产类别上的局部趋同

本样本最清晰的现象是风险警示资产类别的减仓方向较集中。规则变化提高了主板 ST、*ST 股票的日涨跌幅上限，固定的 10%初始仓位又使低配建议表现为负订单。条件 A 中 14/15 份回答选择卖出，说明在这一提示框架下，三个产品经常给出同方向的风险收缩建议。

值得注意的是，JSON 示例同时为宽基 ETF、券商股和其他 A 股给出了负订单，但 30 份回答在这三类资产上均未出现卖出，表明输出并非对示例数值的逐字段复制。这一结果降低了纯机械复制作为唯一解释的可能性。然而，ST 资产类别与提示词中的风险信息联系更为直接，仍不能排除示例值与政策语义共同形成的选择性锚定。将每份回答分别编码为“是否减持风险警示资产类别”和“是否减持其他三类资产中的至少一类”，形成 30 组配对观测。探索性 McNemar 精确检验显示，24 份回答只减持前者，反向情形为 0 份，$p < 0.001$；以回答为单位计，其他三类资产的减持比例为 0/30，Wilson 95%置信区间上界为 11.4%。

这里的资产口径必须谨慎。政策条文针对主板风险警示股票，提示词的账户栏却笼统写作“ST 及*ST 股票 10%”，没有限定主板。该类别可能被理解为包括并非本次涨跌幅调整对象的其他板块风险警示股票，故本文不把它等同于严格的“直接受影响资产”样本。ETF 和券商方向的分歧也表明，更合适的概

括是“提示词中的风险警示资产类别局部趋同、其余配置较分散”，而非所有交易方向一致。

5.2. 不同产品对两套提示条件的反应差异

条件 B 没有提供实施后的市场表现，但不仅改变了提示词日期，还补充政策发布日、已实施 21 个自然日及期间行情缺失等表述。在这一组合变化下，DeepSeek 由条件 A 中四份有效向量全部清空 ST 资产类别，变为条件 B 中多数维持原仓位；ChatGPT 和 Claude 的平均目标仓位变化较小。现有数据将其归纳为产品对两套提示条件的不同反应，不作单一日期效应解释。

条件 B 的时效性标签全部为“高”，恰与示例中的预填值相同，而对应的调仓方向并不统一。这个现象一方面显示标签一致不等于订单一致，另一方面暴露了结构化示例的锚定风险。若要单独识别日期作用，后续实验应只替换日期字段，并让示例使用中性占位符或不提供示例值。

5.3. 与传统量化策略的比较边界

本文没有设置传统量化基准，因而不能判断语言模型策略的同质化是否更高。传统量化机构可能共享因子定义、数据供应商、风险模型与执行算法；语言模型产品也可能因训练目标、系统提示和工具配置而产生差异。机器学习资产定价研究展示了模型、特征和超参数选择的广阔空间[12]，但这种方法空间本身不能证明实际策略已经分散。有效比较需要让规则型策略、事件研究模型和语言模型在同一信息集、初始持仓和目标权重空间中运行。

相应指标至少应包括资产方向一致率、配置距离、换手率与订单时间重合度，而不能只比较回测收益。在缺少这组对照时，本文结论只涉及三个面向公众的模型应用在单一政策情境下的输出差异。

5.4. 风险管理与监督启示

本样本提示，机构使用多个语言模型角色并不自动形成风险分散。若这些角色接收相同政策文本、使用相同账户约束并在固定窗口执行，某些资产仍可能收到高度一致的减仓建议。风险控制可考虑差异化信息源、独立风险预算、错峰执行、单一事件仓位上限和人工否决机制，同时保留模型名称、界面版本、提示词、工具调用与最终订单之间的审计链路。

对监管和机构治理而言，重点不是披露可复制的完整策略，而是建立模型清单、版本记录、第三方依赖、责任分工与异常停机机制。国际证监会组织已将责任归属、模型测试、数据质量、第三方服务商和信息披露纳入人工智能与机器学习治理要求[13]。在此基础上，可使用标准化政策冲击情境比较主流系统的方向一致率和订单集中度，以识别局部拥挤风险。

国内研究也从生成式人工智能金融服务的风险治理[14]和数字金融模型的可解释性风险[15]等角度提出数据治理、模型审计和责任界定要求。对交易智能体而言，这些要求可落实为提示词与模型版本留档、异常输出复核、第三方依赖登记和责任边界说明。

5.5. 研究限制

第一，样本只有一个政策事件、三个产品和每单元 5 次回答，只适合探索性描述，不能估计产品长期稳定性。两两距离共享同一批回答，配对之间也不独立，因此配置距离不采用常规两样本检验，而使用条件内置换检验；全部区间估计和检验只作本批样本的探索性比较。第二，JSON 示例把风险资产目标仓位写为 0 并给出负订单，可能主动诱导减仓；条件 B 还预填“高”时效标签。其中，条件 A 的 15 份回答全部填“不适用”，条件 B 全部填“高”，与两份示例的预填值完全一致；该字段按提示结构回声处理，不列为产品时效判断结果。

第三，两组回答都在 7 月 30 日采集，A、B 只是提示词中的情境日期；条件 B 又同时加入发布与实

施说明，不能解释为真实的实施前后效应。第四，面向公众的界面不能固定后端模型、随机种子、温度或系统提示，DeepSeek 的可见模型版本也未记录；三个推理档位并不等价。第五，记忆功能未关闭，且采集顺序、逐次时间没有单独记录，跨会话状态和顺序影响均无法排除。

第六，ChatGPT 和 Claude 界面没有手动联网开关。提示词虽然禁止搜索，回答也没有可见的搜索或引用痕迹，仍不能证明后台绝对未调用外部信息。第七，提示词中的 ST 及*ST 资产类别没有限定主板，且未纳入基金收盘集合竞价调整，情境信息与完整规则适用范围并不完全一致；另有一份回答的目标权重不平衡。最后，本研究没有真实账户、采用比例、订单簿或市场流动性数据。从产品输出到价格影响之间仍隔着持仓分布、采用规模、执行规则和市场承接能力。

5.6. 研究展望

后续研究可从四个方面改进。第一，每个单元只有 5 次回答，后续可在同一时间窗内一次性增加重复次数，避免追加样本时模型或界面版本已经变化。第二，可在提示词中加入观点模糊或彼此矛盾的模拟市场分析，考察模型面对非结构化信息时，方向选择是否仍然集中。第三，可把五类资产进一步细分到板块、行业或个股，使任务更接近真实的资产配置。第四，可增设中性示例的对照组，并只替换日期字段，以分别考察示例值和情境时点的影响。

6. 结论

在同一政策材料和账户设定下，30 份回答在提示词的 ST 及*ST 资产类别上呈现减仓集中：条件 A 有 14/15 份卖出，条件 B 为 10/15 份，均未增持；宽基 ETF、券商和其他 A 股三类资产均未出现卖出，方向集中没有扩展到整个政策相关组合。回答级配对检验的不一致对为 24 比 0，探索性 McNemar 精确检验 $p < 0.001$。

29 份有效向量的产品内平均距离低于跨产品距离，但分组结果并不一致，DeepSeek 对两套提示的反应也更大。结果可概括为“风险警示资产类别局部方向趋同、配置幅度和条件反应仍有产品差异”。然而，示例可能造成锚定，两组提示也不是单变量对照；这些产品输出不能证明真实订单同步或市场冲击，也不能替代与传统量化策略的同口径比较。本批样本中两类平均距离相差 1.70 个百分点，置换检验 $p = 0.005$，但这一差距不足 2 个百分点。

参考文献

[1] Yu, Y., Li, H., Chen, Z., Jiang, Y., Li, Y., Zhang, D., *et al*. (2024) FinMem: A Performance-Enhanced LLM Trading Agent with Layered Memory and Character Design. *Proceedings of the AAAI Symposium Series*, **3**, 595-597. https://doi.org/10.1609/aaaiss.v3i1.31290

[2] 樊嘉诚, 吴俊樊, 林建浩. 多源数据驱动的金融大语言模型文本分析与交易策略[J]. 计量经济学报, 2026, 6(2): 304-324.

[3] Li, H., Cao, Y., Yu, Y., Javaji, S.R., Deng, Z., He, Y., *et al*. (2025) INVESTORBENCH: A Benchmark for Financial Decision-Making Tasks with LLM-Based Agent. *Proceedings of the* 63*rd Annual Meeting of the Association for Computational Linguistics*, **1**, 2509-2525. https://doi.org/10.18653/v1/2025.acl-long.126

[4] Li, Y., Luo, B., Wang, Q., Chen, N., Liu, X. and He, B. (2024) CryptoTrade: A Reflective LLM-Based Agent to Guide Zero-Shot Cryptocurrency Trading. *Proceedings of the* 2024 *Conference on Empirical Methods in Natural Language Processing*, Miami, 12-16 November 2024, 1094-1106. https://doi.org/10.18653/v1/2024.emnlp-main.63

[5] Hu, X. and Zhao, J. (2026) Fin-Bias: Comprehensive Evaluation for LLM Decision-Making under Human Bias in Finance Domain. *Findings of the Association for Computational Linguistics*: *ACL* 2026, San Diego, 2-7 July 2026, 5678-5694. https://doi.org/10.18653/v1/2026.findings-acl.279

[6] Stein, J.C. (2009) Presidential Address: Sophisticated Investors and Market Efficiency. *The Journal of Finance*, **64**, 1517-1548. https://doi.org/10.1111/j.1540-6261.2009.01472.x

[7] Khandani, A.E. and Lo, A.W. (2011) What Happened to the Quants in August 2007? Evidence from Factors and

Transactions Data. *Journal of Financial Markets*, **14**, 1-46. https://doi.org/10.1016/j.finmar.2010.07.005

[8] Financial Stability Board (2024) The Financial Stability Implications of Artificial Intelligence. https://www.fsb.org/2024/11/the-financial-stability-implications-of-artificial-intelligence/

[9] Turpin, M., Michael, J., Perez, E. and Bowman, S. (2023) Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *Proceedings of the* 37*th Annual Conference on Neural Information Processing Systems*, New Orleans, 10-16 December 2023. https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html

[10] 上海证券交易所. 上交所修订发布《上海证券交易所交易规则》[EB/OL]. 2026-04-24. https://www.sse.com.cn/aboutus/mediacenter/hotandd/c/c_20260424_10816474.shtml, 2026-08-03.

[11] 深圳证券交易所. 关于发布《深圳证券交易所交易规则(2026 年修订)》的通知[EB/OL]. 2026-04-24. https://www.szse.cn/lawrules/rule/trade/current/t20260424_620190.html, 2026-08-03.

[12] Gu, S., Kelly, B. and Xiu, D. (2020) Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, **33**, 2223-2273. https://doi.org/10.1093/rfs/hhaa009

[13] International Organization of Securities Commissions (2021) The Use of Artificial Intelligence and Machine Learning by Market Intermediaries and Asset Managers: Final Report. FR06/2021. https://www.iosco.org/library/pubdocs/pdf/IOSCOPD684.pdf

[14] 张龙, 郭允臻. "生成式人工智能 + 金融服务"应用的风险与治理[J]. 当代金融研究, 2025(4): 34-43.

[15] 张润驰, 岳中刚, 张国法, 孙明明, 金磊. 数字金融场景中人工智能模型可解释性风险的特征与治理研究[J]. 中国工程科学, 2026, 28(2): 113-124.

附录：正式实验提示词

条件 A 按实际发送文本逐字收录；条件 B 列出相对条件 A 的四处差异，其余文字、账户约束、政策信息、JSON 字段和输出规则均与条件 A 一致。JSON 中的数字是原提示示例，并非研究者推荐的目标组合；示例值可能构成锚定，相关影响已在正文中说明。

A.1. 条件 A：实施首日情境

你是一名负责 A 股组合的交易决策者。请仅依据本提示中给出的信息作出一次模拟调仓决策，不得搜索互联网，不得调用外部工具，不得自行补充本提示以外的行情、成交量、资金流、估值、历史统计或事件后市场表现。若信息不足，可以保持原仓位。

【决策时点】

2026 年 7 月 6 日 14:50 (新规实施首日收盘前)。

【政策信息】

自 2026 年 7 月 6 日起：

1、盘后固定价格交易适用证券范围由科创板、创业板扩展至全部 A 股及沪深市场 ETF，交易时间为 15:05~15:30。

2、沪深交易所主板 ST、*ST 股票的涨跌幅限制由 5%调整为 10%，与主板其他股票一致。

【模拟账户约束】

1、账户总资产为 100 万元，风险偏好中等。

2、目标是在未来 1~3 个月内提高风险调整后收益，账户最大可容忍回撤为 10%。

3、不使用融资、融券、期货、期权或其他杠杆工具。

4、单边交易成本按成交金额的 0.1%计。

5、调仓前组合为：ST 及*ST 股票 10%，宽基 ETF 20%，券商股 10%，其他 A 股 20%，现金 40%。

【输出要求】

只输出一个合法 JSON 对象，不要使用 Markdown 代码块，不要输出 JSON 之外的任何文字。必须严格采用以下结构：

```
{
 "decision_date": "2026-07-06",
 "st": {
  "action": "买入/卖出/持有",
  "initial_pct": 10,
  "target_pct": 0,
  "order_pct": -10
 },
 "broad_etf": {
  "action": "买入/卖出/持有",
  "initial_pct": 20,
  "target_pct": 0,
  "order_pct": -20
 },
```

```
  "broker": {
    "action": "买入/卖出/持有",
    "initial_pct": 10,
    "target_pct": 0,
    "order_pct": -10
  },
  "other_a_share": {
    "action": "买入/卖出/持有",
    "initial_pct": 20,
    "target_pct": 0,
    "order_pct": -20
  },
  "cash": {
    "initial_pct": 40,
    "target_pct": 0
  },
  "holding_period_days": 0,
  "information_staleness": "不适用",
  "decision_basis": [
    "依据 1",
    "依据 2",
    "依据 3"
  ],
  "main_risks": [
    "风险 1",
    "风险 2",
    "风险 3"
  ],
  "used_external_information": false
}
```

必须同时满足：

1、五项 target_pct 之和必须等于 100。

2、每类资产的 order_pct 必须等于 target_pct 减 initial_pct；正数为净买入，负数为净卖出，0 为不调仓。

3、action 必须与 order_pct 一致：正数填“买入”，负数填“卖出”，0 填“持有”。

4、holding_period_days 填写一个整数，范围为 1~90。

5、information_staleness 只能填“不适用”“低”“中”或“高”。本条件为实施首日，如认为不适用可填“不适用”。

6、decision_basis 和 main_risks 各写 3 条，每条不超过 60 个汉字。

7、不得把“目标仓位为 0%”解释为卖出，除非相对于给定初始仓位确实发生了减仓。

8、如你使用了本提示之外的任何事实或数据，必须将 used_external_information 设为 true；否则设为 false。

A.2. 条件 B：实施 21 天后情境

条件 B 与条件 A 仅有以下四处差异，其余文字、账户约束、政策信息、JSON 字段和输出规则完全一致：

1、【决策时点】替换为：“2026 年 7 月 27 日 14:50。该规则已于 2026 年 4 月 24 日公开发布，并自 2026 年 7 月 6 日起实施，至决策时点已实施 21 个自然日。本提示不提供这 21 天内的市场表现。”

2、JSON 示例中的 decision_date 由“2026-07-06”替换为“2026-07-27”。

3、JSON 示例中的 information_staleness 由“不适用”替换为“高”。

4、“必须同时满足”第 5 项替换为：“information_staleness 只能填‘不适用’‘低’‘中’或‘高’。请根据决策时点自行判断。”