

金工专题报告 20260110

深度学习系列之一：AI 重塑量化，基于大语言模型驱动的因子改进与情绪 Alpha 挖掘

2026 年 01 月 10 日

证券分析师 于明明

执业证书：S0600525120002

yumm@dwzq.com.cn

■ 本文是深度学习系列报告第一篇。本文系统性地构建了一套基于大语言模型 (LLM) 与提示工程 (Prompt Engineering) 的自动化因子研究框架，旨在探索 AI 在量化投资全链条中的应用潜力。本文首先在低频量价因子领域，通过人机交互的迭代循环，实现了对经典 Alpha158 因子库的深度优化，并进一步从“优化”范式升级为“生成”范式。随后，该自动化框架被成功迁移并应用于基本面和高频数据领域，系统性地挖掘并扩充了传统因子库，验证了 AI 在多数据源、多频率场景下的因子发现能力。最后，本文探索了 AI 在非结构化文本分析中的应用，利用 Gemini 大模型对海量调研纪要进行情绪解读，构建了独特的文本情绪指标，并将其作为增量信息有效融入最终的选股策略。

■ 该框架首先被应用于低频量价因子的挖掘与优化。本文以 Alpha158 因子库为基础进行优化实验，通过案例剖析 (如波动率因子 std20) 展示了 AI 能够识别原始因子逻辑缺陷并提出有效改进方案，且优化效果在 5 至 60 日的多个窗口期下均具备普适性。进而，通过为模型提供已验证的“成功案例”作为先验知识，本文实现了从零生成新因子的突破，成功挖掘出多个与样例因子相关性低、ICIR 在 0.8 以上的新因子。这些 AI 生成因子在样本外跟踪中表现持续稳健，部分因子的样本外 ICIR 能达到 1.0 以上，证明了该方法论的有效性与鲁棒性。

■ 将研究拓宽至基本面与高频领域后，AI 同样展现出强大的因子发现能力。在基本面维度，AI 不仅能生成经典因子的增强版本 (如现金毛利 CGP_TTM)，更能从留存收益 (REP_LF)、应收账款周转率 (ART_QR) 等新颖视角，对价值、质量、成长三类因子进行有效拓展与创新。在高频维度，通过赋予 AI 直接生成 Python 代码的能力，本文挖掘出一批逻辑新颖且表现优异的高频因子，其中部分强信号因子 (如投机波动因子) 的多空组合年化收益超过 60%。将 AI 高频因子库融入融合了日 K 与周 K 行情数据的 AGRU 神经网络模型后，模型的年化多头超额收益由 18.24% 显著提升至 25.28%，RankIC 均值提升了 0.71 个百分点。

■ 本文还探索了 AI 在另类数据处理上的潜力，利用 Gemini 2.5 Pro 大模型对近百万字的上市公司调研纪要进行深度解析，并通过双速动态衰减模型构建了周度情绪因子。研究发现，该因子呈现出独特的非对称预测能力，即正面情绪与股价上涨关系不强，但负面情绪是未来股价下跌的强预警信号，其空头组合年化超额收益达 8.26%，显著优于传统因子。更重要的是，该情绪因子与传统量价及基本面因子的相关性极低，可作为独立且有效的补充信息源。

■ 最终，本文打造了一个多维信息融合的综合策略：将 AI 挖掘的高频因子与低频行情数据融合进 AGRU 神经网络，形成核心 Alpha；再利用 AI 文本情绪因子对该核心 Alpha 进行空头端风险调整，构建了最终的中证 800 指数增强策略。与未加入调研因子调整的策略相比，最终策略在保持换手率基本不变的情况下，年化超额收益由 11.15% 提升至 11.81%，信息比率由 2.18 提升至 2.31，且近三年超额提升皆超过 1 个点，进一步验证了 AI 在量化研究多环节赋能并实现“1+1>2”效果的巨大潜力。

■ 风险提示：

- 1) 结论基于历史数据，在市场环境转变时模型存在失效的风险。本报告所有结论均基于历史数据的回测分析，历史规律不代表未来，当市场风格、投资者结构、宏观环境等发生重大变化时，报告中构建的因子及模型可能失效。
- 2) 模型过拟合风险。本报告构建的因子及深度学习模型，可能对历史数据存在过度拟合的问题，即模型可能学习到了历史数据中的噪音而非真实的规律，从而在未来市场的实际应用中表现不佳。
- 3) 大语言模型（LLM）自身局限性的风险。本报告的研究依赖于特定版本的大语言模型，且大语言模型的输出具有随机性，同时可能存在模型幻觉问题，导致报告结果无法复现。
- 4) 回测设置与实际存在差异的风险。报告中的回测分析基于一系列理想化的假设（如特定的成交价格、固定的交易费用、未考虑市场冲击等），实际交易环境中的成本、滑点和各种限制可能导致策略的真实表现不及预期。
- 5) 因子失效风险。报告中通过 AI 挖掘的因子，其有效性可能随着时间的推移而衰减。特别是在因子逻辑被市场广泛认知或被大量资金使用后，其获取超额收益的能力可能会显著下降。

内容目录

1. 深度学习系列之一：AI 重塑量化，基于大语言模型驱动的因子改进与情绪 Alpha 挖掘	8
2. 低频量价因子的优化改进与生成	9
2.1. Alpha158 因子体系解析与优化框架	9
2.1.1. Qlib 算子与 Alpha158 因子集概述	9
2.1.2. 动态交互框架：基于 Prompt Engineering 的因子优化流程设计	12
2.1.3. 优化效果验证：RankIC 的跨周期普适性与稳定性突破	17
2.1.4. 以波动率因子为例：洞察模型优化方向	22
2.1.5. 改进波动率因子样本外跟踪：持续稳健且优于原始波动因子	27
2.2. 从优化到创造：基于大语言模型的新因子生成范式	28
2.2.1. 从零生成的困境：独立探索下的效率与效果瓶颈	28
2.2.2. 案例驱动的突破：基于先验知识的低相关性因子挖掘	31
2.2.3. AI 生成因子样本外跟踪：IC 出色超额上行	36
3. AI 基本面因子挖掘：构建自动化研究框架	38
3.1. 实验设计：基础数据与处理算子	38
3.2. 人机交互框架：AI 因子生成流程详解	41
3.3. AI 价值因子挖掘：经典框架的有效拓展与创新	44
3.4. AI 质量因子挖掘：从盈利能力到运营效率的多维度探索	51
3.5. AI 成长因子挖掘：探索盈利增长的多维定义	57
4. AI 驱动的高频因子挖掘：从分钟数据到选股信号	61
4.1. 方法论构建：高频因子自动生成的工作流	61
4.2. AI 高频因子库：整体表现与分类解析	65
4.3. 高频因子案例剖析：从逻辑到代码至效果	68
4.3.1. 投机波动因子 (speculative_frenzy_instability)	68
4.3.2. 极端加速度方差占比因子 (extreme_gamma_burst_ratio)	72
4.3.3. 动量与加速度相关性因子 (momentum_acceleration_corr)	75
4.3.4. 收益与价格偏离的交互效应因子 (dual_stress_rolling_idio_ret_dev_product)	78
4.4. AI 高频因子的增量提升	81
4.5. 高频因子对多频率融合的提升	83
4.5.1. 数据处理与模型参数	83
4.5.2. 以日 K 与周 K 为特征的 AGRU	85
4.5.3. 以日 K+周 K+高频因子为特征的 AGRU	88
4.5.4. AGRU 融合日 K+周 K+高频因子的周频 800 指增	91
5. AI 文本分析：对量化选股的增益	93
5.1. 通过 Gemini 模型读取调研文本情绪	93
5.1.1. 近年来调研数量与字数有所提升	93
5.1.2. 综合考虑调研情况，选用 Gemini 2.5 Pro 模型	94
5.1.3. 打分结果展示	95
5.2. 从调研评分到调研因子：双速动态衰减模型	96
5.3. 调研因子特色：独特的空头规避能力	97
5.4. 调研因子对成熟指增框架的再提升	100
6. AI 赋能量化研究：范式优势与当前局限	103
7. 风险提示	105

图表目录

图 1: 因子优化流程示意图.....	14
图 2: std20 原始因子与第一次改进因子累计 RankIC	23
图 3: std20 原始因子与第一次改进因子多头超额净值.....	23
图 4: std20 原始因子与第一二次改进因子累计 RankIC	24
图 5: std20 原始因子与第一二次改进因子多头超额净值.....	24
图 6: std20 原始因子与第一二三次改进因子累计 RankIC	25
图 7: std20 原始因子与第一二三次改进因子多头超额净值.....	25
图 8: std20 原始因子与第二次第三次改进因子样本内与样本外累计 RankIC	27
图 9: std20 原始因子与第二次第三次改进因子样本内与样本外多空超额收益.....	27
图 10: 新生成因子累计 RankIC	35
图 11: 新生成因子多头超额净值.....	35
图 12: 生成因子样本内与样本外累计 RankIC	36
图 13: 生成因子样本内与样本外多空超额收益.....	36
图 14: 基本面因子生成流程示意图.....	41
图 15: REP_LF 因子每期 RankIC 与累计 RankIC.....	48
图 16: REP_LF 因子多头组合净值与多头超额净值.....	48
图 17: REP_LF 因子多空超额净值.....	49
图 18: CGP_TTM 因子每期 RankIC 与累计 RankIC	50
图 19: CGP_TTM 因子多头组合净值与多头超额净值	50
图 20: CGP_TTM 因子多空超额净值	51
图 21: ART_QR 因子多空超额净值	57
图 22: 高频因子生成流程示意图.....	62
图 23: 投机波动因子每期 RankIC 与累计 RankIC.....	70
图 24: 投机波动因子多头组合净值与多头超额净值.....	70
图 25: 投机波动因子多空超额净值.....	71
图 26: 极端加速度方差占比因子每期 RankIC 与累计 RankIC.....	73
图 27: 极端加速度方差占比因子多头组合净值与多头超额净值.....	74
图 28: 极端加速度方差占比因子多空超额净值.....	74
图 29: 动量加速度相关因子每期 RankIC 与累计 RankIC.....	76
图 30: 动量加速度相关因子多头组合净值与多头超额净值.....	77
图 31: 动量加速度相关因子多空超额净值.....	77
图 32: 收益偏离交互因子每期 RankIC 与累计 RankIC.....	79
图 33: 收益偏离交互因子多头组合净值与多头超额净值.....	80
图 34: 收益偏离交互因子多空超额净值.....	80
图 35: 样例高频因子叠加 AI 高频因子 RankIC 提升	81
图 36: 样例高频因子叠加 AI 高频因子多头超额提升	82
图 37: 样例高频因子叠加 AI 高频因子多空超额提升	82
图 38: 回测路径合并方式示意图.....	84
图 39: 以日 K+周 K 为特征的 AGRU 融合模型	86
图 40: Attention 模块内部结构	86

图 41:	日 K+周 K 因子每期 RankIC 与累计 RankIC	87
图 42:	日 K+周 K 因子多头组合净值与多头超额净值	87
图 43:	以日 K+周 K+高频因子为特征的 AGRU 模型	89
图 44:	Attention 模块内部结构	89
图 45:	日 K+周 K+高频因子每期 RankIC 与累计 RankIC	90
图 46:	日 K+周 K+高频因子多头组合净值与多头超额净值	90
图 47:	AGRU 融合日 K+周 K+高频因子的周频 800 指增	92
图 48:	调研文本因子在中证 800 中 RankIC	97
图 49:	调研文本因子在中证 800 中分层超额收益	97
图 50:	调研文本因子在中证 800 中分层年化超额收益	97
图 51:	调研文本因子各组数量	98
图 52:	调研文本因子在中证 800 中按得分分层超额收益	98
图 53:	调研文本因子在中证 800 中按得分分层年化超额收益	98
图 54:	调研文本因子与其它因子空头超额收益	99
图 55:	调研文本因子与其它因子的因子值相关性	99
图 56:	AGRU 融合日 K+周 K+高频因子（空头调整）的周频 800 指增	100
图 57:	原始 800 指增与空头调整 800 指增收益提升对比	101
图 58:	调研因子(50 分以下)调整分数与受影响股票数量（单位：个）	102

表 1:	Alpha 158 因子列表	9
表 2:	Qlib 部分算子	11
表 3:	原始因子与改进因子 RankIC 均值和 ICIR	18
表 4:	不同因子类型下原始表达式与增强表达式 RankIC 均值	21
表 5:	Beta20 原始因子与历次优化因子表达式及 IC 统计量	25
表 6:	min20 原始因子与历次优化因子表达式及 IC 统计量	26
表 7:	qtlu20 原始因子与历次优化因子表达式及 IC 统计量	26
表 8:	std20 原始因子与第二次第三次改进因子样本内与样本外选股效果统计	27
表 9:	历史模型生成的因子表达式与 IC 统计量	29
表 10:	从 Alpha158 生成的因子表达式与 IC 统计量	32
表 11:	新因子与 ICIR 大于 0.7 的样例因子之间的相关性	34
表 12:	新因子之间相关性	34
表 13:	生成因子样本内与样本外选股效果统计	36
表 14:	财务数据字段名称与字段释义	38
表 15:	基本面因子算子名称与释义	40
表 16:	样例价值因子释义与表达式	44
表 17:	样例价值因子统计指标	44
表 18:	AI 价值因子释义与表达式	45
表 19:	AI 价值因子统计指标	46
表 20:	AI 价值因子与样例价值因子的因子值秩相关性	47
表 21:	REP_LF 因子 IC 统计指标	48
表 22:	REP_LF 因子分年度多头超额收益风险绩效指标	48
表 23:	REP_LF 因子分年度多空超额收益风险绩效指标	49
表 24:	CGP_TTM 因子 IC 统计指标	50

表 25:	CGP_TTM 因子分年度多头超额收益风险绩效指标	50
表 26:	CGP_TTM 因子分年度多空超额收益风险绩效指标	51
表 27:	样例质量因子释义与表达式	51
表 28:	样例质量因子统计指标	52
表 29:	AI 质量因子释义	52
表 30:	AI 质量因子统计指标	54
表 31:	GPS_QR、COPPS_QR 因子与样例质量因子秩相关性	56
表 32:	GPS_QR、COPPS_QR 因子与 EPS_QR 因子统计指标对比	56
表 33:	ART_QR 因子分年度多空超额收益风险绩效指标	57
表 34:	ART_QR 因子与样例质量因子秩相关性	57
表 35:	样例成长因子释义与表达式	58
表 36:	样例成长因子统计指标	58
表 37:	AI 成长因子释义与表达式	58
表 38:	AI 成长因子统计指标	59
表 39:	预置分域函数名称与释义	61
表 40:	样例高频因子列表与 IC 统计值	61
表 41:	AI 高频因子统计指标	65
表 42:	投机波动因子 IC 统计指标	70
表 43:	投机波动因子分年度多头超额收益风险绩效指标	70
表 44:	投机波动因子分年度多空超额收益风险绩效指标	71
表 45:	极端加速度方差占比因子 IC 统计指标	73
表 46:	极端加速度方差占比因子分年度多头超额绩效指标	74
表 47:	极端加速度方差占比因子分年度多空超额绩效指标	74
表 48:	动量加速度相关因子 IC 统计指标	76
表 49:	动量加速度相关因子分年度多头超额绩效指标	77
表 50:	动量加速度相关因子分年度多空超额绩效指标	77
表 51:	收益偏离交互因子 IC 统计指标	79
表 52:	收益偏离交互因子分年度多头超额绩效指标	80
表 53:	收益偏离交互因子分年度多空超额绩效指标	80
表 54:	样例高频因子叠加 AI 高频因子 RankIC 与 ICIR 提升	81
表 55:	样例高频因子叠加 AI 高频因子多头超额提升	82
表 56:	样例高频因子叠加 AI 高频因子多空超额提升	82
表 57:	日 K+周 K 因子 IC 统计指标	87
表 58:	日 K+周 K 因子分年度多头超额绩效指标	87
表 59:	日 K+周 K+高频因子 IC 统计指标	90
表 60:	日 K+周 K+高频因子分年度多头超额绩效指标	90
表 61:	中证 800 指增组合收益风险特征指标	92
表 62:	中证 800 指增组合超额收益指标	92
表 63:	调研数量与字数	93
表 64:	全市场每月调研数量 (个)	93
表 65:	全市场每月调研字数(万)	94
表 66:	大模型参数设定	95
表 67:	大模型运行时间及输入输入文本长度	96
表 68:	调研文本因子与其它因子分年度空头超额收益	99

表 69: 调研文本因子与其它因子分年度收益波动比.....99

表 70: 中证 800 指增组合(空头调整)收益风险特征指标.....101

表 71: 中证 800 指增组合(空头调整)超额收益指标.....101

表 72: 原始 800 指增与空头调整 800 指增收益提升对比.....101

1. 深度学习系列之一：AI 重塑量化，基于大语言模型驱动的因子改进与情绪 Alpha 挖掘

在传统的量化因子研究范式中，研究员的先验知识、领域经验以及海量的试错实验构成了因子发现的核心驱动力。这一过程不仅研发周期长、人力成本高，且往往受限于研究员既有的认知框架与方法论，难以产生真正颠覆性的创新。近年来，以大语言模型（LLM）为代表的人工智能技术取得了革命性进展，其强大的自然语言理解、逻辑推理与代码生成能力，为重塑量化研究的全流程提供了前所未有的机遇。

在将这一前沿技术引入严谨的量化投资体系之前，一系列根本性的问题亟待我们深入探索与解答：

在经典的低频量价因子领域，对于 Alpha158 这类成熟的因子库，AI 能否超越简单的参数调优，深入其内在逻辑进行深度重构与创新？更进一步，我们能否摆脱对初始表达式的依赖，让 AI 真正实现从零到一的创造？AI 生成的因子与人类专家的发现有何异同，能否有效挖掘出与现有体系低相关性的新 Alpha 源？

在传统上高度依赖财务专长的基本面研究领域，AI 能否自主学习并理解财务报表各项科目间的复杂勾稽关系？它能否突破市盈率、市净率等经典框架的束缚，通过对利润、现金流、资产等项目的创新性组合，构建出对企业价值、质量与成长性更深刻、更多维度的刻画？我们能否构建一个自动化的 AI 基本面研究框架？

面对数据量巨大、信噪比极低的高频分钟级行情，AI 是否具备直接编写高效、稳健的 Python 代码以进行复杂时序与截面计算的能力？这些捕捉瞬时市场微观结构的因子，又该如何与处理中低频趋势的深度学习模型（如 AGRU）进行有效融合，最终为策略带来显著的增量信息？

跳出纯粹的数值数据，AI 能否深入非结构化的文本海洋，例如海量的上市公司调研纪要，并精准捕捉其中隐藏的、难以量化的情绪与预期？这种“情绪 Alpha”的信号特性是什么？是独特的空头规避能力，还是对称的多空预测力？其信号的衰减周期有多长，与传统因子又呈现怎样的相关性？

当 AI 在量价、基本面、高频、文本等多个维度都贡献了有效的 Alpha 因子后，我们应如何设计一个统一的框架，将这些来源多样、逻辑各异的 AI 生成信号进行有机整合？最终构建出的综合策略，其实战效果（如指数增强）如何？是否真正实现了“1+1>2”的协同效应？

本篇报告作为深度学习系列的第一篇，将以构建一个可落地、可迭代的 AI 驱动研究体系为核心，围绕上述问题展开系统性的实证研究，旨在全面展示大语言模型如何重塑因子挖掘与策略构建的全过程，并量化其在现代量化投资中的实际增益。

2. 低频量价因子的优化改进与生成

2.1. Alpha158 因子体系解析与优化框架

2.1.1. Qlib 算子与 Alpha158 因子集概述

对于原始因子的积累，我们借助 Github 上的开源项目 Qlib，采用其所集成的 Alpha158 量价因子开展了模型层面的基础研究。Alpha158 量价因子是一个丰富且多样化的因子集合，它涵盖了基于价格和成交量等多维度信息构建的多种因子。从因子构成来看，若不考虑窗口期，该因子集包含 42 个基础因子。而当纳入窗口期考量时，部分因子综合考虑了 5 日、10 日、20 日、30 日以及 60 日等不同长度的交易日窗口期，由此，Alpha158 因子集扩展为共计 158 个因子。为了更系统地理解和分析这些因子，我们依据其特征与计算逻辑，将 Alpha158 因子细致地划分为以下 5 大类别：

K 线因子：这些因子仅使用当天的开盘价、收盘价、最高价、最低价以及均价数据，共包含 13 个因子。这些因子捕捉了市场在单个交易日内的波动和变化特征。

波动因子：波动因子主要衡量股票价格的波动性，共包含 5 个因子。波动性因子可以帮助我们理解股票价格的变动幅度和频率，从而更好地评估风险和收益。

价因子：价因子是基于股票的价格信息计算得出的，共包含 100 个因子。这些因子涉及多种价格计算方法和统计指标，能够反映出股票价格的长期和短期趋势。

量因子：量因子基于成交量数据，共包含 30 个因子。成交量是市场交易活动的直接反映，量因子能够提供关于市场流动性和投资者行为的重要信息。

量价相关性因子：这些因子同时考虑了成交量和价格的关系，共包含 10 个因子。通过分析量价关系，可以更深入地了解市场的供需动态和价格变动的内在驱动力。

表1: Alpha 158 因子列表

因子大类	因子名称	窗口参数	因子释义	算式
K 线	HIGH0		最高价除以收盘价	$\$high/\$close$
	KLEN		K 线长度	$(\$high-\$low)/\$open$
	KLOW1		下影线长度相对开盘价的比例	$(\$open-\$low)/\$open$
	KLOW2		下影线长度相对 K 线整体的比例	$(\$open-\$low)/(\$high-\$low+1e-12)$
	KMID1		K 线实体长度相对开盘价的比例	$(\$close-\$open)/\$open$
	KMID2		K 线实体长度相对 K 线整体的比例	$(\$close-\$open)/(\$high-\$low+1e-12)$
	KSFT1		收盘价在整个价格区间（最高价到最低价）中的位置，相对开盘价的比例	$(\$close-\$low)/(\$high-\$low+1e-12)$
	KSFT2		收盘价在整个价格区间（最高价到最低价）中的位置，相对 K 线整体的比例	$(\$close-\$low)/(\$high-\$low+1e-12)$
	KUP1		上影线长度相对开盘价的比例	$(\$high-\$close)/\$open$
	KUP2		上影线长度相对 K 线整体的比例	$(\$high-\$close)/(\$high-\$low+1e-12)$

	LOW0		最低价除以收盘价	$\$low/\$close$
	OPEN0		开盘价除以收盘价	$\$open/\$close$
	VWAP0		均价除以收盘价	$\$vwap/\$close$
波动	STD	[5, 10, 20, 30, 60]	过去 d 天的收盘价标准差，除以最新的收盘价以去除单位。	$Std(\$close, \%d)/\$close$
	BETA	[5, 10, 20, 30, 60]	过去 d 天的收盘价变化率，除以最新的收盘价以去除单位。例如，过去 d 天每天价格上涨 10 美元，那么斜率将是 10。	$Slope(\$close, \%d)/\$close$
价	CNTD	[5, 10, 20, 30, 60]	过去上涨天数与过去下跌天数之间的差异。	$Mean(\$close > Ref(\$close, 1), \%d) - Mean(\$close < Ref(\$close, 1), \%d)$
	CNTN	[5, 10, 20, 30, 60]	过去 d 天内价格下跌的天数百分比。	$Mean(\$close < Ref(\$close, 1), \%d)$
	CNTP	[5, 10, 20, 30, 60]	过去 d 天内价格上涨的天数百分比。	$Mean(\$close > Ref(\$close, 1), \%d)$
	IMAX	[5, 10, 20, 30, 60]	当前日期与之前最高价日期之间的天数，属于 Aroon 指标的一部分。该指标衡量一段时间内高点之间和低点之间的时间。强劲的上升趋势会定期看到新的高点，而强劲的下降趋势会定期看到新的低点。	$IdxMax(\$high, \%d)/\%d$
	IMIN	[5, 10, 20, 30, 60]	当前日期与之前最低价日期之间的天数，属于 Aroon 指标的一部分。	$IdxMin(\$low, \%d)/\%d$
	IMXD	[5, 10, 20, 30, 60]	之前最低价日期出现在最高价日期之后的时间段。较大的值表示向下的动量。	$(IdxMax(\$high, \%d) - IdxMin(\$low, \%d))/\%d$
	MA	[5, 10, 20, 30, 60]	简单移动平均线，过去 d 天的简单移动平均线，除以最新的收盘价以去除单位。	$Mean(\$close, \%d)/\$close$
	MAX	[5, 10, 20, 30, 60]	过去 d 天的最高价，除以最新的收盘价以去除单位。	$Max(\$high, \%d)/\$close$
	MIN	[5, 10, 20, 30, 60]	过去 d 天的最低价，除以最新的收盘价以去除单位。	$Min(\$low, \%d)/\$close$
	QTLT	[5, 10, 20, 30, 60]	过去 d 天收盘价的 20% 分位数，除以最新的收盘价以去除单位。	$Quantile(\$close, \%d, 0.2)/\$close$
	QTLU	[5, 10, 20, 30, 60]	过去 d 天收盘价的 80% 分位数，除以最新的收盘价以去除单位。	$Quantile(\$close, \%d, 0.8)/\$close$
	RANK	[5, 10, 20, 30, 60]	当前收盘价在过去 d 天收盘价中的百分位数，表示当前价格水平相对于过去 d 天的比较，提供额外的信息给移动平均线。	$Rank(\$close, \%d)$
	RESI	[5, 10, 20, 30, 60]	过去 d 天线性回归的残差，表示过去 d 天的趋势线性程度。	$Resi(\$close, \%d)/\$close$
	ROC	[5, 10, 20, 30, 60]	变化率，过去 d 天的价格变化，除以最新的收盘价以去除单位。	$Ref(\$close, \%d)/\$close$
	RSQR	[5, 10, 20, 30, 60]	过去 d 天线性回归的 R 平方值，表示趋势的线性程度。	$Rsquare(\$close, \%d)$
	RSV	[5, 10, 20, 30, 60]	表示过去 d 天内价格在上下阻力价格之间的位置。	$(\$close - Min(\$low, \%d)) / (Max(\$high, \%d) - Min(\$low, \%d) + 1e-12)$

量	SUMD	[5, 10, 20, 30, 60]	总收益与总损失之间的差异比率，类似于 RSI 指标。	$(\text{Sum}(\text{Greater}(\$close - \text{Ref}(\$close, 1), 0), \%d) - \text{Sum}(\text{Greater}(\text{Ref}(\$close, 1) - \$close, 0), \%d)) / (\text{Sum}(\text{Abs}(\$close - \text{Ref}(\$close, 1)), \%d) + 1e-12)$
	SUMN	[5, 10, 20, 30, 60]	总损失与绝对总价格变化的比率，可以通过 $\text{SUMN} = 1 - \text{SUMP}$ 得到，类似于 RSI 指标。	$\text{Sum}(\text{Greater}(\text{Ref}(\$close, 1) - \$close, 0), \%d) / (\text{Sum}(\text{Abs}(\$close - \text{Ref}(\$close, 1)), \%d) + 1e-12)$
	SUMP	[5, 10, 20, 30, 60]	总收益与绝对总价格变化的比率，类似于 RSI 指标。	$\text{Sum}(\text{Greater}(\$close - \text{Ref}(\$close, 1), 0), \%d) / (\text{Sum}(\text{Abs}(\$close - \text{Ref}(\$close, 1)), \%d) + 1e-12)$
	WVMA	[5, 10, 20, 30, 60]	交易量加权的价格变化波动率。	$\text{Std}(\text{Abs}(\$close / \text{Ref}(\$close, 1) - 1) * \$volume, \%d) / (\text{Mean}(\text{Abs}(\$close / \text{Ref}(\$close, 1) - 1) * \$volume, \%d) + 1e-12)$
	VMA	[5, 10, 20, 30, 60]	简单交易量移动平均线。	$\text{Mean}(\$volume, \%d) / (\$volume + 1e-12)$
	VSTD	[5, 10, 20, 30, 60]	过去 d 天的交易量标准差。	$\text{Std}(\$volume, \%d) / (\$volume + 1e-12)$
	VSUMD	[5, 10, 20, 30, 60]	总交易量增加与总交易量减少之间的差异比率，类似于交易量的 RSI 指标。	$(\text{Sum}(\text{Greater}(\$volume - \text{Ref}(\$volume, 1), 0), \%d) - \text{Sum}(\text{Greater}(\text{Ref}(\$volume, 1) - \$volume, 0), \%d)) / (\text{Sum}(\text{Abs}(\$volume - \text{Ref}(\$volume, 1)), \%d) + 1e-12)$
	VSUMP	[5, 10, 20, 30, 60]	总交易量增加与绝对总交易量变化的比率。	$\text{Sum}(\text{Greater}(\$volume - \text{Ref}(\$volume, 1), 0), \%d) / (\text{Sum}(\text{Abs}(\$volume - \text{Ref}(\$volume, 1)), \%d) + 1e-12)$
	VSUMN	[5, 10, 20, 30, 60]	总交易量减少与绝对总交易量变化的比率，可以通过 $\text{VSUMN} = 1 - \text{VSUMP}$ 得到。	$\text{Sum}(\text{Greater}(\text{Ref}(\$volume, 1) - \$volume, 0), \%d) / (\text{Sum}(\text{Abs}(\$volume - \text{Ref}(\$volume, 1)), \%d) + 1e-12)$
	VSUMD	[5, 10, 20, 30, 60]	总交易量增加与总交易量减少之间的差异比率，类似于交易量的 RSI 指标。	$(\text{Sum}(\text{Greater}(\$volume - \text{Ref}(\$volume, 1), 0), \%d) - \text{Sum}(\text{Greater}(\text{Ref}(\$volume, 1) - \$volume, 0), \%d)) / (\text{Sum}(\text{Abs}(\$volume - \text{Ref}(\$volume, 1)), \%d) + 1e-12)$
量价相关性	CORD	[5, 10, 20, 30, 60]	价格变化率与交易量变化率之间的相关性。	$\text{Corr}(\$close / \text{Ref}(\$close, 1), \text{Log}(\$volume / \text{Ref}(\$volume, 1) + 1), \%d)$
	CORR	[5, 10, 20, 30, 60]	绝对收盘价与对数交易量之间的相关性。	$\text{Corr}(\$close, \text{Log}(\$volume + 1), \%d)$

数据来源：Qlib、东吴证券研究所

表2: Qlib 部分算子

算子	算子释义
Ref	变量检索，N=0，检索第一个数据；N>0，检索 N 个周期前的数据；N<0，检索未来数据
Max	单变量滚动 N 个窗口期内最大值
Min	单变量滚动 N 个窗口期内最小值
Sum	单变量滚动 N 个窗口期求和
Mean	单变量滚动 N 个窗口期均值
Std	单变量滚动 N 个窗口期标准差
Var	单变量滚动 N 个窗口期方差
Skew	单变量滚动 N 个窗口期偏度
Kurt	单变量滚动 N 个窗口期峰度
Med	单变量滚动 N 个窗口期中位数
Mad	单变量滚动 N 个窗口期内和均值偏离的绝对值
Slope	单变量与 T(1,2,3...) 的滚动回归的回归系数项

Rsquare	单变量与 T(1,2,3...)的滚动回归的 R 方
Resi	单变量与 T(1,2,3...)的滚动回归的残差
Rank	单变量滚动 N 个窗口期排名
Quantile	单变量滚动 N 个窗口期百分位
Count	单变量滚动 N 个窗口期内非空数值
EMA	单变量滚动 N 个窗口期的指数移动平均
WMA	单变量滚动 N 个窗口期的加权移动平均
Corr	两个变量在滚动 N 个窗口期的相关性
Cov	两个变量在滚动 N 个窗口期的协方差
Delta	单变量在滚动 N 个窗口期的最后值减开始值
Abs	单变量的绝对值
Sign	单变量大于 0 的值置为 1，小于 0 的值置为 -1
Log	单变量的自然对数
Power	单变量的指数幂
Greater	比较两个变量，返回最大值
Less	比较两个变量，返回最小值
IdxMax	单变量滚动 N 个窗口期内最大值的索引
IdxMin	单变量滚动 N 个窗口期内最小值的索引
If	条件判断

数据来源：Qlib、东吴证券研究所

Qlib 中集成了大量基于 Cython 的算子，这些算子与高开低收均价成交量数据巧妙结合，共同构成了因子表达式。将因子表达式输入 Qlib 框架，即可高效地对因子进行高性能计算。在此背景下，如何巧妙运用 DeepSeek 对因子表达式进行改进，进而实现对因子选股能力的显著增强，成为本章着重研究的核心内容。

2.1.2. 动态交互框架：基于 Prompt Engineering 的因子优化流程设计

由于 Alpha158 量价因子支持多种窗口期的灵活计算，为确保研究结果的一致性与可比性，在测算因子 RankIC 均值与 ICIR 时，我们统一将窗口期参数设定为 20 个交易日。同时，明确以下的回测细节：

- 回测区间：2013 年 12 月 31 日至 2024 年 12 月 31 日。
- 剔除：剔除上市不满 365 个自然日的新股，剔除 ST 股。
- 中性化：对因子进行市值行业中性化
- 交易频率：周频调仓，以下周第一个交易日的 VWAP 价格成交，计算 VWAP 收益率的 RankIC 均值与 ICIR。

- 方向调整：根据 RankIC 均值的正负，对因子方向进行调整，使得 RankIC 为正，便于比较。

因子优化流程如下：

- 1) 对于每个因子，至少进行 3 次深度优化。在这 3 次优化过程中，让 DeepSeek 挖掘因子改进的潜力。若在 3 次优化后，优化因子的 RankIC 均值最大值达到原始因子的 1.5 倍，这意味着该因子已实现显著优化，此时将直接终止优化流程，并输出历次优化结果，以保留最佳优化路径与成果。
- 2) 若 3 次优化未能达到上述目标，即优化因子的 RankIC 均值未提升至原始因子的 1.5 倍，则继续进行优化。在整个优化过程中，最多尝试 5 次。这是为了在尽可能挖掘因子潜力的同时，避免过度优化导致的复杂性增加与收益递减，以及无谓的 token 消耗。若经过共 5 次尝试后仍未达标，则终止优化并输出历次优化结果。

图1：因子优化流程示意图



数据来源：东吴证券研究所绘制

如何巧妙编写 prompt 是实现与大语言模型高效交互的核心要点。在与 DeepSeek 模型的交互中，我们需要精准地告知模型任务的具体内容、可用的关键信息以及一些不容忽视的重要注意事项，以下是 prompt_init 的主干内容：

假如你是一位资深的量化选股因子专家，你将根据现有的截面日频量化选股因子的相关信息，对以下因子进行改进以提升其 RankIC 均值：

[[factor_algo]]

这个因子是{direction}因子。

可用变量：

\$open: 开盘价; \$close: 收盘价; \$high: 最高价; \$low: 最低价; \$vwap: 均价; \$volume: 成交量。

可用算子代码:

[{code_content}]

表达式支持不同的窗口期, 若窗口期为 20 个交易日, 调整因子方向并进行市值行业中性化后, 因子 2014 年以来周度 RankIC 均值为 {rankic}, ICIR 为 {icir}。

以提升因子的 RankIC 均值为目标, 对这个因子的表达式进行改进, 先列出至少 5 个改进方案, 比较每个方案可能对因子 IC 提升的潜力, 返回你认为最好的因子表达式及其优化逻辑, 因子表达式格式参考提供的样例。

注意以下几点:

1. 只使用提供的算子, 且保证调用方法正确, 可用算子见代码中 OpsList 变量。
2. 对因子进行正确的去量纲操作, 转换成比例的形式, 使得不同股票间可比。
3. 除了 Ref 外, 每个算子的窗口期相同, 在表达式中仍以 %d 表示。
4. 从逻辑出发进行改进, 不需要计算因子的 RankIC 或 ICIR 等指标。
5. 因子根本逻辑不能被改变, 例如波动类因子不能被改成动量类因子。

按照以下格式返回结果, 其中改进后因子表达式写在中括号内部, 不要换行, 表达式内部括号只能用小括号, 优化逻辑要分点罗列, 并对改进后的因子表达式进行解释:

"

改进后因子表达式: [表达式]

因子优化逻辑: [优化逻辑]

因子解释: [因子解释]

"

以上 prompt 主要在于实现以下目的:

- **任务与信息传达:** 我们首先将 AI 带入资深量化选股因子专家的角色, 简洁明了地阐述任务目标, 即根据现有的截面日频量化选股因子相关信息, 对特定因子进行改进, 以显著提升其 RankIC 均值。同时, 通过 [{factor_algo}] 明确原始因子表达式, 使 AI 清晰理解原始因子的量化逻辑; 借助 {direction} 变量说明因子方向, 帮助 AI 进一步区分不同类型因子, 如反转因子与动量因子, 从而为因子改进提供精准方向。

- **数据与算子说明：**详细告知 AI 可用变量，包括 open（开盘价）、close（收盘价）、high（最高价）、low（最低价）、vwap（均价）、volume（成交量），这些变量作为因子计算的基础数据，为模型改进因子表达式提供丰富素材。同时，将 Qlib 项目中的 ops.py 文件代码输入进大语言模型，以[{code_content}]的形式让 AI 清楚了解有哪些算子可用，以及每个算子的具体使用方法，为因子表达式的创新改进提供技术支持。
- **测试条件与评估告知：**向 AI 明确后文中将以 20 个交易日为窗口期进行测试，并且在测试过程中会对因子进行方向调整、市值行业中性化等预处理操作。同时，告知 AI 原始因子在特定条件下的 RankIC 均值与 ICIR，使其对原始因子的选股效果形成初步评估，从而在改进过程中有针对性地提升因子表现。
- **任务要求与思路引导：**再次强调以提升因子的 RankIC 均值为核心目标，要求对因子表达式进行改进。先列出至少 5 个改进方案，并对每个方案可能提升因子 IC 的潜力进行比较，最终返回认为最好的因子表达式及其优化逻辑，因子表达式格式参考提供的样例。这不仅为模型提供了清晰的任务要求，还通过推荐的初始思路引导 Deepseek 在思维链中进行系统、全面的初步思考。
- **注意事项明确：**为确保模型改进的准确性与有效性，我们还明确了一系列注意事项。例如，限定只使用提供的算子，且保证调用方法正确，防止大模型幻觉导致 AI 使用不存在的算子；提示对因子进行正确的去量纲操作，转换成比例形式，以保证不同股票间可比；统一除 Ref 外每个算子的窗口期以%d 表示，避免 AI 将某些算子的窗口期设置为其他值；要求从逻辑出发进行改进，避免 AI 因大模型幻觉“猜测”优化后的因子 IC 统计量；强调因子根本逻辑不能被改变，防止为追求高 RankIC 而改变因子本质类型。
- **返回格式约束：**按照特定格式要求 AI 返回结果，其中改进后因子表达式写在中括号内部，不得换行，表达式内部括号统一使用小括号。优化逻辑要分点罗列，并对改进后的因子表达式进行详细解释。这种严格的格式约束，方便了 Python 代码对返回内容进行准确解析。

基于上述针对 prompt_init 的设计思路，我们进一步构建了 prompt_opti。prompt_opti 主要用于收集模型历次优化结果，并据此告知模型继续优化。

以提升因子的 RankIC 均值为目标，比较之前表达式得到的 RankIC 均值与 ICIR，继续对这个因子的表达式进行改进，如果有必要的话可以推翻过去的方案重新思考。先列出至少 5 个改进方案，比较每个方案可能对因子 IC 提升的潜力，返回你认为最好的因子表达式及其优化逻辑，改进后的表达式需要不同于之前的表达式，因子表达式格式参考提供的样例。

按照以下格式返回结果，其中改进后因子表达式写在中括号内部，不要换行，表达式内部括号只能用小括号，优化逻辑要分点罗列，并对改进后的因子表达式进行解释：

"

改进后因子表达式：[表达式]

因子优化逻辑：[优化逻辑]

因子解释：[因子解释]

"

原始因子表达式为：{}，RankIC 均值为{}，ICIR 为{}，因子方向为{}；

第 1 次改进后因子表达式为{}，RankIC 均值为{}，ICIR 为{}，因子方向为{}；

第 2 次改进后因子表达式为{}，RankIC 均值为{}，ICIR 为{}，因子方向为{}；

第 3 次改进后因子表达式为{}，RankIC 均值为{}，ICIR 为{}，因子方向为{}；

.....

其主干内容设计思路为：首先以提升因子的 RankIC 均值为目标，引导模型比较之前表达式得到的 RankIC 均值与 ICIR，在此基础上继续对因子表达式进行改进。特别强调在必要情况下，可以推翻过去的方案重新思考，避免优化结果陷入局部最优解，确保模型能够持续探索更优的因子表达式。同时，通过收集原始因子和历次优化因子的表达式与预测效果，以详细罗列的方式（如原始因子表达式为：{}，RankIC 均值为{}，ICIR 为{}，因子方向为{}；第 1 次改进后因子表达式为{}，RankIC 均值为{}，ICIR 为{}，因子方向为{}.....）让 AI 全面、深入地分析过去的优化方向与效果，从而有针对性地进一步改进或重新思考因子表达式。通过 prompt_init 与 prompt_opti 的协同配合，我们构建了一个完整、高效的与大语言模型交互的体系，形成因子优化的底层框架。

2.1.3. 优化效果验证：RankIC 的跨周期普适性与稳定性突破

经过对 DeepSeek 改进后因子的测试，在设定的 5 次迭代范围内，29 个窗口期因子的表现呈现出积极态势。具体而言，共有 22 个因子的 RankIC 有所提升，其中 15 个因子的 RankIC 均值提升至 1.2 倍以上，10 个因子的 RankIC 均值更是提升至 1.5 倍以上，这些因子在选股能力的提升幅度上较为突出。

在衡量因子稳定性及预测能力的 ICIR 指标方面，同样有 23 个因子实现提升。其中 14 个因子的 ICIR 提升至 1.2 倍以上，10 个因子的 ICIR 提升至 1.5 倍以上。这一系列数据体现出 DeepSeek 对多数因子的优化卓有成效，切实增强了因子在选股策略中的有效

性与可靠性。

表3：原始因子与改进因子 RankIC 均值和 ICIR

因子类型	原始因子表达式	原始 RankIC 均值	原始 ICIR	改进次数	改进 RankIC 均值	改进 ICIR	RankIC 提升幅度	与原始因子相关性
std20	Std(\$close, %d)/\$close	4.03%	0.31	1	6.18%	0.42	2.14pct	69.30%
				2	7.80%	0.55	3.76pct	70.15%
				3	7.01%	0.79	2.98pct	6.21%
beta20	Slope(\$close, %d)/\$close	4.99%	0.45	1	4.49%	0.39	-0.49pct	72.75%
				2	3.66%	0.35	-1.32pct	-2.10%
				3	3.78%	0.38	-1.20pct	79.64%
				4	3.90%	0.38	-1.08pct	87.55%
				5	3.68%	0.36	-1.31pct	79.18%
cntd20	Mean(\$close>Ref(\$close, 1), %d)- Mean(\$close<Ref(\$close, 1), %d)	2.25%	0.29	1	3.99%	0.49	1.74pct	80.56%
				2	7.02%	0.70	4.77pct	40.43%
				3	1.23%	0.21	-1.02pct	-3.22%
cntn20	Mean(\$close<Ref(\$close, 1), %d)	2.07%	0.27	1	2.38%	0.39	0.31pct	9.99%
				2	4.59%	0.52	2.52pct	31.58%
				3	5.01%	0.54	2.94pct	24.05%
cntp20	Mean(\$close>Ref(\$close, 1), %d)	2.13%	0.29	1	3.30%	0.44	1.17pct	89.12%
				2	5.11%	0.56	2.98pct	55.19%
				3	4.20%	0.51	2.07pct	36.29%
imax20	IdxMax(\$high, %d)/%d	3.37%	0.44	1	1.48%	0.18	-1.89pct	36.99%
				2	2.04%	0.22	-1.33pct	48.64%
				3	1.18%	0.14	-2.20pct	-6.46%
				4	3.80%	0.51	0.43pct	48.55%
				5	3.16%	0.46	-0.22pct	41.35%
imin20	IdxMin(\$low, %d)/%d	2.59%	0.35	1	4.08%	0.53	1.50pct	-25.36%
				2	1.33%	0.17	-1.25pct	-39.09%
				3	5.11%	0.68	2.53pct	-5.69%
imxd20	(IdxMax(\$high, %d)-IdxMin(\$low, %d))/%d	3.86%	0.46	1	5.36%	0.53	1.50pct	75.29%
				2	1.67%	0.24	-2.19pct	-4.33%
				3	4.26%	0.53	0.40pct	76.22%
				4	2.85%	0.40	-1.02pct	65.02%
				5	5.59%	0.55	1.73pct	51.10%
ma20	Mean(\$close, %d)/\$close	6.89%	0.56	1	7.32%	0.57	0.43pct	97.36%
				2	4.60%	0.50	-2.30pct	46.88%
				3	4.62%	0.49	-2.27pct	50.04%
				4	6.04%	0.58	-0.85pct	42.40%
				5	1.89%	0.22	-5.00pct	21.53%
max20	Max(\$high, %d)/\$close	1.07%	0.09	1	1.89%	0.23	0.82pct	48.85%
				2	1.80%	0.23	0.73pct	11.34%

				3	1.88%	0.23	0.81pct	48.42%
min20	Min(\$low, %d)/\$close	7.48%	0.72	1	4.63%	0.38	-2.85pct	78.38%
				2	3.62%	0.29	-3.86pct	62.27%
				3	3.63%	0.49	-3.85pct	21.83%
				4	3.40%	0.38	-4.08pct	72.10%
				5	0.14%	0.03	-7.34pct	4.55%
qtd20	Quantile(\$close, %d, 0.2)/\$close	7.34%	0.66	1	7.37%	0.65	0.03pct	99.56%
				2	7.31%	0.57	-0.03pct	87.47%
				3	0.62%	0.11	-6.72pct	-0.08%
				4	7.31%	0.57	-0.03pct	87.14%
				5	7.08%	0.54	-0.26pct	77.67%
qtl20	Quantile(\$close, %d, 0.8)/\$close	5.73%	0.47	1	3.87%	0.45	-1.85pct	27.59%
				2	3.93%	0.56	-1.80pct	-4.98%
				3	1.24%	0.28	-4.48pct	4.75%
				4	3.80%	0.54	-1.93pct	-5.65%
				5	2.81%	0.49	-2.92pct	1.26%
rank20	Rank(\$close, %d)	5.64%	0.55	1	2.71%	0.33	-2.93pct	61.99%
				2	2.59%	0.32	-3.05pct	21.88%
				3	2.24%	0.34	-3.40pct	15.43%
				4	2.66%	0.32	-2.99pct	21.36%
				5	3.92%	0.44	-1.73pct	68.29%
resi20	Resi(\$close, %d)/\$close	3.68%	0.37	1	3.88%	0.40	0.19pct	81.93%
				2	3.87%	0.40	0.19pct	85.69%
				3	3.46%	0.42	-0.22pct	71.67%
				4	1.90%	0.25	-1.78pct	14.12%
				5	3.87%	0.41	0.19pct	88.68%
roc20	Ref(\$close, %d)/\$close	6.68%	0.56	1	1.07%	0.09	-5.61pct	52.04%
				2	6.89%	0.56	0.21pct	82.54%
				3	7.32%	0.57	0.63pct	83.12%
				4	3.68%	0.37	-3.00pct	-4.80%
				5	6.94%	0.57	0.26pct	83.06%
rsqr20	Rsquare(\$close, %d)	0.15%	0.02	1	0.04%	0.01	-0.12pct	-13.30%
				2	0.68%	0.12	0.53pct	-18.55%
				3	0.46%	0.09	0.31pct	-17.22%
rsv20	(\$close-Min(\$low, %d))/(Max(\$high, %d)-Min(\$low, %d)+1e-12)	4.51%	0.42	1	5.54%	0.51	1.03pct	91.54%
				2	5.62%	0.48	1.11pct	83.88%
				3	3.93%	0.40	-0.58pct	42.00%
				4	5.95%	0.51	1.44pct	80.91%
				5	5.82%	0.50	1.31pct	82.33%
sumd20	(Sum(Greater(\$close-Ref(\$close, 1), 0), %d)-Sum(Greater(Ref(\$close, 1)-\$close, 0), %d))/(Sum(Abs(\$close-Ref(\$close, 1)), %d)+1e-12)	5.24%	0.49	1	6.18%	0.69	0.94pct	31.87%
				2	4.42%	0.62	-0.81pct	15.96%
				3	3.31%	0.43	-1.92pct	39.40%
				4	5.26%	0.62	0.03pct	10.02%
				5	2.99%	0.56	-2.25pct	11.82%

sumn20	Sum(Greater(Ref(\$close, 1)-\$close, 0), %d)/(Sum(Abs(\$close-Ref(\$close, 1)), %d)+1e-12)	5.24%	0.49	1	1.23%	0.17	-4.01pct	-13.20%
				2	1.39%	0.18	-3.85pct	-14.97%
				3	0.91%	0.13	-4.32pct	-14.91%
				4	1.05%	0.15	-4.19pct	-13.00%
				5	0.91%	0.13	-4.32pct	-14.91%
sump20	Sum(Greater(\$close-Ref(\$close, 1), 0), %d)/(Sum(Abs(\$close-Ref(\$close, 1)), %d)+1e-12)	5.24%	0.49	1	5.94%	0.55	0.70pct	40.16%
				2	6.29%	0.54	1.06pct	86.48%
				3	3.56%	0.47	-1.68pct	27.56%
				4	3.10%	0.40	-2.13pct	-3.94%
				5	5.73%	0.59	0.50pct	60.27%
vma20	Mean(\$volume, %d)/(\$volume+1e-12)	3.61%	0.51	1	3.64%	0.52	0.03pct	94.00%
				2	3.27%	0.47	-0.34pct	-23.54%
				3	0.29%	0.06	-3.33pct	52.82%
				4	2.85%	0.47	-0.76pct	88.07%
				5	6.25%	0.80	2.64pct	3.93%
vstd20	Std(\$volume, %d)/(\$volume+1e-12)	0.59%	0.11	1	4.61%	0.77	4.02pct	-40.40%
				2	4.09%	0.72	3.50pct	-56.15%
				3	4.35%	0.75	3.76pct	-17.28%
vsumd20	(Sum(Greater(\$volume-Ref(\$volume, 1), 0), %d)-Sum(Greater(Ref(\$volume, 1)-\$volume, 0), %d))/(Sum(Abs(\$volume-Ref(\$volume, 1)), %d)+1e-12)	4.28%	0.67	1	4.17%	0.62	-0.11pct	36.50%
				2	2.42%	0.52	-1.86pct	-5.40%
				3	1.53%	0.31	-2.75pct	40.46%
				4	4.25%	0.70	-0.03pct	91.27%
				5	2.13%	0.39	-2.15pct	59.45%
vsumn20	Sum(Greater(Ref(\$volume, 1)-\$volume, 0), %d)/(Sum(Abs(\$volume-Ref(\$volume, 1)), %d)+1e-12)	4.28%	0.67	1	4.28%	0.67	0.00pct	99.99%
				2	5.59%	0.80	1.31pct	54.73%
				3	2.32%	0.54	-1.96pct	56.72%
				4	3.61%	0.52	-0.67pct	38.54%
				5	5.57%	0.64	1.30pct	49.60%
vsump20	Sum(Greater(\$volume-Ref(\$volume, 1), 0), %d)/(Sum(Abs(\$volume-Ref(\$volume, 1)), %d)+1e-12)	4.28%	0.67	1	5.22%	0.81	0.94pct	12.81%
				2	5.30%	0.81	1.02pct	12.94%
				3	5.67%	0.86	1.39pct	12.32%
				4	3.03%	0.43	-1.25pct	7.33%
				5	3.77%	0.58	-0.51pct	5.11%
wvma20	Std(Abs(\$close/Ref(\$close, 1)-1)*\$volume, %d)/(Mean(Abs(\$close/Ref(\$close, 1)-1)*\$volume, %d)+1e-12)	1.77%	0.32	1	3.42%	0.60	1.65pct	72.01%
				2	3.31%	0.60	1.54pct	69.17%
				3	3.07%	0.54	1.30pct	75.16%
cord20	Corr(\$close/Ref(\$close, 1), Log(\$volume/Ref(\$volume, 1)+1), %d)	5.09%	0.76	1	1.76%	0.40	-3.33pct	9.64%
				2	0.23%	0.05	-4.86pct	-3.55%
				3	1.00%	0.21	-4.09pct	-0.10%
				4	2.57%	0.65	-2.52pct	14.77%
				5	3.99%	0.59	-1.10pct	70.53%
corr20	Corr(\$close, Log(\$volume+1), %d)	4.58%	0.68	1	4.51%	0.62	-0.07pct	70.62%
				2	3.93%	0.75	-0.65pct	32.57%
				3	5.17%	0.77	0.59pct	50.06%

	4	5.25%	0.79	0.67pct	49.30%
	5	4.46%	0.64	-0.12pct	63.78%

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2024 年 12 月 31 日

以上测算基于周频调仓，并将表达式中的窗口期%d 设置为 20。为进一步探究 DeepSeek 优化效果的稳定性，我们思考若将窗口期设置为其他参数，优化出来的因子表达式是否仍能保持提升态势？为此，本文选取 Alpha158 中优化后 RankIC 均值有提升的因子，对原始表达式与改进后 20 日 RankIC 均值最高的表达式，在 5/10/20/30/60 个交易日的窗口期下测算其 RankIC 均值。

表4：不同因子类型下原始表达式与增强表达式 RankIC 均值

因子类型	表达式状态	5	10	20	30	60
std	原始	5.55%	5.19%	4.03%	3.42%	2.33%
	增强	8.43%	8.37%	7.79%	7.23%	5.95%
cntd	原始	3.04%	2.42%	2.25%	2.04%	1.55%
	增强	6.33%	6.40%	7.02%	6.92%	6.36%
cntn	原始	2.78%	2.17%	2.07%	1.82%	1.21%
	增强	3.86%	4.43%	5.01%	4.86%	4.26%
cntp	原始	3.01%	2.40%	2.13%	1.88%	1.42%
	增强	4.77%	4.62%	5.11%	5.04%	4.57%
imax	原始	3.47%	3.11%	3.37%	3.47%	2.67%
	增强	4.00%	3.68%	3.80%	3.91%	3.32%
imin	原始	2.83%	2.51%	2.59%	2.62%	1.76%
	增强	6.63%	6.16%	5.11%	4.62%	4.29%
imxd	原始	3.96%	3.63%	3.86%	3.89%	2.83%
	增强	5.48%	5.30%	5.59%	5.27%	3.60%
ma	原始	5.63%	6.17%	6.89%	7.19%	7.03%
	增强	5.95%	6.72%	7.32%	7.47%	7.31%
max	原始	0.14%	0.14%	1.07%	1.64%	2.14%
	增强	2.52%	2.35%	1.89%	2.30%	3.08%
qtld	原始	6.53%	6.98%	7.34%	7.57%	7.31%
	增强	6.61%	6.98%	7.37%	7.59%	7.31%
resi	原始	2.21%	3.51%	3.68%	4.06%	5.34%
	增强	2.63%	3.97%	3.88%	3.86%	4.69%
roc	原始	5.86%	5.99%	6.68%	6.60%	5.68%
	增强	5.95%	6.72%	7.32%	7.47%	7.31%
rsqr	原始	0.54%	0.47%	0.15%	0.34%	0.00%
	增强	0.84%	0.77%	0.68%	0.49%	0.07%
rsv	原始	4.39%	4.27%	4.51%	4.66%	4.63%
	增强	5.43%	5.78%	5.95%	5.82%	5.36%

sumd	原始	4.66%	4.66%	5.24%	5.23%	4.75%
	增强	5.33%	5.84%	6.18%	6.19%	5.94%
sump	原始	4.66%	4.66%	5.24%	5.23%	4.75%
	增强	5.40%	5.87%	6.29%	6.42%	6.31%
vma	原始	2.23%	2.84%	3.61%	4.14%	5.01%
	增强	6.21%	6.32%	6.25%	6.15%	5.93%
vstd	原始	1.68%	0.65%	0.59%	1.48%	3.15%
	增强	3.66%	4.26%	4.61%	4.43%	3.81%
vsumn	原始	2.69%	3.10%	4.28%	4.47%	4.97%
	增强	4.66%	4.95%	5.59%	5.42%	5.09%
vsump	原始	2.69%q	3.10%	4.28%	4.47%	4.97%
	增强	6.57%	6.27%	5.68%	5.31%	4.50%
wvma	原始	0.77%	1.21%	1.77%	1.74%	1.35%
	增强	2.79%	3.31%	3.42%	3.17%	2.35%
corr	原始	4.83%	5.25%	4.58%	4.01%	2.60%
	增强	4.33%	4.92%	5.25%	4.86%	4.10%

数据来源：Wind、东吴证券研究所

从表 4 数据可见，绝大部分原始因子表达式经 DeepSeek 优化后，不仅在 20 日窗口期下 RankIC 均值显著提升，在其他窗口期下同样展现出增强态势。这充分证明 AI 所优化的因子表达式在不同时间窗口下具备普适性，能够在多种市场时间尺度下，有效提升因子对股票收益的预测能力。

2.1.4. 以波动率因子为例：洞察模型优化方向

我们选取波动率因子 std20 展开深入剖析，以此探究 DeepSeek 对因子的改进方式，并明晰因子选股效果的提升究竟源于大语言模型的内在实力还是偶然因素。

std20 因子的原始表达式为“Std(\$close,%d)/\$close”，其含义明确，旨在计算过去 20 个交易日收盘价的标准差，并通过除以收盘价实现去量纲处理，以此衡量价格的波动程度。

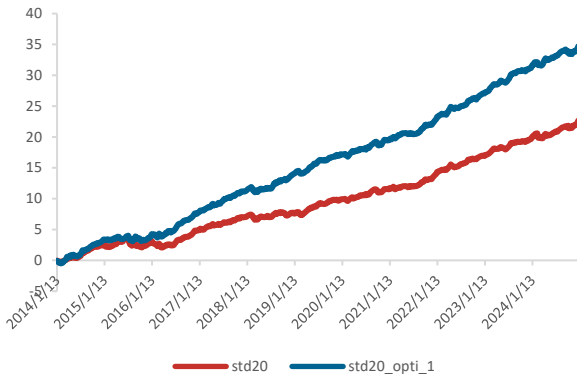
在第一次改进中，DeepSeek 将因子表达式修改为：

$$\text{Mean}(\text{Greater}(\text{Shigh}-\text{Slow}, \text{Greater}(\text{Abs}(\text{Shigh}-\text{Ref}(\text{Sclose},1)), \text{Abs}(\text{Slow}-\text{Ref}(\text{Sclose},1)))), \%d)/\text{Sclose}$$

该表达式在原始基础上，对分子进行了创新调整，引入了平均真实波幅 ATR 的概念。此 ATR 捕捉了价格波动中的日内波动、向上跳空和向下跳空三种模式，相较于单纯依赖收盘价标准差，能更为敏锐地识别价格剧烈波动的股票。分母依旧维持除以最新收盘价的形式，保持去量纲的操作。引入 ATR 计算波动率后，因子的 RankIC 均值从 4.03% 提升至 6.18%，ICIR 从 0.31 提升至 0.42，年化多头超额从 -2.23% 提升至 0.71%，年化

多空收益从 7.33%大幅提升至 17.42%。

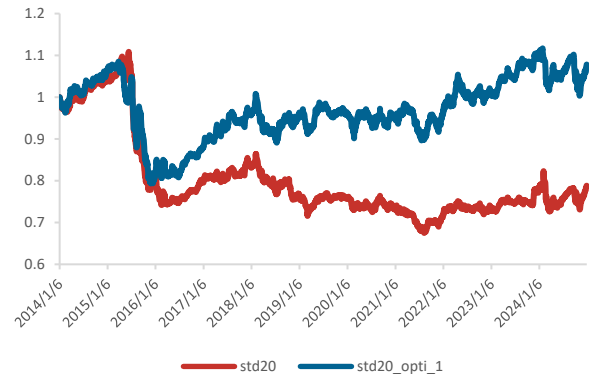
图2: std20 原始因子与第一次改进因子累计 RankIC



数据来源: Wind、东吴证券研究所

数据日期: 2014 年 1 月 13 日至 2024 年 12 月 31 日

图3: std20 原始因子与第一次改进因子多头超额净值



数据来源: Wind、东吴证券研究所

数据日期: 2014 年 1 月 6 日至 2024 年 12 月 31 日

在第二次改进中, DeepSeek 将因子表达式修改为:

$$\text{EMA}(\text{Greater}(\text{High}-\text{Low}, \text{Greater}(\text{Abs}(\text{High}-\text{Ref}(\text{Close}, 1)), \text{Abs}(\text{Low}-\text{Ref}(\text{Close}, 1)))) * \text{Volume}, \%d) / \text{EMA}(\text{Volume}, \%d) / \text{Close}$$

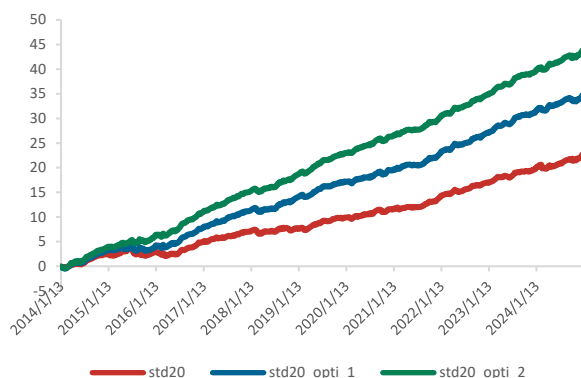
该表达式在第一次优化表达式的基础上, 主要实施了两点关键改进:

- 1) 引入成交量加权机制: 将真实波动幅度(TR)与当日成交量相乘, 强化量价共振效应, 高成交量伴随的波动更具信息含量, 通过 $\text{EMA}(\text{Volume}, \%d)$ 消除成交量绝对值影响, 构建单位成交量波动比率。
- 2) 采用 EMA 双重平滑: 对分子分母同时进行指数加权平均, 既保留成交量加权特性又加强近期数据的权重。

DeepSeek 对该因子的逻辑解释为: “该因子通过成交量加权的指数移动平均真实波动率, 捕捉资金流动驱动的价格不稳定性。相比简单平均 TR, 成交量加权能识别主力资金参与的异常波动, EMA 处理强化了近期市场情绪的敏感性, 双重 EMA 标准化有效剥离了量价纲差异。该设计同时满足波动测量的全面性、量价关系的协同性以及市场记忆衰减的特性, 预期能更精准识别高投机风险个股”, 可以看出模型对该因子的理解基本无误。

经此改进, 因子的 RankIC 均值由第一次优化的 6.18%提升至 7.8%, ICIR 由 0.42 提升至 0.55, 年化多头超额从 0.71%提升至 5.24%, 年化多空收益由 17.42%提升至 33.81%。从相关对比图中能够明显观察到, 该波动率因子在 2015 年的回撤得到显著改善。

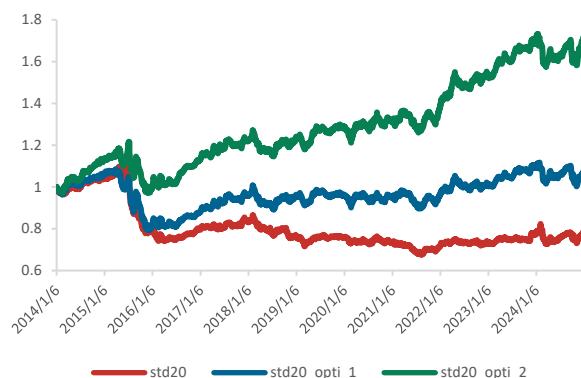
图4: std20 原始因子与第一二次改进因子累计 RankIC



数据来源: Wind、东吴证券研究所

数据日期: 2014 年 1 月 13 日至 2024 年 12 月 31 日

图5: std20 原始因子与第一二次改进因子多头超额净值



数据来源: Wind、东吴证券研究所

数据日期: 2014 年 1 月 6 日至 2024 年 12 月 31 日

在第三次改进中, DeepSeek 将因子表达式优化为:

$$\text{EMA}(\text{Greater}(\text{Greater}(\$high-\$low, \text{Greater}(\text{Abs}(\$high-\text{Ref}(\$close, 1))), \text{Abs}(\$low-\text{Ref}(\$close, 1)))), \text{Abs}(\$open-\text{Ref}(\$close, 1))) * \text{Sign}(\$close-\text{Ref}(\$close, 1) + 1e-5) * \$volume, \%d) / \text{EMA}(\$volume, \%d) / \$close$$

此表达式在第二次优化的基础上, 又增添了两点改进:

1) 四维波动极值捕捉: 在原有三重比较基础上新增开盘跳空缺口 (\$open 与前收盘价差), 形成 high-low、high-prev_close、low-prev_close、open-prev_close 四维波动极值筛选。

2) 引入方向敏感系数: 通过 $\text{Sign}(\$close-\text{Ref}(\$close, 1))$ 捕捉收盘方向, 下跌日波动赋予负权重强化空头波动信号。

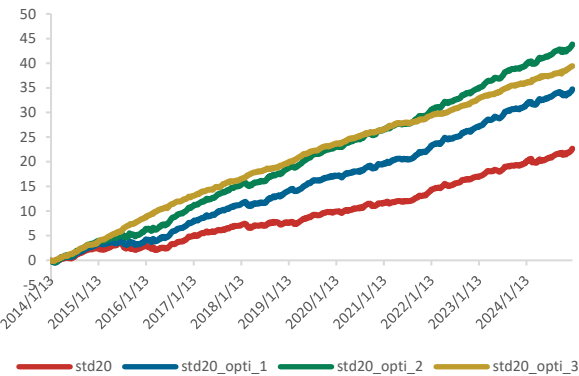
DeepSeek 对该因子的逻辑解释为: “该因子通过整合日内波动、隔夜跳空和方向性量价信息, 构建多维波动冲击指标。在捕捉最大价格变动的基础上, 通过收盘方向符号强化下跌波动的负面效应, 结合成交量加权突出资金驱动的不稳定性。相比前两版改进, 该设计更精准识别空头力量主导的异常波动, 预期可提升对风险溢价补偿要求的定价能力。”

第三次优化后, 因子的 RankIC 均值由第二次优化的 7.8% 略微降低至 7.01%, 但 ICIR 由 0.55 提升至 0.79, 年化多头超额由 5.24% 提升至 6.29%, 年化多空收益由 33.81% 大幅提升至 51.10%。从测算效果来看, 虽然第三次优化的 RankIC 均值有所下降, 但 ICIR、年化多头超额收益与年化多空超额收益均有显著提升。然而, 从人为主观理解因

子的角度，我们认为 DeepSeek 对因子的理解存在一定偏差。

原始波动因子与历次优化因子均为负向因子，意味着因子值越大，后续可能跑输市场整体；因子值越小，后续可能战胜市场整体。第三次的因子表达式中
“Greater(Greater(\$high-\$low, Greater(Abs(\$high-Ref(\$close,1)), Abs(\$low-Ref(\$close,1)))) , Abs(\$open-Ref(\$close,1)))” 部分会返回一个正数，而表达式中
“Sign(\$close-Ref(\$close,1))” 部分则根据当天的涨跌调整前者的方向（涨为 1，跌为-1）。因此，若股票长期下跌，分子端可能为负数，因子值较小；若长期盘整，分子端趋近于 0；若呈上涨趋势，分子端为正数，因子值较大。所以，我们理解该因子在原有波动率因子的基础上，叠加了反转因子，从而使因子 ICIR 及收益得到一定提升。

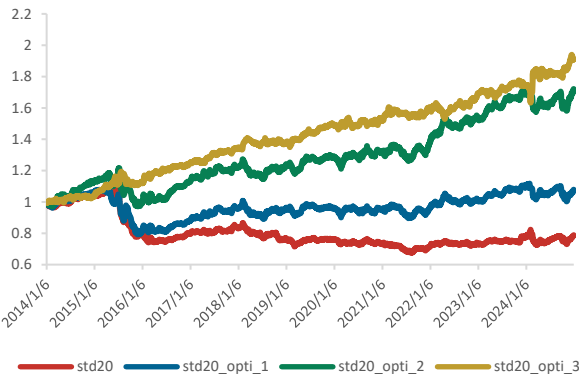
图6：std20 原始因子与第一二三次改进因子累计 RankIC



数据来源：Wind、东吴证券研究所

数据日期：2014 年 1 月 13 日至 2024 年 12 月 31 日

图7：std20 原始因子与第一二三次改进因子多头超额净值



数据来源：Wind、东吴证券研究所

数据日期：2014 年 1 月 6 日至 2024 年 12 月 31 日

除优化成功的案例外，我们也对优化无明显提升的案例进行了总结与分析。以 Beta20 因子为例，DeepSeek 在该因子上尝试了 5 次优化，但 RankIC 均值和 ICIR 均未得到提升。Beta20 原始因子构造逻辑为过去 N 个交易日收盘价的斜率除以收盘价，因子方向为负向，本质仍为反转因子。由于受到 prompt 中“因子根本逻辑不能被改变”的限制，DeepSeek 主要对分母端进行改进尝试，但效果不佳。类似的情况还出现在 min20 因子与 qtlu20 因子上，同样因分子端逻辑限制，导致改进效果不显著。

表5：Beta20 原始因子与历次优化因子表达式及 IC 统计量

因子名称	因子表达式	RankIC 均值	ICIR
beta20	$\text{Slope}(\$close, \%d) / \$close$	4.99%	0.45
beta20_opti_1	$\text{Slope}(\text{EMA}(\$close, \%d), \%d) / \text{EMA}(\$close, \%d)$	4.49%	0.39
beta20_opti_2	$\text{Slope}(\text{WMA}(\$close, \%d), \%d) / \text{WMA}(\$close, \%d)$	3.66%	0.35
beta20_opti_3	$\text{Slope}(\$close, \%d) / \text{Std}(\$close, \%d)$	3.78%	0.38

beta20_opti_4	Slope(\$vwap, %d) / Mean(Greater(\$high - \$low, Abs(\$vwap - Ref(\$vwap, 1))), %d)	3.90%	0.38
beta20_opti_5	Slope(\$close, %d) / Std(\$high - \$low, %d)	3.68%	0.36

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2024 年 12 月 31 日

表6：min20 原始因子与历次优化因子表达式及 IC 统计量

因子名称	因子表达式	RankIC 均值	ICIR
min20	Min(\$low, %d)/\$close	7.48%	0.72
min20_opti_1	Min(\$low, %d)/EMA(\$close, %d)	4.63%	0.38
min20_opti_2	Min(\$low, %d)/Ref(Mean(Greater(\$close, \$low), %d), 1)	3.62%	0.29
min20_opti_3	Min(\$low, %d)/(WMA(\$close, %d)*Std(\$close, %d))	3.63%	0.49
min20_opti_4	Min(\$low, %d)/(EMA(\$close, %d)-Std(\$close, %d))	3.40%	0.38
min20_opti_5	Min(\$low, %d)*Mean(Greater(\$close, Ref(\$low, 1)), %d)/\$close	0.14%	0.03

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2024 年 12 月 31 日

表7：qtl20 原始因子与历次优化因子表达式及 IC 统计量

因子名称	因子表达式	RankIC 均值	ICIR
qtl20	Quantile(\$close, %d, 0.8)/\$close	5.73%	0.47
qtl20_opti_1	Quantile(\$close, %d, 0.9)*Slope(\$close, %d)/\$close	3.87%	0.45
qtl20_opti_2	EMA(Quantile(Greater(\$close,EMA(\$close,%d)),%d,0.8),%d)/(Std(\$close,%d)*\$close)	3.93%	0.56
qtl20_opti_3	Resi(Quantile(Greater(\$close*\$volume,Mean(\$close*\$volume,%d)),%d,0.85),%d)/(\$close*Std(\$close,%d))	1.24%	0.28
qtl20_opti_4	Greater(Quantile(\$close,%d,0.8),EMA(\$close,%d))/(Std(\$close,%d)*\$close)	3.80%	0.54
qtl20_opti_5	Quantile(Greater(\$close,EMA(\$close,%d)),%d,0.75)*Rsquare(\$close,%d)/(\$close*Std(\$close,%d))	2.81%	0.49

数据来源：Wind、东吴证券研究所

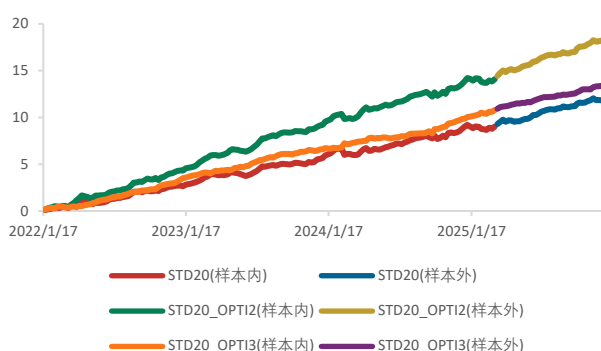
数据日期：2013 年 12 月 31 日至 2024 年 12 月 31 日

2.1.5. 改进波动率因子样本外跟踪：持续稳健且优于原始波动因子

我们以 2022 年初至 2025 年 3 月 21 日为样本内，以 2025 年 3 月 21 日至 2025 年 12 月 15 日为样本外，跟踪观察第二次和第三次改进后的波动因子在样本外的表现。

可见，改进后的波动率因子，不论是样本内还是样本外，皆优于原始的波动率因子，其中 std20_opti_2 样本外 RankIC 均值为 10.57%，std20_opti_3 样本 ICIR 为 1.04，年化多空超额收益为 33.16%。

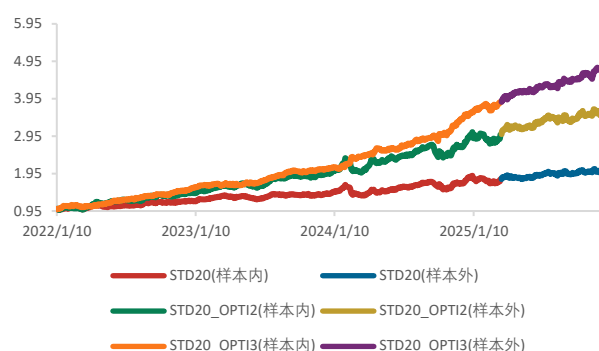
图8：std20 原始因子与第二次第三次改进因子样本内与样本外累计 RankIC



数据来源：Wind、东吴证券研究所

数据日期：2022 年 1 月 17 日至 2025 年 12 月 15 日

图9：std20 原始因子与第二次第三次改进因子样本内与样本外多空超额收益



数据来源：Wind、东吴证券研究所

数据日期：2022 年 1 月 10 日至 2025 年 12 月 15 日

表8：std20 原始因子与第二次第三次改进因子样本内与样本外选股效果统计

因子名称	RankIC 均值		ICIR		多头超额年化收益		多空超额年化收益	
	样本内	样本外	样本内	样本外	样本内	样本外	样本内	样本外
STD20	5.61%	7.16%	0.40	0.59	1.38%	-0.54%	21.10%	13.56%
STD20_OPTI2	8.75%	10.57%	0.55	0.80	7.52%	-2.33%	42.85%	27.20%
STD20_OPTI3	6.70%	6.62%	0.71	1.04	5.09%	0.84%	55.31%	33.16%

数据来源：Wind、东吴证券研究所

数据日期：2022 年 1 月 17 日至 2025 年 12 月 15 日

2.2. 从优化到创造：基于大语言模型的新因子生成范式

2.2.1. 从零生成的困境：独立探索下的效率与效果瓶颈

在第一章里，我们以 Alpha158 原始因子表达式为依托，借助 DeepSeek 开展因子改进工作，取得了一定成果，但也察觉到部分因子受限于原有逻辑，难以进一步优化。由此引发思考：若给予 DeepSeek 更自由的发挥空间，使其从因子改进转向从零进行因子生成的任务，它能否巧妙组合各类算子与可用数据，挖掘出具备出色选股能力的因子？这正是本节着力探讨的核心问题。

我们依旧从 prompt 入手，清晰告知模型所需执行的任务。因子生成任务的 prompt 由 prompt_init 与 prompt_opti 构成，以下详述 prompt_init 的主干内容：

假如你是一位资深的量化选股因子专家，你将根据现有的算子，挖掘周度 ICIR 绝对值在 0.8 以上的新因子，你返回因子表达式之后，我会调整因子方向并进行市值行业中性化，自行测算因子 IC 和 ICIR。

可用变量：

\$open：开盘价；\$close：收盘价；\$high：最高价；\$low：最低价；\$vwap：均价；\$volume：成交量。

可用算子代码：

[[code_content]]

注意以下几点：

1. 只使用提供的算子，且保证调用方法正确，可用算子见代码中 OpsList 变量。
2. 对因子进行正确的去量纲操作，转换成比例的形式，使得不同股票间可比。
3. 除了 Ref 外，每个算子的窗口期相同，且都用 %d 表示，不要出现算子窗口期为固定常数的情况，之后在测算过程中会将 %d 设置为 20。
4. 从逻辑出发进行挖掘，不需要计算因子的 RankIC 或 ICIR 等指标。

按照以下格式返回结果，其中挖掘的因子表达式写在中括号内部，不要换行，表达式内部括号只能用小括号，并对挖掘的因子表达式进行解释：

"

改进后因子表达式：[表达式]

因子解释：[因子解释]

"

prompt_opti 仍然起到收集反馈的作用，内容为：

第 1 次尝试因子表达式为 {}, RankIC 均值为 {}, ICIR 为 {}, 因子方向为 {}。

第 2 次尝试因子表达式为 {}, RankIC 均值为 {}, ICIR 为 {}, 因子方向为 {}。

第 3 次尝试因子表达式为 {}, RankIC 均值为 {}, ICIR 为 {}, 因子方向为 {}。

.....

在此基础上继续改进或者重新思考，如果表达式越来越复杂且效果越来越差，则可能需要对表达式进行简化或者完全重新思考新的方案。

因子生成部分的 prompt_init 与因子优化部分的 prompt_init 大致相似，只是去除了对 Alpha158 表达式和 IC 统计量的描述，同时为模型设定了生成周度 ICIR 在 0.8 以上因子的明确目标。prompt_opti 部分则着重将历次生成的表达式与 IC 统计量反馈给模型，促使模型在前述基础上不断改进。在实际操作过程中，我们发现模型可能会将表达式修改得过于复杂，而实际效果却不尽人意，所以在 prompt_opti 中特别提示模型，可适时简化表达式或者重新构思全新方案。

表9：历史模型生成的因子表达式与 IC 统计量

迭代次数	因子表达式	方向	RankIC 均值	ICIR
1	$\text{Log}(\text{Greater}(\$close, \text{Mean}(\$high, \%d)) / \text{Less}(\$close, \text{Mean}(\$low, \%d))) * \text{Sign}(\text{Slope}(\$volume, \%d))$	负向	3.74%	0.46
2	$\text{Slope}((\$close - \text{Mean}(\$low, \%d)) / (\text{Mean}(\$high, \%d) - \text{Mean}(\$low, \%d)), \%d) * \text{Sign}(\text{Resi}(\$volume, \%d))$	负向	1.09%	0.21
3	$\text{Sign}(\text{Slope}(\$close, \%d, \%d)) * (\text{Mean}(\$volume, \%d) / \text{Ref}(\$volume, \%d) - 1)$	负向	2.33%	0.46
4	$\text{Slope}(\$close, \%d, \%d) * \text{Slope}(\text{Log}(\$volume, \%d))$	负向	2.37%	0.41
5	$\text{Slope}(\$close, \%d) * (\text{Ref}(\$volume, \%d) / \text{Mean}(\$volume, \%d) - 1) / \text{Std}(\$volume, \%d)$	正向	0.91%	0.21
6	$\text{Slope}((\$close - \text{Mean}(\$low, \%d)) / (\text{Mean}(\$high, \%d) - \text{Mean}(\$low, \%d)), \%d) * ((\$volume - \text{Mean}(\$volume, \%d)) / \text{Std}(\$volume, \%d))$	负向	1.73%	0.26
7	$\text{Slope}((\$close / \text{Mean}(\$close, \%d)), \%d) * (\text{Rank}(\$volume, \%d) - \text{Ref}(\text{Rank}(\$volume, \%d), \%d))$	负向	1.97%	0.32
8	$\text{Log}((\$close - \text{Mean}(\$low, \%d)) / (\text{Mean}(\$high, \%d) - \text{Mean}(\$low, \%d))) * (\$volume / \text{Mean}(\$volume, \%d) - 1)$	负向	3.46%	0.52
9	$(\$close - \text{Mean}(\$low, \%d)) / (\text{Mean}(\$high, \%d) - \text{Mean}(\$low, \%d)) * \text{Sign}((\$volume / \text{Ref}(\$volume, \%d) - 1) - \text{Mean}((\$volume / \text{Ref}(\$volume, \%d) - 1), \%d))$	负向	1.55%	0.24
10	$(\$close / \text{Mean}(\$close, \%d) - 1) / \text{Std}(\$close, \%d) * (\$volume - \text{Mean}(\$volume, \%d)) / \text{Std}(\$volume, \%d)$	负向	1.92%	0.32
11	$\text{Log}((\$close - \text{Mean}(\$low, \%d)) / (\text{Mean}(\$high, \%d) - \text{Mean}(\$low, \%d))) * (\text{Slope}(\$volume, \%d) / \text{Std}(\$volume, \%d))$	负向	3.35%	0.49
12	$\text{Log}(\$close / \text{Mean}(\$high, \%d)) * (\text{Ref}(\$volume, \%d) / \text{Mean}(\$volume, \%d) - 1)$	正向	2.13%	0.35
13	$\text{Log}(\$close / \text{Mean}(\$high, \%d)) * ((\$volume - \text{Mean}(\$volume, \%d)) / \text{Std}(\$volume, \%d))$	负向	2.18%	0.27
14	$\text{Log}(\$close / \text{Mean}(\$close, \%d)) * \text{Slope}(\$volume, \%d)$	负向	2.99%	0.46
15	$\text{Log}(\$close / \text{Mean}(\$close, \%d)) * \text{Sign}(\text{Slope}(\$volume, \%d))$	负向	2.70%	0.35
16	$\text{Log}((\$close - \text{Mean}(\$low, \%d)) / (\text{Mean}(\$high, \%d) - \text{Mean}(\$low, \%d))) * (\$volume / \text{Mean}(\$volume, \%d) - 1)$	负向	3.46%	0.52
17	$((\$close - \text{Min}(\$low, \%d)) / (\text{Max}(\$high, \%d) - \text{Min}(\$low, \%d))) * ((\$volume - \text{Mean}(\$volume, \%d)) / \text{Std}(\$volume, \%d))$	负向	3.88%	0.59
18	$\text{Log}((\$close - \text{Min}(\$low, \%d)) / (\text{Max}(\$high, \%d) - \text{Min}(\$low, \%d))) * (\text{Slope}(\$volume, \%d) / \text{Std}(\$volume, \%d))$	正向	2.28%	0.33

19	$\text{Log}((\$close - \text{Min}(\$low, \%d)) / (\text{Max}(\$high, \%d) - \text{Min}(\$low, \%d))) * ((\$volume - \text{Min}(\$volume, \%d)) / (\text{Max}(\$volume, \%d) - \text{Min}(\$volume, \%d)))$	负向	0.68%	0.09
20	$\text{Log}((\$close - \text{Mean}(\$low, \%d)) / (\text{Mean}(\$high, \%d) - \text{Mean}(\$low, \%d))) * \text{Sign}(\text{Slope}(\$volume, \%d))$	负向	2.80%	0.44

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2024 年 12 月 31 日

从上表可知，我们让 DeepSeek 对因子表达式进行了 20 次迭代，然而生成的因子预测效果普遍一般。其中，RankIC 均值最大值为 3.88%，最小值仅 0.68%；ICIR 最大值为 0.59，与 prompt 中要求 0.8 的 ICIR 仍存在一定差距。

尽管在这 20 次迭代中，ICIR 未达预设目标，但 DeepSeek 给出的因子表达式并非毫无逻辑。比如第一次生成的因子表达式为：

$$\text{Log}(\text{Greater}(\$close, \text{Mean}(\$high, \%d)) / \text{Less}(\$close, \text{Mean}(\$low, \%d))) * \text{Sign}(\text{Slope}(\$volume, \%d))$$

可以看出 DeepSeek 试图通过该因子捕捉量价协同效应，对放量上涨或缩量下跌的股票赋予因子值权重，而长期盘整的股票因子值则相对较小。

第二次生成的因子表达式为：

$$\text{Slope}((\$close - \text{Mean}(\$low, \%d)) / (\text{Mean}(\$high, \%d) - \text{Mean}(\$low, \%d)), \%d) * \text{Sign}(\text{Resi}(\$volume, \%d))$$

是在第一次表达式基础上的升级。DeepSeek 对该表达式的解释为：

1) 价格标准化：(收盘价-最低价均值)/(最高价均值-最低价均值)形成 0-1 区间波动，消除价格绝对水平影响；

2) 价格趋势：对标准化价格序列计算斜率捕捉趋势强度；

3) 量能异常：用成交量残差的符号表征非趋势性放量。该设计同时捕捉价格突破强度与异常量能信号，且标准化过程使跨股票可比性增强。

不过我们也发现，DeepSeek 对该表达式的理解存在一定误区，例如模型提到“(收盘价 - 最低价均值)/(最高价均值 - 最低价均值)形成 0 - 1 区间波动”，实际上对于趋势下跌的股票，最新收盘价可能小于过去 N 个交易日最低价均值，波动区间可能低于 0。但同时也能看出，DeepSeek 有一些创新构造方式，比如用“Sign(Resi(\$volume,%d))”捕捉异常的放量与缩量，在高频场景下，这种方式对收益预测确有一定帮助。

2.2.2. 案例驱动的突破：基于先验知识的低相关性因子挖掘

在上一节中，我们让 DeepSeek 从零开始生成因子，历经 20 次迭代，挖掘出的因子 ICIR 仍未达预期。那么，若为模型提供一些成功案例，使其能总结这些案例的优点，是否能实现更出色的因子挖掘效果呢？基于此，本节我们将第一章中的 Alpha158 原始因子表达式、优化后的表达式以及对应的 IC 统计量作为“成功案例”输入模型，探究在此基础上能否取得更好成果。

我们依旧从 prompt 着手，清晰告知模型需要执行的任务。因子生成任务的 prompt 仍由 prompt_init 与 prompt_opti 构成。以下是 prompt_init 的主干内容：

假如你是一位资深的量化选股因子专家，你将根据现有的截面日频量化选股因子的相关信息，根据样例因子表达式，挖掘周度 ICIR 绝对值在 0.8 以上的新因子：

以下是通过算子计算得到的因子与对应的 RankIC 均值及 ICIR，名称中含_opti 的因子是由原始因子(不含 opti)尝试改进后得到，RankIC 均值与 ICIR 是假设表达式中 %d 为 20 个交易日，调整因子方向并进行市值行业中性化后计算得到的：

[

因子名称：std20；因子表达式：Std(\$close,%d)/\$close；RankIC 均值：0.0403；ICIR：0.3105；因子方向：负向；

因子名称：std20_opti_1；因子表达式：Mean(Greater(\$high-\$low, Greater(Abs(\$high-Ref(\$close,1)), Abs(\$low-Ref(\$close,1)))), %d)/\$close；RankIC 均值：0.0618；ICIR：0.4152；因子方向：负向；

因子名称：std20_opti_2；因子表达式：EMA(Greater(\$high-\$low, Greater(Abs(\$high-Ref(\$close,1)), Abs(\$low-Ref(\$close,1))))*\$volume,%d)/EMA(\$volume,%d)/\$close；RankIC 均值：0.0780；ICIR：0.5520；因子方向：负向；

.....

]

可用变量：

\$open：开盘价；\$close：收盘价；\$high：最高价；\$low：最低价；\$vwap：均价；\$volume：成交量。

可用算子代码：

[{code_content}]

注意以下几点：

1. 只使用提供的算子，且保证调用方法正确，可用算子见代码中 OpsList 变量。
2. 对因子进行正确的去量纲操作，转换成比例的形式，使得不同股票间可比。

- 除了 Ref 外，每个算子的窗口期相同，且都用 %d 表示，不要出现算子窗口期为固定常数的情况。
- 从逻辑出发进行挖掘，不需要计算因子的 RankIC 或 ICIR 等指标。
- 生成的新因子与样例因子不能相同，相关性尽可能低。

按照以下格式返回结果，其中挖掘的因子表达式写在中括号内部，不要换行，表达式内部括号只能用小括号，并对挖掘的因子表达式进行解释：

"

改进后因子表达式：[表达式]

因子解释：[因子解释]

"

prompt_opti 仍然起到收集反馈的作用，内容为：

第 1 次尝试因子表达式为 {}, RankIC 均值为 {}, ICIR 为 {}, 因子方向为 {}。

第 2 次尝试因子表达式为 {}, RankIC 均值为 {}, ICIR 为 {}, 因子方向为 {}。

第 3 次尝试因子表达式为 {}, RankIC 均值为 {}, ICIR 为 {}, 因子方向为 {}。

.....

在此基础上继续改进或者重新思考，如果表达式越来越复杂且效果越来越差，则可能需要对表达式进行简化或者完全重新思考新的方案。

本节的 prompt 与从零生成因子的 prompt 相比，新增了将 Alpha158 原始与优化表达式及其 IC 统计量输入给模型这一内容，prompt_opti 则与前文保持一致。

表10：从 Alpha158 生成的因子表达式与 IC 统计量

因子 序号	迭代 次数	因子表达式	方向	RankIC 均值	ICIR
1	1	EMA(Greater(\$close-Ref(\$close,1),0)*Power(\$close/Ref(\$close,1)-1,2)*\$volume,%d)/EMA(\$volume,%d)-EMA(Greater(Ref(\$close,1)-\$close,0)*Power(Ref(\$close,1)/\$close-1,2)*\$volume,%d)/EMA(\$volume,%d)	负向	6.74%	0.83
2	1	EMA(Greater(Greater(\$high-Ref(\$close,1),Abs(\$open-Ref(\$close,1))),\$low-Ref(\$close,1))*Sign(\$volume-EMA(\$volume,%d))*Power(\$close/Ref(\$close,1),2,%d)/EMA(Power(\$close/Ref(\$close,1),2)*\$volume,%d)	负向	3.89%	0.65
	2	EMA((\$close/Ref(\$close,1)-1)*Sign(\$volume-EMA(\$volume,%d))*Log(\$volume+1,%d)/Std(\$close/Ref(\$close,1)-1,%d)	负向	4.67%	0.77

3	EMA((((\$close/Ref(\$close,1)-1)*Sign(\$volume-Ref(\$volume,1)) - Mean((\$close/Ref(\$close,1)-1)*Sign(\$volume-Ref(\$volume,1)),%d)),%d)/Std(\$close/Ref(\$close,1)-1,%d)	负向	3.52%	0.64
	EMA((\$close/Ref(\$close,1)-1)*Sign(\$volume-EMA(\$volume,%d))*Abs(\$close/Ref(\$close,1)-1,%d)/Std(\$close/Ref(\$close,1)-1,%d)	负向	6.56%	0.82
3	EMA(Greater(\$high-Ref(\$close,1),\$close-Ref(\$low,1))*(\$close/Ref(\$close,1)-1)*Sign(\$volume-EMA(\$volume,%d)),%d)/(Std(\$close,%d)*EMA(\$volume,%d))	负向	2.56%	0.40
	EMA((Sign(Slope(\$close,%d)) * (\$close/Ref(\$close,1)-1) * (\$volume/EMA(\$volume,%d))),%d)/Std(\$close/Ref(\$close,1)-1,%d)	负向	2.36%	0.29
	EMA((\$close/Ref(\$close,1)-1)*Power(\$volume/EMA(\$volume,%d),2),%d)/Std(\$close/Ref(\$close,1)-1,%d)	负向	6.59%	0.72
	EMA((\$close/Ref(\$close,1)-1)*Power(\$volume/EMA(\$volume,%d),2)*Sign(Slope(\$volume,%d)),%d)/Std(Resi(\$close,%d),%d)	负向	2.48%	0.47
	EMA((\$close/Ref(\$close,1)-1)*Power(\$volume/EMA(\$volume,%d),2),%d)/Std(Resi(\$close,%d),%d)	负向	5.17%	0.65
	EMA((\$close/Ref(\$close,1)-1)*Power(\$volume/EMA(\$volume,%d),2),%d)/(Std(Resi(\$close,%d),%d)*Std(\$volume/EMA(\$volume,%d),%d))	负向	4.60%	0.59
	EMA((\$close/Ref(\$close,1)-1)*Power(\$volume/EMA(\$volume,%d),1.5),%d)/Std(Resi(\$close,%d),%d)	负向	5.38%	0.64
	EMA(Greater(\$high-Ref(\$close,1),Abs(\$close-Ref(\$low,1)))*(\$close/Ref(\$close,1)-1)*(\$volume/EMA(\$volume,%d)),%d)/Std(Resi(\$close,%d),%d)	负向	8.76%	0.87
4	EMA(Greater(\$high-Ref(\$close,1),Abs(\$close-Ref(\$low,1)))*Sign(\$close-EMA(\$close,%d))*Power(\$volume/EMA(\$volume,%d),0.5),%d)/Std(Resi(\$close,%d),%d)	负向	5.49%	0.54
	EMA(Greater(\$high-Ref(\$close,1),Abs(\$close-Ref(\$low,1)))*Sign(\$close-EMA(\$close,%d))*(\$volume/EMA(\$volume,%d)),%d)/Std(Resi(\$close,%d),%d)	负向	5.91%	0.59
	EMA((Greater(\$high-Ref(\$close,1),Abs(\$close-Ref(\$low,1)))*Sign(\$close-Ref(\$close,1))*(\$volume/EMA(\$volume,%d)),%d)/Std(Greater(\$high-\$low,Abs(\$high-Ref(\$close,1))),%d)	负向	5.24%	0.59
	EMA((Slope(\$close,%d)*Sign(\$volume-EMA(\$volume,%d))),%d)/Std(Resi(\$volume,%d),%d)	正向	0.09%	0.02
	EMA((Slope(\$close,%d)*Sign(\$volume/Ref(\$volume,1)-1)),%d)/Std(\$close/Ref(\$close,1)-1,%d)	正向	0.60%	0.13
	EMA((\$close/Ref(\$close,1)-1)*(\$volume/Ref(\$volume,1)-1),%d)/Std(\$close/Ref(\$close,1)-1,%d)	负向	6.20%	0.87
5	EMA(Greater(Greater(\$high-\$low,Abs(\$high-Ref(\$close,1))),Abs(\$low-Ref(\$close,1)))*Sign(\$close-Ref(\$close,1)+1e-5)*\$volume,%d)/EMA(\$volume,%d)/Std(\$close,%d)	负向	5.12%	0.65
	EMA(((\$high-\$low)*(\$close/Ref(\$close,1)-1)*(\$volume/EMA(\$volume,%d))),%d)/Std(\$close/Ref(\$close,1)-1,%d)	负向	7.27%	0.87
	EMA(Resi(\$close,%d)*Sign(\$volume-EMA(\$volume,%d))*(Max(\$high,%d)/\$close-1),%d)/Std(\$close,%d)	负向	2.90%	0.43
	EMA((\$close/Ref(\$close,1)-1)*Sign(\$volume/Ref(\$volume,1)-1),%d)/Std(\$close/Ref(\$close,1)-1,%d)	负向	5.87%	0.90

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2024 年 12 月 31 日

经过有限测试，最终生成了数个 ICIR 在 0.8 以上的因子。其中，第一次尝试经一次迭代便得到满足条件的因子表达式，有时则需历经 8 次迭代才得到较好的因子。

尽管我们在 prompt 中要求新因子与参考样例因子相关性尽可能低，但 DeepSeek 作为大语言模型自身不具备计算能力。因此，我们测算了新因子与样例因子中 ICIR 大于 0.7 的因子之间的相关性，以及新因子之间的相关性。

表11：新因子与 ICIR 大于 0.7 的样例因子之间的相关性

	Factor_1	Factor_2	Factor_3	Factor_4	Factor_5
std20_opti_3	31.85%	16.81%	42.47%	21.85%	20.32%
cntd20_opti_2	29.35%	22.24%	44.47%	24.76%	22.70%
min20	-24.01%	-16.18%	-34.03%	-15.51%	-15.98%
vma20_opti_5	6.01%	-2.90%	3.51%	-0.31%	-1.07%
vstd20_opti_1	10.58%	9.81%	22.56%	13.39%	10.74%
vstd20_opti_2	7.28%	6.08%	15.79%	9.72%	7.36%
vstd20_opti_3	6.09%	9.02%	19.50%	14.73%	9.47%
vsumn20_opti_2	19.01%	13.67%	26.94%	16.71%	21.44%
vsump20_opti_1	44.38%	3.83%	8.01%	4.52%	4.12%
vsump20_opti_2	45.28%	3.97%	8.29%	4.65%	4.23%
vsump20_opti_3	43.98%	7.72%	18.73%	12.38%	10.62%
cord20	14.60%	8.25%	14.39%	7.54%	15.56%
corr20_opti_2	7.29%	24.52%	24.25%	17.13%	71.36%
corr20_opti_3	14.22%	8.19%	14.28%	7.34%	15.94%
corr20_opti_4	12.19%	8.92%	9.00%	4.18%	13.68%

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2024 年 12 月 31 日

表12：新因子之间相关性

	Factor_1	Factor_2	Factor_3	Factor_4	Factor_5
Factor_1		8.92%	20.91%	11.64%	9.42%
Factor_2	8.92%		11.17%	33.16%	48.19%
Factor_3	20.91%	11.17%		41.20%	29.32%
Factor_4	11.64%	33.16%	41.20%		46.75%
Factor_5	9.42%	48.19%	29.32%	46.75%	

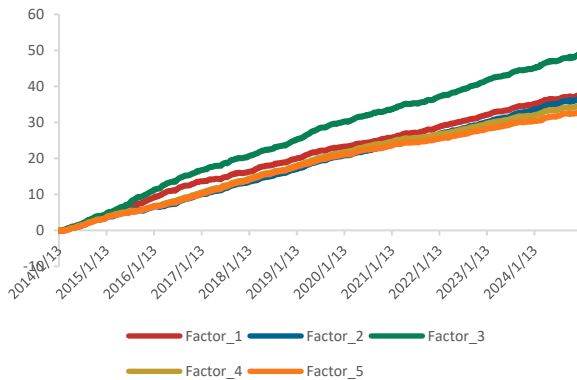
数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2024 年 12 月 31 日

从表中可见，新生成的因子与样例因子中效果较好的因子相关性较低，仅 Factor_5 与 corr20_opti_2 之间的相关性达到 71.36%，其他因子之间的相关性均在 50%

以下，其中 Factor_2 与所有出色样例因子相关性均在 25% 以下。新因子之间的相关性也在可接受范围内，所有因子之间相关性最大值为 48.19%，均值为 26.07%，其中 Factor_1 与其它 4 个因子的相关性均在 21% 以下。

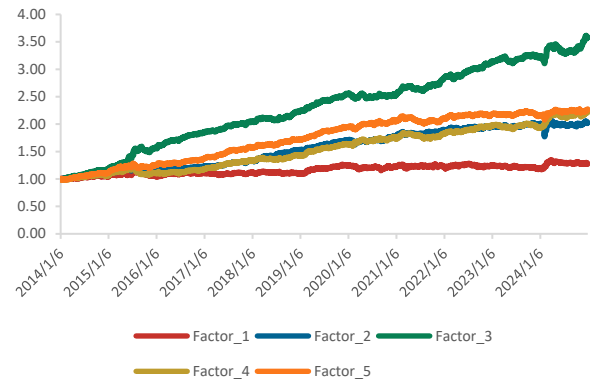
图10：新生成因子累计 RankIC



数据来源：Wind、东吴证券研究所

数据日期：2014年1月13日至2024年12月31日

图11：新生成因子多头超额净值



数据来源：Wind、东吴证券研究所

数据日期：2014年1月6日至2024年12月31日

以新生成的前两个因子为例，分析 DeepSeek 所生成因子的底层逻辑。

第一个有效因子表达式为：

$$\text{EMA}(\text{Greater}(\$close - \text{Ref}(\$close, 1), 0) * \text{Power}(\$close / \text{Ref}(\$close, 1) - 1, 2) * \$volume, \%d) / \text{EMA}(\$volume, \%d) - \text{EMA}(\text{Greater}(\text{Ref}(\$close, 1) - \$close, 0) * \text{Power}(\text{Ref}(\$close, 1) / \$close - 1, 2) * \$volume, \%d) / \text{EMA}(\$volume, \%d)$$

AI 对该因子的逻辑解释为：“该因子捕捉量价协同增强效应。分子端计算上涨日价格涨幅平方与成交量的加权均值（EMA），突出强势股的量能驱动特征；分母端进行成交量标准化消除规模影响。通过对称计算下跌日的量价惩罚项并做差值，有效区分多头与空头动能。价格涨幅平方处理放大突破行情的区分度，EMA 平滑避免短期噪声，最终形成兼具趋势强度和量能验证的双向驱动因子。”

可见 DeepSeek 对因子表达式的理解基本正确。通过学习样例表达式，AI 在第一次迭代中便能够使用“ $\$volume / \text{EMA}(\$volume, \%d)$ ”表达成交量加权。在创新性上，AI 使用“上涨日价格变动×涨幅平方 - 下跌日价格变动×跌幅平方”区分多头与空头动能，使得因子 ICIR 达到 0.83。

第二个有效因子表达式为：

$$\text{EMA}((\$close / \text{Ref}(\$close, 1) - 1) * \text{Sign}(\$volume - \text{EMA}(\$volume, \%d)) * \text{Abs}(\$close / \text{Ref}(\$close, 1) - 1, \%d) / \text{Std}(\$close / \text{Ref}(\$close, 1) - 1, \%d))$$

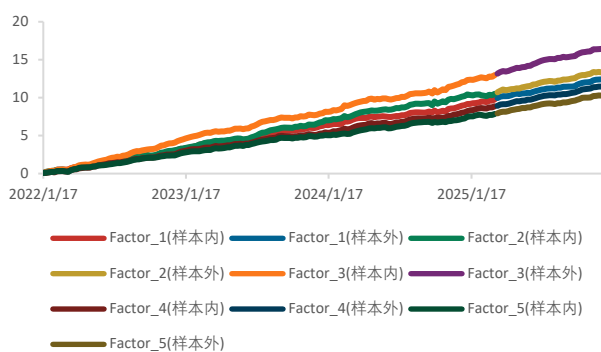
AI 对该因子的逻辑解释为：“该因子聚焦量价共振的强度维度：1) 价格收益率与成交量突破 EMA 的方向符号相乘，捕捉量价同向动量；2) 乘以收益率绝对值强化大波动日的信号权重；3) EMA 平滑保留趋势持续性特征；4) 用收益率波动率标准化消除市场整体波动影响。通过“方向×幅度×量能验证”三维度构建纯净动量信号。”

人为拆分该因子，表达式分子由 3 个部分相乘组成 “ $(\text{\$close}/\text{Ref}(\text{\$close},1)-1)$ ”、“ $\text{Sign}(\text{\$volume}-\text{EMA}(\text{\$volume},\%d))$ ” 和 “ $\text{Abs}(\text{\$close}/\text{Ref}(\text{\$close},1)-1)$ ”。不难发现，对于连续放量大涨，或者连续缩量大跌，又或者处于盘整过程中涨时放量跌时缩量的股票，在该表达式中的分子端值会偏大，反之则偏小。而在分母端除以近期的波动率，则进一步放大了近期连续放量大涨或缩量大跌的因子值。该因子周频 ICIR 达到 0.82，且与第一个因子的相关性仅为 8.92%。

2.2.3. AI 生成因子样本外跟踪：IC 出色超额上行

对新生成的 5 个因子，我们同样以 2022 年初至 2025 年 3 月 21 日为样本内，以 2025 年 3 月 21 日至 2025 年 12 月 15 日为样本外，对其进行样本外跟踪。DeepSeek 挖掘的 5 个新因子，近期样本外 RankIC 相对样本内 RankIC 基本没有衰减，均在 6% 以上，ICIR 皆在 1.0 左右。其中 Factor_4 多头超额年化收益达到 7.3%，Factor_3 的多空超额年化收益达到 33.79%。证明了 DeepSeek 生成的日频价量因子在样本内与实际样本外的效果稳定且持久。

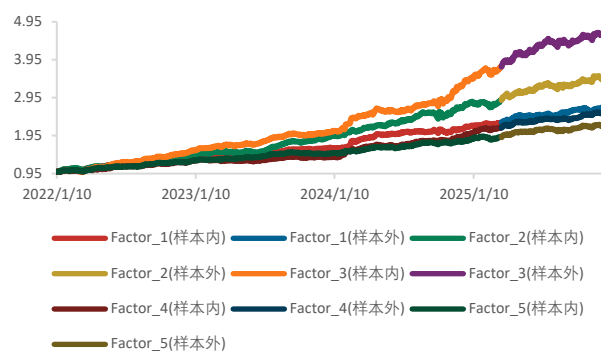
图12：生成因子样本内与样本外累计 RankIC



数据来源：Wind、东吴证券研究所

数据日期：2022 年 1 月 17 日至 2025 年 12 月 15 日

图13：生成因子样本内与样本外多空超额收益



数据来源：Wind、东吴证券研究所

数据日期：2022 年 1 月 10 日至 2025 年 12 月 15 日

表13：生成因子样本内与样本外选股效果统计

因子名称	RankIC 均值		ICIR		多头超额年化收益		多空超额年化收益	
	样本内	样本外	样本内	样本外	样本内	样本外	样本内	样本外

Factor_1	6.07%	6.76%	0.79	1.03	-0.30%	0.12%	30.82%	25.48%
Factor_2	6.53%	7.27%	0.69	0.92	2.07%	4.24%	41.23%	27.55%
Factor_3	8.06%	8.84%	0.77	1.13	6.62%	5.00%	54.01%	33.79%
Factor_4	5.51%	6.68%	0.79	1.26	5.22%	7.30%	28.45%	26.51%
Factor_5	4.88%	6.24%	0.80	1.14	1.71%	2.16%	23.56%	22.18%

数据来源：Wind、东吴证券研究所

数据日期：2022 年 1 月 17 日至 2025 年 12 月 15 日

3. AI 基本面因子挖掘：构建自动化研究框架

3.1. 实验设计：基础数据与处理算子

为实现人工智能（AI）模型在选股因子挖掘领域的自动化应用，首要步骤是构建一个结构化且逻辑严谨的实验环境。该环境由两个核心部分构成：标准化的基础数据集与一系列预定义的因子处理算子。本节将详细阐述这些基础设定，它们是整个 AI 基本面因子生成流程的基石。

本研究的基础数据层涵盖了财务数据与行情数据两大类。在财务数据方面，我们从上市公司的利润表、资产负债表与现金流量表中，筛选出字段覆盖率高于 75% 的核心财务项目，以确保因子的有效性和覆盖广度。此外，为将公司的基本面状况与市场估值相联系，我们引入了日度频率的总市值（mv）与滚动十二个月（TTM）的分红总额（dividend_ttm）数据。这些经过筛选和整理的数据共同构成了可供 AI 调用的原始数据池。

表14：财务数据字段名称与字段释义

表类型	字段名称	字段释义	表类型	字段名称	字段释义
利润表	totaloperatingrevenue	营业总收入	资产负债表	accountreceivable	其中:应收账款
利润表	operatingrevenue	营业收入	资产负债表	advancepayment	预付款项
利润表	totaloperatingcost	营业总成本	资产负债表	inventories	存货
利润表	operatingcost	营业成本	资产负债表	totalcurrentassets	流动资产合计
利润表	operatingtaxsurcharges	营业税金及附加	资产负债表	fixedassets	其中:固定资产
利润表	operatingexpense	销售费用	资产负债表	intangibleassets	无形资产
利润表	totaladminexpense	管理费用	资产负债表	deferredtaxassets	递延所得税资产
利润表	financialexpense	财务费用	资产负债表	totalnoncurrentassets	非流动资产合计
利润表	operatingprofit	营业利润	资产负债表	totalassets	资产总计
利润表	nonoperatingincome	加:营业外收入	资产负债表	accountspayable	其中:应付账款
利润表	nonoperatingexpense	减:营业外支出	资产负债表	salariespayable	应付职工薪酬
利润表	totalprofit	利润总额	资产负债表	taxspayable	应交税费
利润表	netprofit	净利润	资产负债表	otherpayable	其中:其他应付款
利润表	npparentcompanyowners	归属于母公司所有者的净利润	资产负债表	totalcurrentliability	流动负债合计
利润表	totalcompositeincome	综合收益总额	资产负债表	totalnoncurrentliability	非流动负债合计
利润表	ciparentcompanyowners	归属于母公司所有者的综合收益总额	资产负债表	totalliability	负债合计
利润表	basiceps	基本每股收益	资产负债表	paidincapital	实收资本(或股本)
利润表	dilutedeps	稀释每股收益	资产负债表	capitalreservefund	资本公积
利润表	nonrecurringprofitloss	非经常性损益	资产负债表	surplusreservefund	盈余公积
利润表	npeductnonrecurringpl	扣除非经常性损益后的归母净利润	资产负债表	retainedprofit	未分配利润
利润表	grossprofit	毛利	资产负债表	sewwithoutmi	归属母公司所有者权益(或股东权益)合计

利润表	netincomefromoperating	经营活动净收益	资产负债表	totalshareholderequity	所有者权益(或股东权益)合计
利润表	ebit	息税前利润	资产负债表	totalliabilityandequity	负债和所有者权益(或股东权益)总计
利润表	ebitda	息税折旧摊销前利润	资产负债表	billacreceivable	应收票据及应收账款
现金流量表	goodssalesrendercash	销售商品、提供劳务收到的现金	资产负债表	notaccountspayable	应付票据及应付账款
现金流量表	othercashinrelatedoperate	收到其他与经营活动有关的现金	资产负债表	otherreceivablelead	其他应收款(含利息和股利)
现金流量表	subtotaloperatecashinflow	经营活动现金流入小计	资产负债表	totalfixedasset	固定资产合计
现金流量表	goodsservicescashpaid	购买商品、接受劳务支付的现金	资产负债表	tconstruinprocess	在建工程合计
现金流量表	staffbehalfpaid	支付给职工以及为职工支付的现金	资产负债表	otherpayablelead	其他应付款(含利息和股利)
现金流量表	alltaxespaid	支付的各项税费	资产负债表	interestfreeclabilities	无息流动负债
现金流量表	otheroperatecashpaid	支付其他与经营活动有关的现金	资产负债表	interestfreenoncl	无息非流动负债
现金流量表	subtotaloperatecashoutflow	经营活动现金流出小计	资产负债表	interestbeardebt	带息债务
现金流量表	netoperatecashflow	经营活动产生的现金流量净额	资产负债表	netdebt	净债务
现金流量表	subtotalinvestcashinflow	投资活动现金流入小计	资产负债表	totalpaidincapital	全部投入资本
现金流量表	fixintanotherassetacquitcash	购建固定资产、无形资产和其他长期资产支付的现金	资产负债表	workingcaital	营运资本
现金流量表	subtotalinvestcashoutflow	投资活动现金流出小计	资产负债表	networkingcaital	净营运资本
现金流量表	netinvestcashflow	投资活动产生的现金流量净额	资产负债表	nettangibleassets	有形资产净值
现金流量表	subtotalfinancecashinflow	筹资活动现金流入小计	资产负债表	retainedearnings	留存收益
现金流量表	dividendinterestpayment	分配股利、利润或偿付利息支付的现金	资产负债表	cashequivalents	货币资金/现金及存放中央银行款项
现金流量表	subtotalfinancecashoutflow	筹资活动现金流出小计			
现金流量表	netfinancecashflow	筹资活动产生的现金流量净额			
现金流量表	cashequivalentincrease	现金及现金等价物净增加额			
现金流量表	beginperiodcash	加:期初现金及现金等价物余额			
现金流量表	endperiodcashequivalent	期末现金及现金等价物余额			
现金流量表	fcff	企业自由现金流量 FCFF			
现金流量表	fcfe	股权自由现金流量 FCFE			
现金流量表	currentaccruedda	当期计提折旧与摊销			
现金流量表	goodssalesrendercash	销售商品、提供劳务收到的现金			

数据来源: Wind、东吴证券研究所

在原始数据池的基础上,我们为 AI 提供了一套标准化的处理算子库。这些算子封装了量化研究中常用的数据处理逻辑,例如,ttm 算子可将累计财务数据转换为滚动四个季度的口径,而 yoy 算子则用于计算同比增长率。通过提供这些预定义函数,AI 能够直接调用高级运算,从而专注于探索数据间的深层财务逻辑,而非基础的数据处理与计算。

表15：基本面因子算子名称与释义

算子名称	算子释义
get	取财务或市值数据
quarter	将财务数据累计值转为单季度值
ttm	将财务数据累计值转为 TTM 值
diff	当期财务数据减去上个季度财务数据
yoy	计算财务数据同比
qoq	计算财务数据环比
refq	财务数据前移 N 个季度
op	对多个财务数据进行运算
op2	用于财务数据与市值数据之间的运算

数据来源：东吴证券研究所整理

根据前文的研究结论，为 AI 提供有效的因子范例（即先验知识），能够显著提升其生成新因子的效率与质量。因此，本研究延续此方法，将一批经典的基本面因子及其历史回测表现作为“样例”输入给 AI 模型作为参考。

为确保所有因子（无论是作为样例的因子还是 AI 新生成的因子）均在统一、可比的基准下进行评估，我们设定了如下标准化的回测框架：

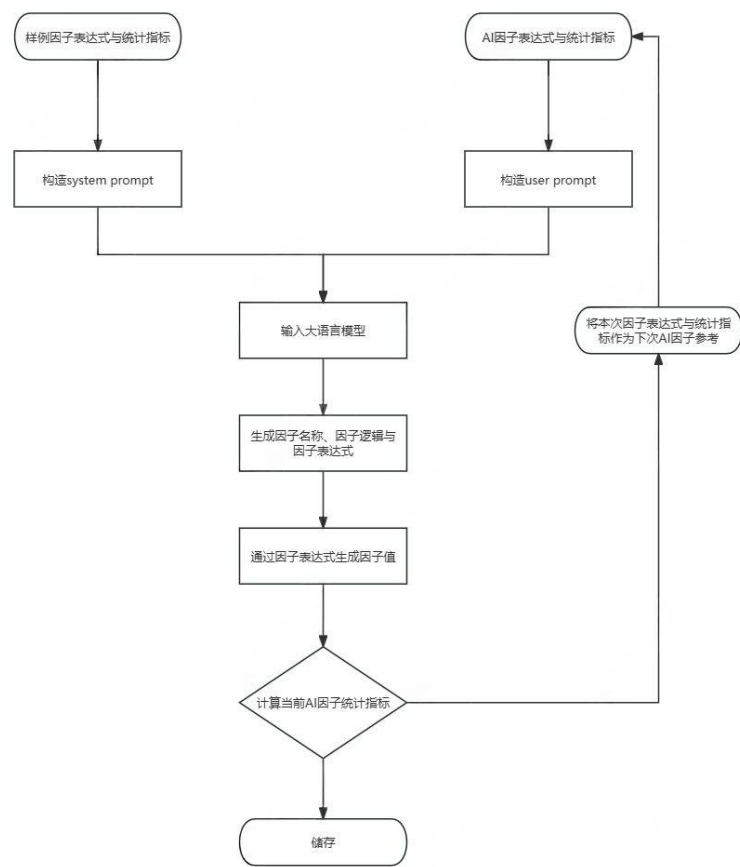
- **回测区间：**2013 年 12 月 31 日至 2025 年 6 月 30 日。
- **样本空间：**剔除上市不满 365 个自然日的新股，剔除 ST 股。
- **中性化处理：**对因子进行市值行业中性化。
- **测试方式：**周频调仓，以下周第一个交易日的 VWAP 价格成交，计算 VWAP 收益率的 RankIC 均值与 ICIR。
- **方向调整：**根据 RankIC 均值的正负，对因子方向进行调整，使得 RankIC 为正，便于比较。

通过上述数据与算子体系，AI 能够在标准化、结构化的环境下生成新的基本面因子。这为后续的因子生成流程奠定了坚实基础。在下一小节，我们将详细介绍 AI 生成基本面因子的操作流程，并结合具体案例分析其有效性。

3.2. 人机交互框架：AI 因子生成流程详解

为将前述的基础数据与算子转化为有效的选股因子，我们设计了一套人机协同的自动化 workflow。此流程旨在系统性地引导 AI 模型进行因子挖掘与创新，其核心是基于精心设计的提示工程（Prompt Engineering）构建的迭代循环机制。该流程的宏观逻辑如下图所示。

图14：基本面因子生成流程示意图



数据来源：东吴证券研究所绘制

如图 1 所示，整个因子生成流程是一个结构化的闭环系统，主要包含以下几个关键步骤：

- 1) 初始化：研究员首先定义实验的元规则，包括设定 AI 的角色、提供可用的财务数据与处理算子列表，并输入一组经过验证的“样例因子”及其历史表现作为 AI 的初始参考知识库。
- 2) AI 生成：在接收到包含上述所有信息的指令后，AI 模型开始进行创造性工作，生成一个全新的因子，并以标准化的 JSON 格式输出其名称、逻辑释义及表达式。

3) 验证与评估：研究员获取 AI 生成的因子表达式，并利用 1.1 节中定义的标准化回测框架，对新因子的历史选股效果（如 RankIC、ICIR、超额收益等）进行量化评估。

4) 反馈与迭代：经过评估后，添加至“已生成成功因子列表”中。该列表会作为动态更新的知识库，在下一轮交互中反馈给 AI，并明确要求其生成与所有已知样例及已成功因子均不重复的新因子。

通过这一“初始化 → 生成 → 评估 → 反馈”的迭代循环，AI 的学习过程被有效引导。它不仅能从最初的样例中学习，还能从自身每一轮的成功创造中汲取经验，从而持续地向更深、更广的因子空间进行探索。

该流程的具体实现，依赖于系统提示（System Prompt）与用户提示（User Prompt）的协同作用。

系统提示为 AI 设定了恒定的角色、知识边界与行为准则，是整个流程的稳定基石。其设计的核心思想在于：

- **角色设定**：“假如你是一位资深的量化选股因子专家”，这一指令旨在激活模型在特定专业领域的知识储备与推理能力。
- **知识边界定义**：通过明确提供可用的数据字段（`available_data`）和算子库（`process_code_prompt`），我们为 AI 圈定了其可以操作的“工具箱”，确保了生成表达式的可执行性，并防止了模型凭空捏造不存在的数据或函数。
- **提供范例**：样例因子（`sample_factor`）及其评价指标的展示，为 AI 提供了具体的“成功案例”，使其直观地理解何为有效的因子，并学习其结构与逻辑表达方式。
- **硬性规则约束**：诸如“因子名称与表达式不能相同”、“进行正确的去量纲操作”等规则，旨在保证生成因子的新颖性、可比性和逻辑严谨性。特别是要求 AI“从逻辑出发进行挖掘，不需要计算因子的 RankIC”，这实现了人机分工，让 AI 专注逻辑创新，将计算验证交由外部系统，提高了整体效率。

假如你是一位资深的量化选股因子专家，你将根据样例基本面因子的相关信息，挖掘在所有 A 股样本空间中，周度选股效果出色的基本面因子，以因子表达式的方式返回。

以下是算子相关代码：

```
[[process_code_prompt]]
```

以下是构造基本面因子可用的数据：

```
[[available_data]]
```

以下是现有基本面因子表达式，以及周度调仓下的因子评价指标，指标为调整因子方向并进行市值行业中性化后计算得到的：

```
[{sample_factor}]
```

注意以下几点：

1. 用 get 函数获取可用数据，仅从提供的范围内选择。
2. 对因子进行正确的去量纲操作，转换成比例的形式，使得不同股票间可比。
3. 返回的因子和样例因子及已生成因子的名称与表达式不能相同。
4. 只使用提供的算子，且保证调用方法正确。
5. 利润表与现金流量表数据为时期数据可以从累计值转为季度值或 TTM 值，但是资产负债表数据为时点值不能转换。
6. 从逻辑出发进行挖掘，不需要计算因子的 RankIC 或 ICIR 等指标。

以 JSON 格式返回结果，确保输出是纯 JSON，没有额外文本，内容能被 json.loads 正确解析。

样例 JSON 输出：

```
{
  '因子名称': 'EPS_QR_PLUS',
  '因子逻辑': '归属于母公司所有者的净利润_单季度 / 最新总股本，EPS 越高，说明公司单位股本创造的利润越多，未来股价上涨的可能性越大。',
  '因子表达式': 'op([quarter(get('npparentcompanyowners')), get('paidincapital']), 'val0 / val1')'
}
```

与静态的系统提示不同，用户提示是驱动流程迭代的关键。它在每一轮交互中动态变化，其核心作用是提供即时反馈，构建 AI 的“记忆”。

- **传递新知：**将上一轮 AI 生成并被验证有效的因子及其完整的绩效指标反馈给 AI。这相当于一个动态更新的“经验库”，让 AI 知道“已经做过什么”以及“做得怎么样”。
- **驱动创新：**在反馈了最新成果后，给出“继续寻找...”的指令。结合系统提示中“不能相同”的规则，AI 被激励去探索与所有已知因子(包括原始样例和自身已生成的)逻辑和表达均不相同的全新方向，避免了在已有成果附近的重复探索。

以下是 AI 已生成的因子：

```
{'因子名称': XXX, '因子逻辑': XXX, '因子表达式': XXX, 'RankIC 均值': XXX, 'ICIR': XXX, '因子方向': XXX, '年  
化多头超额': XXX, '多头超额收益波动比': XXX, '年化多空超额': XXX, '多空超额收益波动比': XXX}
```

```
{'因子名称': XXX, '因子逻辑': XXX, '因子表达式': XXX, 'RankIC 均值': XXX, 'ICIR': XXX, '因子方向': XXX, '年  
化多头超额': XXX, '多头超额收益波动比': XXX, '年化多空超额': XXX, '多空超额收益波动比': XXX}
```

```
.....
```

继续寻找选股能力出色的基本面因子，按照要求的格式返回。

综上，这一人机协同的闭环框架将 AI 的计算创造力与人类研究员的专业验证能力有机结合，构建了一个可扩展、高效率的因子发现引擎，为后续大规模、系统性的因子生成实验奠定了坚实的方法论基础。

3.3. AI 价值因子挖掘：经典框架的有效拓展与创新

为启动 AI 的因子挖掘流程，我们首先需要为其提供一套高质量的“先验知识”。在价值投资领域，我们选取了市场上广泛认可的经典价值因子，如市盈率倒数（EP）、市净率倒数（BP）、市销率倒数（SP）、市现率倒数（CFP）以及股息率（DP）等，作为 AI 学习和模仿的基准。这些“样例因子”及其历史回测表现构成了 AI 进行价值因子探索的初始知识库，旨在引导其理解价值投资的核心逻辑——寻找被市场低估的优质资产。

表16：样例价值因子释义与表达式

因子名称	因子释义	表达式
EP_TTM	归属于母公司所有者的净利润_TTM / 总市值	op2([ttm(get('npapparentcompanyowners')), get('mv')], 'val0 / val1')
EP_TTM_DEDUCTED	扣除非经常性损益后的归母净利润_TTM / 总市值	op2([ttm(get('npdeductnonrecurringpl')), get('mv')], 'val0 / val1')
BP_LF	归属母公司所有者权益(或股东权益)合计 / 总市值	op2([get('sewithoutmi'), get('mv')], 'val0 / val1')
SP_TTM	营业收入_TTM / 总市值	op2([ttm(get('operatingrevenue')), get('mv')], 'val0 / val1')
OCFP_TTM	经营活动现金净流量_TTM / 总市值	op2([ttm(get('netoperatecashflow')), get('mv')], 'val0 / val1')
DIVIDEND_YIELD_TTM	股息率（近 12 个月）	op2([get('dividend_ttm'), get('mv')], 'val0 / val1')

数据来源：Wind、东吴证券研究所

表17：样例价值因子统计指标

因子名称	RankIC 均值	ICIR	年化多头超额	多头超额收益波动比	年化多空超额	多空超额收益波动比
EP_TTM	3.59%	0.39	7.39%	0.90	12.53%	1.22
EP_TTM_DEDUCTED	3.34%	0.37	7.82%	0.97	10.77%	1.07
BP_LF	4.13%	0.36	7.17%	0.84	20.91%	2.12
SP_TTM	2.78%	0.29	4.88%	0.78	15.87%	2.27

OCFP_TTM	2.80%	0.41	5.17%	0.66	9.54%	1.72
DIVIDEND_YIELD_TTM	3.21%	0.40	7.16%	1.08	17.41%	2.80

数据来源：Wind、东吴证券研究所

接收到初始指令和样例后，AI 开始自主探索。在我们的实验中，模型成功生成了一个包含 20 个全新因子的多样化组合。这些由 AI 创造的因子不仅局限于对样例的简单模仿，而是展现出对财务数据更深层次的理解和创造性的组合。它们广泛地从股权自由现金流（FCFEP）、企业自由现金流（FCFFP）、息税折旧摊销前利润（EBITDA）、留存收益（Retained Earnings）等多个维度，对企业的内在价值进行了更为精细和立体的刻画。

表18：AI 价值因子释义与表达式

因子名称	因子释义	因子表达式
FCFEP_TTM	股权自由现金流 TTM / 总市值	op2([ttm(get('fcfe')), get('mv')], 'val0 / val1')
EBITDA_P_TTM	EBITDA_TTM / 总市值	op2([ttm(get('ebitda')), get('mv')], 'val0 / val1')
FCFFP_TTM	企业自由现金流 TTM / 总市值	op2([ttm(get('fcff')), get('mv')], 'val0 / val1')
GP_TTM	毛利 TTM / 总市值	op2([ttm(get('grossprofit')), get('mv')], 'val0 / val1')
TBPR_LF	有形资产净值 / 总市值	op2([get('nettangibleassets'), get('mv')], 'val0 / val1')
OPP_TTM	营业利润 TTM / 总市值	op2([ttm(get('operatingprofit')), get('mv')], 'val0 / val1')
CSHREVP_TTM	销售商品、提供劳务收到的现金 _TTM / 总市值	op2([ttm(get('goodssalesrendercash')), get('mv')], 'val0 / val1')
REP_LF	留存收益 / 总市值	op2([get('retainedearnings'), get('mv')], 'val0 / val1')
OEP_TTM	所有者盈余 TTM / 总市值	op2([ttm(get('npparentcompanyowners')), ttm(get('currentaccruedda')), ttm(get('fixintanotherassetacquitash')), get('mv')], '(val0 + val1 - val2) / val3')
EBITP_TTM	EBIT_TTM / 总市值	op2([ttm(get('ebit')), get('mv')], 'val0 / val1')
TCIP_TTM	综合收益总额 TTM / 总市值	op2([ttm(get('ciparentcompanyowners')), get('mv')], 'val0 / val1')
NWCP_LF	净营运资本 / 总市值	op2([get('networkingcaital'), get('mv')], 'val0 / val1')
CEP_TTM	现金收益 TTM / 总市值	op2([ttm(get('npparentcompanyowners')), ttm(get('currentaccruedda')), get('mv')], '(val0 + val1) / val2')
CGP_TTM	现金毛利 TTM / 总市值	op2([ttm(get('goodssalesrendercash')), ttm(get('goodsservicescashpaid')), get('mv')], '(val0 - val1) / val2')
TPP_TTM	利润总额 TTM / 总市值	op2([ttm(get('totalprofit')), get('mv')], 'val0 / val1')
TPICP_LF	实收资本 / 总市值	op2([get('totalpaidincapital'), get('mv')], 'val0 / val1')
NIEOAP_TTM	经营活动净收益 TTM / 总市值	op2([ttm(get('netincomefromoperating')), get('mv')], 'val0 / val1')
NOAP_LF	净营运资产 / 总市值	op2([get('totalassets'), get('cashequivalents'), get('totalliability'), get('interestbeardebt'), get('mv')], '((val0 - val1) - (val2 - val3)) / val4')
TOTAL_YIELD_PLUS_TTM	(分红 + 净偿还债务) / 总市值	op2([get('dividend_ttm'), op([refq(get('interestbeardebt'), 4), get('interestbeardebt')], 'val0 - val1'), get('mv')], '(val0 + val1) / val2')
REINVESTMENT_YIELD_TTM	(净利润 - 分红) / 总市值	op2([ttm(get('npparentcompanyowners')), get('dividend_ttm'), get('mv')], '(val0 - val1) / val2')

数据来源：Wind、东吴证券研究所

从整体回测表现来看，AI 生成的价值因子组合展现出不错的选股能力。综合统计显示，这 20 个 AI 因子的 RankIC 均值稳定在 3% 以上，ICIR 均值超过 0.35，表明其预测能力具备显著的有效性与持续性。同时，多头组合的年化超额收益波动比均值达到 0.71，而多空对冲组合的收益波动比均值为 1.56，印证了这些因子在区分未来优胜股与劣势股方面的能力。

表19：AI 价值因子统计指标

因子名称	因子方向	RankIC 均值	ICIR	年化 多头超额	多头超额 收益波动比	年化 多空超额	多空超额 收益波动比
FCFEP_TTM	正向	1.58%	0.39	3.96%	0.60	5.04%	1.10
EBITDA_P_TTM	正向	3.71%	0.37	5.17%	0.61	14.90%	1.54
FCFFP_TTM	正向	1.25%	0.30	3.91%	0.56	3.39%	0.76
GP_TTM	正向	3.71%	0.40	6.33%	0.84	18.77%	2.32
TBPR_LF	正向	3.99%	0.39	6.53%	0.76	19.03%	2.22
OPP_TTM	正向	3.47%	0.38	5.89%	0.71	10.62%	1.06
CSHREVP_TTM	正向	2.75%	0.30	5.02%	0.80	17.11%	2.60
REP_LF	正向	3.97%	0.43	5.38%	0.63	11.48%	1.22
OEP_TTM	正向	2.15%	0.34	5.55%	0.67	5.76%	0.81
EBITP_TTM	正向	3.39%	0.36	5.30%	0.64	11.75%	1.21
TCIP_TTM	正向	3.38%	0.38	6.11%	0.76	10.72%	1.09
NWCP_LF	正向	1.21%	0.29	2.21%	0.41	2.21%	0.51
CEP_TTM	正向	4.07%	0.40	7.27%	0.84	17.11%	1.66
CGP_TTM	正向	3.44%	0.39	6.50%	0.85	17.97%	2.51
TPP_TTM	正向	3.55%	0.38	5.81%	0.69	11.32%	1.12
TPICP_LF	正向	3.16%	0.28	3.52%	0.44	14.71%	1.69
NIEOAP_TTM	正向	3.04%	0.36	6.81%	0.82	9.86%	1.05
NOAP_LF	正向	2.78%	0.26	2.86%	0.35	11.68%	1.45
TOTAL_YIELD_PLUS_TTM	正向	3.10%	0.40	6.97%	1.07	15.47%	2.62
REINVESTMENT_YIELD_TTM	负向	3.12%	0.40	6.95%	1.07	16.26%	2.72

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2025 年 6 月 30 日

深入分析 AI 的创造过程，我们发现其贡献主要体现在两大方面：因子创新与因子增强。一方面，AI 能够生成与所有样例因子相关性均较低的新颖因子，开拓了传统价值投资的边界；另一方面，AI 也能在经典因子的逻辑基础上进行优化，生成相关性虽高但选股效果更为出色的增强型因子。下方的相关性矩阵清晰地揭示了 AI 生成因子与样例因子之间的关系，为我们进一步筛选和理解这些新因子提供了实证依据。

表20：AI 价值因子与样例价值因子的因子值秩相关性

因子名称	样例价值因子					
	EP_TTM	EP_TTM_DEDUCTED	BP_LF	SP_TTM	OCFP_TTM	DIVIDEND_YIELD_TTM
FCFEP_TTM	27.65%	25.49%	16.38%	20.48%	25.17%	24.00%
EBITDA_P_TTM	77.99%	67.85%	55.47%	60.80%	49.26%	49.21%
FCFFP_TTM	21.70%	18.98%	9.71%	11.53%	42.02%	17.70%
GP_TTM	54.57%	50.83%	55.80%	75.28%	43.63%	41.71%
TBPR_LF	35.39%	29.56%	89.32%	49.53%	25.59%	36.42%
OPP_TTM	95.04%	91.14%	34.83%	33.97%	37.43%	59.91%
CSHREVP_TTM	28.98%	22.38%	58.88%	96.37%	38.73%	23.77%
REP_LF	61.61%	57.95%	69.12%	49.45%	36.66%	57.67%
OEP_TTM	63.21%	57.95%	18.14%	15.63%	20.07%	38.08%
EBITP_TTM	89.42%	80.18%	42.49%	47.53%	42.08%	53.53%
TCIP_TTM	97.06%	90.22%	33.58%	30.86%	35.18%	59.98%
NWCP_LF	7.91%	4.46%	36.19%	25.52%	-13.61%	9.66%
CEP_TTM	84.25%	75.00%	54.56%	54.15%	48.94%	54.82%
CGP_TTM	34.31%	30.70%	49.98%	65.15%	70.50%	30.83%
TPP_TTM	97.03%	89.59%	36.95%	36.15%	38.15%	60.09%
TPICP_LF	23.89%	14.94%	88.01%	71.05%	32.25%	21.75%
NIEOAP_TTM	86.30%	92.31%	23.31%	27.68%	35.74%	56.78%
NOAP_LF	21.60%	13.51%	83.23%	64.30%	27.29%	19.49%
TOTAL_YIELD_PLUS_TTM	57.43%	56.46%	31.81%	23.62%	32.62%	95.61%
REINVESTMENT_YIELD_TTM	48.26%	49.33%	33.96%	24.84%	27.02%	96.67%

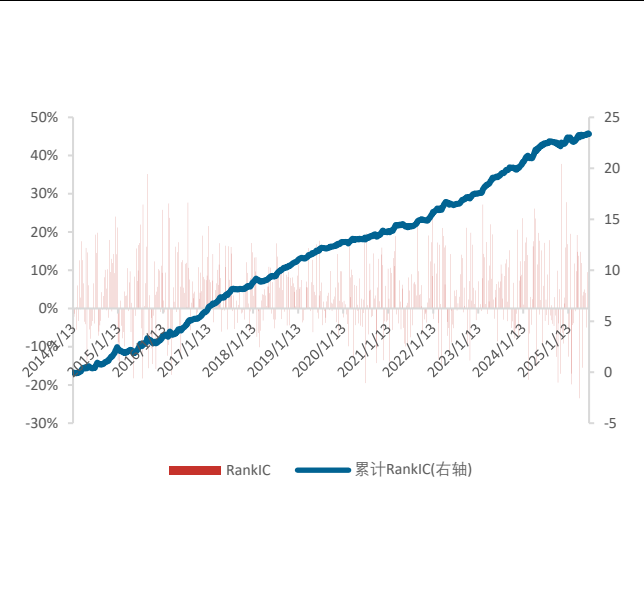
数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2025 年 6 月 30 日

REP_LF 因子（留存收益/总市值）是 AI 因子创新的典型代表。该因子与所有样例因子的相关性均未超过 70%，展现了其独特的逻辑视角。留存收益是公司自创立以来所有净利润的累积，扣除已派发股利后的部分，它直接反映了公司内源性资本的积累和再投资能力。REP_LF 因子将这一历史盈利的“存量”与公司当前的“市值”进行比较，旨在衡量市场对公司长期价值创造能力的定价是否充分。高比率通常意味着公司历史盈利丰厚，安全边际较高，但当前可能被市场低估。

从回测结果来看，REP_LF 因子的表现十分稳健。经市值与行业中性化后，该因子在整个回测区间的周频 RankIC 均值达到 3.97%，ICIR 为 0.43。其多头组合实现了 5.38% 的年化超额收益，而多空对冲组合的年化超额收益更是达到了 11.48%，对应的收益波动比为 1.22，展示了其在全市场范围内的优异选股能力。

图15: REP_LF 因子每期 RankIC 与累计 RankIC



数据来源: Wind、东吴证券研究所
数据日期: 2014 年 1 月 13 日至 2025 年 6 月 30 日

图16: REP_LF 因子多头组合净值与多头超额净值



数据来源: Wind、东吴证券研究所
数据日期: 2014 年 1 月 6 日至 2025 年 6 月 30 日

表21: REP_LF 因子 IC 统计指标

	RankIC	ICIR	t 统计值	胜率
2014	4.41%	0.50	3.57	56.86%
2015	1.90%	0.16	1.13	55.77%
2016	5.41%	0.54	3.83	74.00%
2017	4.95%	0.73	5.20	68.63%
2018	4.56%	0.69	4.89	72.55%
2019	3.78%	0.65	4.71	69.23%
2020	1.96%	0.23	1.61	60.78%
2021	3.13%	0.39	2.79	65.38%
2022	4.24%	0.42	2.98	66.00%
2023	5.55%	0.67	4.76	70.00%
2024	4.96%	0.39	2.79	61.54%
2025	1.93%	0.15	0.78	57.69%
全区间	3.97%	0.43	10.32	65.14%

数据来源: Wind、东吴证券研究所
数据日期: 2013 年 12 月 31 日至 2025 年 6 月 30 日

表22: REP_LF 因子分年度多头超额收益风险绩效指标

年份	年化收益	年化波动	最大回撤	收益波动比	收益回撤比
2014	13.82%	8.17%	-6.63%	1.69	2.09
2015	-16.78%	13.01%	-19.86%	-1.29	-0.85
2016	11.16%	7.51%	-8.45%	1.49	1.32
2017	18.06%	6.46%	-5.65%	2.79	3.20
2018	9.58%	5.86%	-8.77%	1.63	1.09
2019	0.16%	4.96%	-5.23%	0.03	0.03
2020	1.22%	7.17%	-6.64%	0.17	0.18
2021	3.08%	8.21%	-10.26%	0.38	0.30
2022	14.25%	9.85%	-10.70%	1.45	1.33
2023	5.86%	7.21%	-6.50%	0.81	0.90
2024	12.72%	11.41%	-11.00%	1.12	1.16
2025	-7.90%	9.99%	-9.38%	-0.79	-0.84
全区间	5.38%	8.55%	-19.86%	0.63	0.27

数据来源: Wind、东吴证券研究所
数据日期: 2013 年 12 月 31 日至 2025 年 6 月 30 日

图17: REP_LF 因子多空超额净值



数据来源: Wind、东吴证券研究所
数据日期: 2014 年 1 月 6 日至 2025 年 6 月 30 日

表23: REP_LF 因子分年度多空超额收益风险绩效指标

年份	年化 收益	年化 波动	最大 回撤	收益 波动比	收益 回撤比
2014	17.65%	6.79%	-4.85%	2.60	3.64
2015	-4.09%	13.73%	-13.54%	-0.30	-0.30
2016	22.08%	7.56%	-3.81%	2.92	5.79
2017	29.43%	5.15%	-2.32%	5.71	12.70
2018	25.85%	6.17%	-5.28%	4.19	4.89
2019	8.71%	6.19%	-6.23%	1.41	1.40
2020	10.21%	8.45%	-8.30%	1.21	1.23
2021	0.21%	8.30%	-9.68%	0.03	0.02
2022	7.41%	8.75%	-8.67%	0.85	0.85
2023	8.95%	7.65%	-8.94%	1.17	1.00
2024	12.84%	16.39%	-22.33%	0.78	0.58
2025	-6.29%	12.66%	-9.01%	-0.50	-0.70
全区间	11.48%	9.41%	-22.33%	1.22	0.51

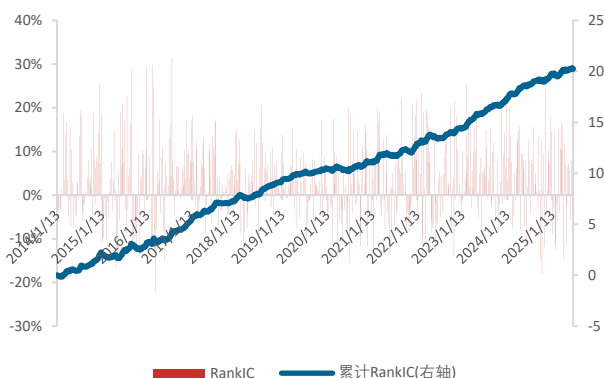
数据来源: Wind、东吴证券研究所
数据日期: 2013 年 12 月 31 日至 2025 年 6 月 30 日

除了创造全新因子, AI 同样擅长对现有因子逻辑进行深度优化, CGP_TTM 因子便是例证。该因子的构建方式为“(销售商品、提供劳务收到的现金 TTM- 购买商品、接受劳务支付的现金 TTM) / 总市值”, 本质上是衡量公司核心购销活动产生的“现金毛利”。

相较于样例因子中基于权责发生制的经营活动现金流(OCFP_TTM), CGP_TTM 剔除了更多非核心经营活动现金流的干扰, 也规避了应收、应付、存货等会计科目的影响, 从而更纯粹、更真实地反映了企业主营业务的“造血”能力。在因子相关性上, 它与 OCFP_TTM 的相关性达到 70.5%, 但在大部分回测指标上实现了超越, 可视作对 OCFP_TTM 的一次成功“增强”。

经过标准化的回测, CGP_TTM 因子全区间的 RankIC 均值为 3.44%, ICIR 为 0.39。其多头组合年化超额收益为 6.50%, 而多空组合表现尤为亮眼, 年化超额收益为 17.97%, 收益波动比为 2.51, 充分说明该因子在识别估值合理的公司方面具有强大的判别力。

图18: CGP_TTM 因子每期 RankIC 与累计 RankIC



数据来源: Wind、东吴证券研究所

数据日期: 2014 年 1 月 13 日至 2025 年 6 月 30 日

图19: CGP_TTM 因子多头组合净值与多头超额净值



数据来源: Wind、东吴证券研究所

数据日期: 2014 年 1 月 6 日至 2025 年 6 月 30 日

表24: CGP_TTM 因子 IC 统计指标

	RankIC	ICIR	t 统计值	胜率
2014	4.04%	0.49	3.53	60.78%
2015	1.44%	0.14	0.98	50.00%
2016	4.90%	0.45	3.21	76.00%
2017	3.98%	0.57	4.07	68.63%
2018	3.57%	0.50	3.61	70.59%
2019	2.61%	0.46	3.28	65.38%
2020	1.27%	0.14	1.02	52.94%
2021	2.87%	0.32	2.34	57.69%
2022	3.60%	0.35	2.45	62.00%
2023	5.26%	0.67	4.73	72.00%
2024	4.98%	0.54	3.87	75.00%
2025	2.37%	0.24	1.20	53.85%
全区间	3.44%	0.39	9.48	64.12%

数据来源: Wind、东吴证券研究所

数据日期: 2013 年 12 月 31 日至 2025 年 6 月 30 日

表25: CGP_TTM 因子分年度多头超额收益风险绩效指标

年份	年化收益	年化波动	最大回撤	收益波动比	收益回撤比
2014	17.59%	7.19%	-4.85%	2.45	3.62
2015	-10.36%	11.47%	-14.04%	-0.90	-0.74
2016	9.60%	6.81%	-9.36%	1.41	1.03
2017	16.67%	5.74%	-3.51%	2.90	4.75
2018	8.53%	5.12%	-6.52%	1.67	1.31
2019	0.95%	4.54%	-5.73%	0.21	0.17
2020	-0.42%	7.19%	-6.27%	-0.06	-0.07
2021	2.42%	8.09%	-10.56%	0.30	0.23
2022	17.33%	9.30%	-9.95%	1.86	1.74
2023	4.09%	6.39%	-5.49%	0.64	0.74
2024	13.93%	9.66%	-8.62%	1.44	1.62
2025	-4.18%	7.99%	-6.79%	-0.52	-0.62
全区间	6.50%	7.69%	-14.88%	0.85	0.44

数据来源: Wind、东吴证券研究所

数据日期: 2013 年 12 月 31 日至 2025 年 6 月 30 日

图20: CGP_TTM 因子多空超额净值



数据来源: Wind、东吴证券研究所
数据日期: 2014 年 1 月 6 日至 2025 年 6 月 30 日

表26: CGP_TTM 因子分年度多空超额收益风险绩效指标

年份	年化 收益	年化 波动	最大 回撤	收益 波动比	收益 回撤比
2014	28.20%	6.65%	-2.46%	4.24	11.47
2015	13.48%	12.23%	-8.13%	1.10	1.66
2016	23.08%	7.01%	-3.56%	3.29	6.47
2017	20.90%	4.65%	-1.70%	4.49	12.29
2018	25.00%	5.08%	-2.61%	4.92	9.59
2019	9.78%	4.59%	-2.80%	2.13	3.50
2020	5.88%	7.33%	-6.89%	0.80	0.85
2021	8.67%	7.44%	-6.31%	1.17	1.37
2022	16.32%	8.08%	-5.71%	2.02	2.86
2023	20.68%	5.37%	-3.84%	3.85	5.39
2024	24.61%	7.39%	-7.01%	3.33	3.51
2025	4.30%	6.37%	-2.90%	0.67	1.48
全区间	17.97%	7.16%	-8.13%	2.51	2.21

数据来源: Wind、东吴证券研究所
数据日期: 2013 年 12 月 31 日至 2025 年 6 月 30 日

3.4. AI 质量因子挖掘：从盈利能力到运营效率的多维度探索

质量因子旨在识别那些具有稳健的盈利能力、健康的财务状况和高效的运营管理能力
力的公司。为引导 AI 在这一领域进行探索，我们向其提供了涵盖盈利能力（如每股收
益 EPS、净资产收益率 ROE）、资产质量（如每股净资产 BPS）和现金流质量（如股
经营现金流 OCFPS）等维度的经典质量因子作为样例。

表27: 样例质量因子释义与表达式

因子名称	因子释义	表达式
EPS_QR	归属于母公司所有者的净利润_单季度 / 最新总股本	op([quarter(get('npparentcompanyowners')), get('paidincapital')], 'val0 / val1')
EPS_TTM	归属于母公司所有者的净利润_TTM / 最新总股本	op([ttm(get('npparentcompanyowners')), get('paidincapital')], 'val0 / val1')
BPS_LF	归属母公司所有者权益(或股东权益)合计 / 最新总股本	op([get('sewithoutmi'), get('paidincapital')], 'val0 / val1')
ROE_QR	2 * 归属于母公司所有者的净利润_单季度 / （当季度 股东权益合计+ 上季度股东权益合计）	op([quarter(get('npparentcompanyowners')), op([get('sewithoutmi'), refq(get('sewithoutmi'), 1)], '(val0 + val1) / 2')], 'val0 / val1')
ROE_TTM	归属于母公司所有者的净利润_TTM / 股东权益合计	op([ttm(get('npparentcompanyowners')), op([get('sewithoutmi'), refq(get('sewithoutmi'), 1)], '(val0 + val1) / 2')], 'val0 / val1')
OCFPS_QR	经营活动现金净流量_单季度 / 最新总股本	op([quarter(get('netoperatcashflow')), get('paidincapital')], 'val0 / val1')
OCFPS_TTM	经营活动现金净流量_TTM / 最新总股本	op([ttm(get('netoperatcashflow')), get('paidincapital')], 'val0 / val1')
OCF_QUALITY_TTM	(过去一年经营活动产生的现金流量净额 - 过去一年 营业利润)/总资产	op([ttm(get('netoperatcashflow')), ttm(get('operatingprofit')), get('totalassets')], '(val0 - val1) / val2')

GROSS_PROFIT_QR	(营业收入_单季度 - 营业成本_单季度) / 总资产	$\frac{\text{op}([\text{quarter}(\text{get}('operatingrevenue')), \text{quarter}(\text{get}('operatingcost'))], \text{get}('totalassets')], '(\text{val0} - \text{val1}) / \text{val2}')$
GROSS_PROFIT_TTM	(营业收入_TTM - 营业成本_TTM) / 总资产	$\frac{\text{op}([\text{ttm}(\text{get}('operatingrevenue')), \text{ttm}(\text{get}('operatingcost'))], \text{get}('totalassets')], '(\text{val0} - \text{val1}) / \text{val2}')$

数据来源：东吴证券研究所整理

表28：样例质量因子统计指标

因子名称	RankIC 均值	ICIR	年化多头超额	多头超额收益波动比	年化多空超额	多空超额收益波动比
EPS_QR	2.82%	0.31	2.96%	0.49	14.56%	1.65
EPS_TTM	2.08%	0.21	-0.94%	-0.15	7.00%	0.73
BPS_LF	1.43%	0.19	-1.58%	-0.29	7.27%	0.95
ROE_QR	2.39%	0.28	3.55%	0.70	11.77%	1.33
ROE_TTM	1.53%	0.17	0.15%	0.03	4.84%	0.50
OCFPS_QR	1.47%	0.36	0.67%	0.12	5.67%	1.24
OCFPS_TTM	2.11%	0.36	0.85%	0.14	9.29%	1.73
OCF_QUALITY_TTM	0.48%	0.10	-0.60%	-0.12	3.62%	0.63
GROSS_PROFIT_QR	1.63%	0.21	3.05%	0.54	11.85%	1.54
GROSS_PROFIT_TTM	1.03%	0.14	1.13%	0.20	7.16%	0.95

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2025 年 6 月 30 日

基于这些样例，AI 系统性地生成了 60 个新的质量因子。通过分析这些因子的逻辑表达式，我们观察到 AI 的探索覆盖了多个维度，不仅包括对传统盈利指标的深化，还延伸至资本结构、资产周转效率以及现金流转换等多个层面。

表29：AI 质量因子释义

因子名称	因子逻辑
ACEPS_QR	$((\text{单季度归母净利润} + \text{单季度经营活动现金流净额}) / 2) / \text{最新总股本}$
CAOPS_QR	$(\text{单季度营业利润} - (\text{单季度归母净利润} - \text{单季度经营活动现金流量净额})) / \text{最新总股本}$
CCEPS_QR	$(\text{单季度扣非归母净利润} + \text{单季度折旧与摊销}) / \text{最新总股本}$
CCPATPS_QR	$(\text{单季度销售商品、提供劳务收到的现金} - \text{单季度购买商品、接受劳务支付的现金} - \text{单季度支付给职工以及为职工支付的现金} - \text{单季度支付的各项税费}) / \text{最新总股本}$
CCPS_QR	$(\text{单季度销售商品、提供劳务收到的现金} - \text{单季度购买商品、接受劳务支付的现金} - \text{单季度支付给职工以及为职工支付的现金}) / \text{最新总股本}$
CFOITPS_QR	$(\text{单季度经营活动现金流量净额} + \text{单季度财务费用} + \text{单季度支付的各项税费}) / \text{最新总股本}$
CGPS_QR	$(\text{单季度销售商品、提供劳务收到的现金} - \text{单季度购买商品、接受劳务支付的现金}) / \text{最新总股本}$
COPPS_QR	$(\text{单季度毛利} - \text{单季度销售费用} - \text{单季度管理费用}) / \text{最新总股本}$
CROPPS_QR	$(\text{单季度营业利润} * (\text{单季度销售商品、提供劳务收到的现金} / \text{单季度营业总收入})) / \text{最新总股本}$
EBITDAPS_QR	$\text{单季度息税折旧摊销前利润} / \text{最新总股本}$
EBITPS_QR	$\text{单季度息税前利润} / \text{最新总股本}$

FCOPPS_QR	(单季度销售商品、提供劳务收到的现金 - 单季度购买商品、接受劳务支付的现金 - 单季度支付给职工以及为职工支付的现金 - 单季度支付其他与经营活动有关的现金) / 最新总股本
GPS_QR	单季度毛利 / 最新总股本
NOPATS_QR	(单季度营业利润 - (单季度利润总额 - 单季度净利润)) / 最新总股本
OEPS_QR	(单季度归母净利润 + 单季度折旧与摊销 - 单季度资本性支出) / 最新总股本
QOP_PS_QR	(单季度营业利润 * (单季度经营活动现金流净额 / 单季度归母净利润)) / 最新总股本
SGPS_QR	(单季度毛利 - 单季度销售费用) / 最新总股本
CORE_ROIC_QR	单季度扣非归母净利润 / 平均(归母股东权益 + 净债务)
CRE_LF	资本公积 / 归母股东权益
CROE_AVG_QR	单季度经营活动现金流量净额 / 平均归母股东权益
CROIC_QR	单季度经营活动现金流量净额 / 平均(归母股东权益 + 带息债务)
DROE_AVG_QR	单季度扣非归母净利润 / 平均归母股东权益
RE_RATIO_LF	留存收益 / 归母股东权益
ROIC_QR	单季度营业利润 / 平均(归母股东权益 + 带息债务)
ATO_QR	单季度营业总收入 / 平均总资产
CATO_AVG_QR	单季度销售商品、提供劳务收到的现金 / 平均总资产
CGPA_AVG_QR	(单季度销售商品、提供劳务收到的现金 - 单季度购买商品、接受劳务支付的现金) / 平均总资产
CORE_CASH_PROFIT_TO_ASSETS_QR	(单季度销售商品提供劳务收到的现金 - 单季度购买商品接受劳务支付的现金 - 单季度支付给职工的现金) / 期末总资产
CORE_CEROA_AVG_QR	(单季度扣非归母净利润 + 单季度折旧与摊销) / 平均总资产
CROA_AVG_QR	单季度经营活动现金流量净额 / 平均总资产
DROA_AVG_QR	单季度扣非净利润 / 平均总资产
GPA_AVG_QR	单季度毛利 / 平均总资产
OPA_AVG_QR	单季度营业利润 / 平均总资产
RNOA_AVG_QR	单季度营业利润 / 平均净经营资产
SFC_ROA_QR	(单季度经营活动现金流净额 + 单季度筹资活动现金流净额) / 平均总资产
ART_QR	单季度营业总收入 / 平均应收账款
CASH_CONVERSION_EFFICIENCY_QR	单季度经营活动现金流量净额 / 单季度销售商品、提供劳务收到的现金
CASH_OPERATING_PROFIT_MARGIN_QR	(销售商品、提供劳务收到的现金 - 购买商品、接受劳务支付的现金 - 支付给职工以及为职工支付的现金 - 支付的各项税费) / 销售商品、提供劳务收到的现金 (单季度)
CFO_NP_RATIO_QR	单季度经营活动现金流量净额 / 单季度归属于母公司所有者的净利润
CFO_PAYOUT_COVERAGE_TTM	经营活动现金流量净额(TTM) / 分配股利、利润或偿付利息支付的现金(TTM)
CGE_QR	(单季度销售商品收到的现金 - 单季度购买商品支付的现金) / (单季度销售费用 + 单季度管理费用)
CGPM_QR	(单季度销售商品、提供劳务收到的现金 - 单季度购买商品、接受劳务支付的现金) / 单季度销售商品、提供劳务收到的现金
COP_QR	(单季度销售商品提供劳务收到的现金 - 单季度购买商品接受劳务支付的现金) / (单季度支付给职工的现金 + 单季度支付其他与经营活动有关的现金)
FCFE_NP_RATIO_TTM	股权自由现金流(TTM) / 归母净利润(TTM)
FCFF_DEBT_COVERAGE_TTM	企业自由现金流量(TTM) / 总负债
MARGINAL_PROFIT_ON_CAPEX_QR	当季营业利润环比变动额 / 上一季度资本性支出
NPM_QR	单季度归母净利润 / 单季度营业总收入
NWC_TURNOVER_QR	单季度营业总收入 / 平均净营运资本
OCF_CAPEX_COVERAGE_QR	单季度经营活动现金流量净额 / 单季度资本性支出
OCF_INV_COVERAGE_QR	单季度经营活动现金流量净额 / 单季度投资活动现金流出小计

OCF_SG_A_PRODUCTIVITY_QR	单季度经营活动现金流净额 / (单季度销售费用 + 单季度管理费用)
OCFC_TTM	经营活动现金流量净额(TTM) / 流动负债合计
OCFGP_RATIO_QR	单季度经营活动现金流量净额 / 单季度毛利
OCFL_TTM	经营活动现金流量净额(TTM) / 负债合计
OCFS_RATIO_QR	单季度经营活动现金流量净额 / 单季度营业总收入
OCFTC_QR	单季度经营活动现金流量净额 / 单季度营业总成本
OPM_QR	单季度营业利润 / 单季度营业总收入
PROFIT_CASH_EFFICIENCY_QR	单季度营业利润 / 单季度销售商品、提供劳务收到的现金
SG_PRODUCTIVITY_QR	单季度毛利 / (单季度销售费用 + 单季度管理费用)
WCROI_QR	单季度营业利润 / 平均净营运资本

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2025 年 6 月 30 日

备注：单季度因子（QR 结尾）分母中的“平均”代表当季度与上季度的均值

对这批 AI 生成因子的回测结果显示，其中一部分因子具备了有效的选股潜力。具体来看，在 60 个因子中，有 5 个因子的年化多头超额收益超过了 3.55%，7 个因子的多头组合收益波动比大于 0.7。在多空对冲组合中，有 3 个因子的年化超额收益超过 14.5%，10 个因子的收益波动比大于 1.73。这些数据表明，AI 生成的因子库中包含了一批值得进一步研究的因子。

表30：AI 质量因子统计指标

因子名称	因子方向	RankIC 均值	ICIR	年化 多头超额	多头超额 收益波动比	年化 多空超额	多空超额 收益波动比
ACEPS_QR	正向	2.03%	0.34	1.04%	0.18	8.23%	1.47
CAOPS_QR	正向	1.64%	0.37	0.73%	0.12	6.68%	1.36
CCEPS_QR	正向	1.94%	0.25	1.40%	0.22	8.16%	1.06
CCPATPS_QR	正向	1.48%	0.32	1.09%	0.21	6.46%	1.34
CCPS_QR	正向	1.70%	0.32	1.12%	0.21	7.19%	1.43
CFOITPS_QR	正向	1.68%	0.35	0.99%	0.19	7.36%	1.50
CGPS_QR	正向	1.83%	0.30	1.41%	0.28	9.70%	1.86
COPPS_QR	正向	2.58%	0.31	2.77%	0.46	15.92%	2.20
CROPPS_QR	正向	2.77%	0.32	2.27%	0.43	12.38%	1.63
EBITDAPS_QR	正向	2.11%	0.28	0.70%	0.12	9.99%	1.39
EBITPS_QR	正向	2.68%	0.32	1.22%	0.21	14.23%	1.89
FCOPPS_QR	正向	1.58%	0.35	1.49%	0.27	7.21%	1.43
GPS_QR	正向	2.15%	0.26	1.08%	0.19	14.62%	2.06
NOPATS_QR	正向	2.72%	0.30	1.98%	0.33	12.66%	1.56
OEPS_QR	正向	1.17%	0.24	0.87%	0.15	3.28%	0.59
QOP_PS_QR	正向	1.48%	0.37	0.21%	0.04	5.19%	1.24
SGPS_QR	正向	2.39%	0.29	2.06%	0.35	16.64%	2.29
CORE_ROIC_QR	正向	2.23%	0.26	2.51%	0.46	9.53%	1.09

CRE_LF	负向	1.34%	0.27	2.14%	0.63	5.50%	1.16
CROE_AVG_QR	正向	1.21%	0.35	1.31%	0.29	5.78%	1.40
CROIC_QR	正向	1.26%	0.34	2.69%	0.80	6.51%	1.51
DROE_AVG_QR	正向	2.44%	0.28	4.32%	0.83	11.93%	1.33
RE_RATIO_LF	正向	2.05%	0.30	2.83%	0.65	6.39%	0.80
ROIC_QR	正向	2.31%	0.26	3.35%	0.64	12.31%	1.39
ATO_QR	正向	1.21%	0.25	1.19%	0.31	8.94%	1.78
CATO_AVG_QR	正向	1.14%	0.27	1.24%	0.31	7.97%	1.78
CGPA_AVG_QR	正向	1.28%	0.26	3.78%	0.90	9.36%	1.75
CORE_CASH_PROFIT_TO_ASSETS_QR	正向	1.28%	0.29	3.22%	0.79	7.89%	1.55
CORE_CEROA_AVG_QR	正向	1.76%	0.22	3.77%	0.71	10.12%	1.20
CROA_AVG_QR	正向	1.27%	0.33	2.51%	0.76	6.61%	1.52
DROA_AVG_QR	正向	2.20%	0.25	4.13%	0.73	11.37%	1.23
GPA_AVG_QR	正向	1.64%	0.21	3.04%	0.54	12.09%	1.56
OPA_AVG_QR	正向	2.24%	0.25	3.61%	0.68	12.15%	1.37
RNOA_AVG_QR	正向	2.20%	0.26	2.14%	0.42	9.99%	1.17
SFC_ROA_QR	正向	0.75%	0.21	1.08%	0.31	4.39%	1.10
ART_QR	正向	1.40%	0.26	1.00%	0.20	8.68%	1.98
CASH_CONVERSION_EFFICIENCY_QR	正向	1.09%	0.34	1.60%	0.48	6.27%	1.48
CASH_OPERATING_PROFIT_MARGIN_QR	正向	0.94%	0.29	1.25%	0.39	5.62%	1.35
CFO_NP_RATIO_QR	正向	0.64%	0.25	-0.67%	-0.16	2.07%	0.61
CFO_PAYOUT_COVERAGE_TTM	正向	0.89%	0.28	2.39%	0.70	5.79%	1.48
CGE_QR	正向	1.28%	0.36	1.47%	0.31	6.88%	1.63
CGPM_QR	正向	0.61%	0.18	1.59%	0.48	5.93%	1.45
COP_QR	正向	1.37%	0.36	0.29%	0.06	6.37%	1.37
FCFE_NP_RATIO_TTM	正向	0.56%	0.23	-0.70%	-0.17	1.04%	0.31
FCFF_DEBT_COVERAGE_TTM	正向	1.07%	0.28	1.16%	0.34	3.38%	0.72
MARGINAL_PROFIT_ON_CAPEX_QR	正向	0.66%	0.22	-0.12%	-0.03	4.32%	1.08
NPM_QR	正向	1.76%	0.24	0.77%	0.20	7.32%	0.92
NWC_TURNOVER_QR	正向	0.61%	0.23	-1.65%	-0.37	-0.14%	-0.04
OCF_CAPEX_COVERAGE_QR	正向	0.99%	0.34	0.36%	0.08	4.14%	1.12
OCF_INV_COVERAGE_QR	正向	0.93%	0.33	1.06%	0.25	5.33%	1.48
OCF_SG_A_PRODUCTIVITY_QR	正向	1.17%	0.35	0.78%	0.17	6.12%	1.46
OCFC_TTM	正向	1.51%	0.30	1.52%	0.38	7.19%	1.36
OCFGP_RATIO_QR	正向	0.90%	0.31	-1.29%	-0.31	3.53%	1.02
OCFL_TTM	正向	1.48%	0.28	2.10%	0.48	7.23%	1.33
OCFS_RATIO_QR	正向	1.15%	0.35	0.27%	0.07	4.76%	1.12
OCFTC_QR	正向	1.21%	0.35	-0.11%	-0.03	4.37%	0.98
OPM_QR	正向	1.87%	0.26	0.89%	0.24	7.66%	0.98
PROFIT_CASH_EFFICIENCY_QR	正向	1.76%	0.24	0.70%	0.19	7.64%	0.97
SG_PRODUCTIVITY_QR	正向	1.86%	0.31	1.26%	0.26	10.43%	1.77
WCROI_QR	正向	1.19%	0.23	-0.78%	-0.16	1.26%	0.30

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2025 年 6 月 30 日

在 AI 生成的因子中，我们发现模型能够基于经典盈利能力指标进行逻辑上的延伸。以 GPS_QR（单季度毛利/最新总股本）和 COPPS_QR（(单季度毛利 - 销售费用 - 管理费用)/最新总股本）为例，这两个因子体现了对盈利能力的不同层次的刻画。

GPS_QR 关注的是企业产品或服务的直接盈利空间，而 COPPS_QR 则在毛利的基础上，进一步扣除了为实现销售和维持公司运营所必需的费用。理论上，后者旨在更聚焦于核心经营环节的盈利结果，能够过滤掉部分毛利高但运营效率不佳的公司。从因子相关性数据看，这两个新因子与作为样例的 EPS_QR 因子有较高的关联度。

表31：GPS_QR、COPPS_QR 因子与样例质量因子秩相关性

	EPS_QR	EPS_TTM	BPS_LF	ROE_QR	ROE_TTM	OCFPS_QR	OCFPS_TTM
COPPS_QR	82.62%	72.33%	55.98%	70.55%	59.06%	30.42%	45.73%
GPS_QR	71.86%	68.48%	64.35%	57.06%	51.12%	25.84%	45.37%

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2025 年 6 月 30 日

在选股效果上，COPPS_QR 与 EPS_QR 的 ICIR 及多头超额收益表现相近。值得注意的是，COPPS_QR 在多空超额收益波动比这一指标上表现出一定优势。这可能表明，相较于 EPS_QR，COPPS_QR 因子在识别潜在表现较差的股票方面具备更强的区分能力。

表32：GPS_QR、COPPS_QR 因子与 EPS_QR 因子统计指标对比

因子名称	RankIC 均值	ICIR	年化多头超额	多头超额收益波动比	年化多空超额	多空超额收益波动比
COPPS_QR	2.58%	0.31	2.77%	0.46	15.92%	2.20
GPS_QR	2.15%	0.26	1.08%	0.19	14.62%	2.06
EPS_QR	2.82%	0.31	2.96%	0.49	14.56%	1.65

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2025 年 6 月 30 日

除了对传统盈利指标的优化，AI 也展现出构建新型运营效率指标的能力。在生成的因子中，我们注意到模型不再局限于以股本、股东权益或总资产作为分母进行标准化，而是开始探索使用其他财务项目。

ART_QR 因子（单季度营业总收入 / 平均应收账款），即应收账款周转率，便是一个很好的例子。该因子衡量企业在一个季度内收回应收账款的效率。较高的周转率通常

意味着企业信用政策得当、回款速度快，从而资产流动性较强、潜在的坏账风险较低。它从资产管理的角度评估了公司的运营质量。

从回测绩效上看，ART_QR 的多空组合表现较为突出，其收益波动比为 1.98。更重要的是，该因子与所有样例质量因子的相关性都处于较低水平，说明它提供了一个相对独立的选股视角，这对于构建多元化的因子模型具有积极意义。

图21: ART_QR 因子多空超额净值



数据来源：Wind、东吴证券研究所

数据日期：2014 年 1 月 6 日至 2025 年 6 月 30 日

表33: ART_QR 因子分年度多空超额收益风险绩效指标

年份	年化收益	年化波动	最大回撤	收益波动比	收益回撤比
2014	-1.54%	3.79%	-4.95%	-0.41	-0.31
2015	3.83%	6.71%	-4.05%	0.57	0.95
2016	13.41%	4.61%	-2.25%	2.91	5.97
2017	11.08%	3.41%	-1.33%	3.25	8.32
2018	18.81%	3.83%	-2.43%	4.91	7.74
2019	13.29%	3.79%	-2.83%	3.51	4.70
2020	6.96%	4.50%	-2.45%	1.55	2.84
2021	5.75%	5.04%	-6.85%	1.14	0.84
2022	4.06%	4.52%	-3.01%	0.90	1.35
2023	11.69%	2.84%	-1.69%	4.12	6.91
2024	6.08%	4.18%	-4.57%	1.45	1.33
2025	4.01%	3.94%	-1.78%	1.02	2.25
全区间	8.68%	4.39%	-6.85%	1.98	1.27

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2025 年 6 月 30 日

表34: ART_QR 因子与样例质量因子秩相关性

	EPS_QR	EPS_TTM	BPS_LF	ROE_QR	ROE_TTM	OCFPS_QR	OCFPS_TTM
ART_QR	21.95%	16.24%	5.39%	22.93%	16.52%	25.77%	28.79%

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2025 年 6 月 30 日

3.5. AI 成长因子挖掘：探索盈利增长的多维定义

成长性驱动股价长期上行的关键因素之一。在传统因子体系中，成长因子通常通过计算企业关键财务指标的同比增长率来构建。为引导 AI 在这一方向上进行探索，我们提供了以各口径下的利润和收入单季度同比增速为代表的一组样例因子。这些基础成长因子代表了市场对企业增长情况的常规度量方式。

表35：样例成长因子释义与表达式

因子名称	因子释义	表达式
NETPROFIT_QR_YOY	归属于母公司所有者的净利润_单季度_同比增速	yoy(quarter(get('npparentcompanyowners')))
DEDUCTED_NETPROFIT_QR_YOY	扣除非经常性损益后的归母净利润_单季度_同比增速	yoy(quarter(get('npdeductnonrecurringpl')))
OPERATINGPROFIT_QR_YOY	营业利润_单季度_同比增速	yoy(quarter(get('operatingprofit')))
REVENUE_QR_YOY	营业收入_单季度_同比增速	yoy(quarter(get('operatingrevenue')))

数据来源：东吴证券研究所整理

表36：样例成长因子统计指标

因子名称	RankIC 均值	ICIR	年化多头超额	多头超额收益波动比	年化多空超额	多空超额收益波动比
NETPROFIT_QR_YOY	1.84%	0.32	5.88%	1.60	12.26%	1.92
DEDUCTED_NETPROFIT_QR_YOY	1.81%	0.33	6.26%	1.71	12.40%	1.95
OPERATINGPROFIT_QR_YOY	1.68%	0.30	4.12%	1.12	10.49%	1.70
REVENUE_QR_YOY	1.21%	0.22	1.20%	0.31	9.82%	1.69

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2025 年 6 月 30 日

在接收了样例因子后，AI 系统生成了超过 20 个新的成长因子。分析这些因子的构造逻辑可以发现，AI 的探索并不仅限于对样例的模仿，而是展现出对“成长”这一概念多维度的理解。其尝试方向涵盖了从利润定义的扩展（如 EBITDA、综合收益等），到盈利质量的调整（如考虑费用、折旧、现金流等），再到部分创新性成长概念的提出。

表37：AI 成长因子释义与表达式

因子名称	因子逻辑
EBITDA_QR_YOY	单季度息税折旧摊销前利润(EBITDA)的同比增长率
GROSSPROFIT_QR_YOY	单季度毛利的同比增长率
CI_QR_YOY	单季度归属于母公司所有者的综合收益总额的同比增长率
CDCFG_QR_YOY	(单季度毛利 - 单季度销售费用 - 单季度管理费用 - 单季度资本性支出) 的同比增长率
CORE_PROFIT_ENGINE_GROWTH_QR	(单季度毛利 - 单季度管理费用) 的同比增长率
CORE_PROFIT_QR_YOY	(单季度毛利 - 单季度销售费用 - 单季度管理费用) 的同比增长率
SGP_QR_YOY	(单季度毛利 - 单季度销售费用) 的同比增长率
CE_QR_YOY	(单季度归母净利润 + 单季度折旧与摊销) 的同比增长率
FINANCING_NEUTRAL_PROFIT_QR_YOY	(单季度归母净利润 + 单季度财务费用) 的同比增长率
GAP_QR_YOY	(单季度归母净利润 + 单季度销售费用 + 单季度管理费用) 的同比增长率
OWNER_EARNINGS_QR_YOY	(单季度归母净利润 + 单季度折旧与摊销 - 单季度资本性支出) 的同比增长率
CORE_CE_QR_YOY	(单季度扣非归母净利润 + 单季度折旧与摊销) 的同比增长率
CORE_EARNINGS_QR_YOY	(单季度扣非归母净利润 + 单季度财务费用) 的同比增长率
CORE_GAP_QR_YOY	(单季度扣非归母净利润 + 单季度销售费用 + 单季度管理费用) 的同比增长率
CORE_OWNER_EARNINGS_QR_YOY	(单季度扣非归母净利润 + 单季度折旧与摊销 - 单季度资本性支出) 的同比增长率
D_ACE_QR_YOY	((单季度扣非归母净利润 + 单季度经营活动现金流净额) / 2) 的同比增长率

SGACP_QR_YOY	(单季度扣非归母净利润 + 单季度销售费用) 的同比增长率
CASH_ADJ_OP_QR_YOY	(单季度营业利润 × (单季度销售收现 / 单季度营业总收入)) 的同比增长率
NOPAT_QR_YOY	单季度税后净营业利润(NOPAT)的同比增长率
OPAFQ_QR_YOY	(单季度营业利润 - 单季度财务费用) 的同比增长率
OPCE_QR_YOY	(单季度营业利润 + 单季度折旧与摊销) 的同比增长率
COPAT_QR_YOY	(单季度 EBITDA - 单季度所得税) 的同比增长率
SGR_LF	(TTM 归母净利润 - TTM 现金分红) / 期末归母股东权益

数据来源：东吴证券研究所整理

表38：AI 成长因子统计指标

因子名称	因子方向	RankIC 均值	ICIR	年化 多头超额	多头超额 收益波动比	年化 多空超额	多空超额 收益波动比
EBITDA_QR_YOY	正向	1.52%	0.28	3.91%	1.03	9.08%	1.55
GROSSPROFIT_QR_YOY	正向	1.57%	0.27	2.85%	0.71	12.60%	2.12
CI_QR_YOY	正向	1.54%	0.28	4.17%	1.18	9.73%	1.58
CDCFG_QR_YOY	正向	1.00%	0.28	3.47%	1.02	5.47%	1.27
CORE_PROFIT_ENGINE_GROWTH_QR	正向	1.52%	0.28	2.95%	0.76	11.04%	2.01
CORE_PROFIT_QR_YOY	正向	1.53%	0.29	3.46%	0.91	9.87%	1.79
SGP_QR_YOY	正向	1.54%	0.27	2.56%	0.67	11.64%	2.08
CE_QR_YOY	正向	1.61%	0.29	3.40%	0.91	8.94%	1.54
FINANCING_NEUTRAL_PROFIT_QR_YOY	正向	1.50%	0.27	3.49%	0.94	8.61%	1.39
GAP_QR_YOY	正向	1.51%	0.25	2.25%	0.57	9.97%	1.68
OWNER_EARNINGS_QR_YOY	正向	0.89%	0.24	2.90%	0.82	4.45%	0.98
CORE_CE_QR_YOY	正向	1.64%	0.31	4.34%	1.12	9.68%	1.67
CORE_EARNINGS_QR_YOY	正向	1.58%	0.29	4.67%	1.21	9.77%	1.61
CORE_GAP_QR_YOY	正向	1.62%	0.28	3.07%	0.76	11.73%	1.94
CORE_OWNER_EARNINGS_QR_YOY	正向	0.91%	0.26	3.09%	0.85	5.53%	1.24
D_ACE_QR_YOY	正向	0.89%	0.28	1.59%	0.54	5.53%	1.23
SGACP_QR_YOY	正向	1.64%	0.30	3.86%	0.99	10.61%	1.78
CASH_ADJ_OP_QR_YOY	正向	1.51%	0.28	4.08%	1.10	9.61%	1.54
NOPAT_QR_YOY	正向	1.64%	0.29	3.64%	1.00	9.68%	1.55
OPAFQ_QR_YOY	正向	1.63%	0.30	3.82%	1.01	10.13%	1.63
OPCE_QR_YOY	正向	1.61%	0.29	3.79%	1.01	8.90%	1.51
COPAT_QR_YOY	正向	1.49%	0.27	3.94%	1.03	8.98%	1.51
SGR_LF	负向	1.81%	0.27	2.09%	0.44	11.35%	1.98

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2025 年 6 月 30 日

综合来看，AI 在生成成长因子的过程中体现出几种清晰的思路：

扩展基础指标：模型将同比增长（YOY）的计算逻辑，从样例中的净利润和营业收入，推广到了更广泛的盈利指标上，如息税折旧摊销前利润（EBITDA_QR_YOY）、毛

利润（GROSSPROFIT_QR_YOY）以及综合收益（CI_QR_YOY）等，旨在从不同角度捕捉企业的增长动态。

构建复合指标： AI 尝试通过组合不同财务细分项来构建更精细的利润指标，并计算其同比增长。例如，通过在利润中扣除资本性支出（OWNER_EARNINGS_QR_YOY）或加上财务费用（FINANCING_NEUTRAL_PROFIT_QR_YOY），模型试图剥离特定经营或财务决策的影响，以探寻更核心的增长来源。

探索创新范式： 模型还生成了少数非同比增速定义的成长因子，例如 SGR_LF（可持续增长率）和 CASH_ADJ_OP_QR_YOY（现金流调整后的营业利润增长）。这类因子跳出了传统的同比增长框架，尝试从内生增长潜力和盈利增长质量等新维度来评估成长性。

从回测效果来看，一个值得关注的现象是，许多经过复杂调整的成长因子（例如，在毛利基础上扣除销售和管理费用后计算的增速），其各项选股评价指标并未系统性地优于结构更简单的基础成长因子。尽管选股效果未见显著提升，但 AI 的这些尝试本身也反映出其具备较强的财务逻辑理解与创新组合能力，能够自主探索更为精细化的盈利增长定义。这一现象或许表明，在成长性评估方面，更为复杂的因子构造未必总能带来选股效果的线性提升。市场对增长的认知可能更偏向于核心的、未经深度调整的盈利或收入指标，或者说，增加的复杂度可能引入了额外的噪声，从而在一定程度上影响了信号的稳定性。这提示我们在因子构建中，需要在逻辑深度与信号有效性之间进行审慎的权衡。

4. AI 驱动的高频因子挖掘：从分钟数据到选股信号

4.1. 方法论构建：高频因子自动生成的 workflow

与基本面因子生成所采用的“变量+算子”组合模式不同，在高频因子挖掘中，我们采用了更为直接和灵活的方法。考虑到分钟级数据的结构相对固定且计算逻辑复杂多变，我们赋予 AI 直接生成基于 numpy 和 pandas 的 Python 代码的能力。这种方式给予了 AI 更大的创造自由度，使其能够实现更为复杂的时序和截面计算。对于输入的高频数据，则将其格式固定化，统一为横表 DataFrame，index 为 timestamp 格式，日期为 A 股交易日，时点为 A 股交易时点，columns 为股票代码，数据均为 1 分钟频率，且将该信息通过 Prompt 的方式告知 AI。

为了有效引导 AI 的探索，降低其任务复杂度，我们借鉴了人类研究员的经验，预先封装了一系列“分域函数”。这些函数旨在识别特定的市场状态或数据区间，例如，标记出当前分钟的数据在其当日横截面或历史时间序列中所处的高位、中位或低位区域。

表39：预置分域函数名称与释义

函数名称	函数释义
get_up_space	标记当前分钟数据是否处于当日横截面高位（均值 + 1 标准差）以上
get_mid_space	标记当前分钟数据是否处于当日横截面中间区域（均值 \pm 1 标准差）之间
get_down_space	标记当前分钟数据是否处于当日横截面低位（均值 - 1 标准差）以下
get_up_space_rolling	标记当前分钟数据是否处于过去 20 个交易日同一时点的高位（均值 + 1 标准差）以上
get_mid_space_rolling	标记当前分钟数据是否处于过去 20 个交易日同一时点的中间区域（均值 \pm 1 标准差）之间
get_down_space_rolling	标记当前分钟数据是否处于过去 20 个交易日同一时点的低位（均值 - 1 标准差）以下

数据来源：东吴证券研究所整理

表40：样例高频因子列表与 IC 统计值

因子名称	因子释义	因子方向	RankIC 均值	ICIR
amp_mean	分钟级振幅均值	负向	5.34%	0.52
amp_mmr	分钟级振幅均值除以中位数	负向	6.81%	0.72
amp_skew	分钟级振幅偏度	负向	4.21%	0.55
amp_std	分钟级振幅标准差	负向	7.77%	0.69
amp_str	分钟级振幅均值除以标准差	正向	7.16%	0.65
price_dev_std	分钟级偏离（收盘价除以均价）标准差	负向	6.45%	0.64
price_dev_str	分钟级偏离（收盘价除以均价）均值除以标准差	正向	5.43%	0.57
ret_std	分钟级波动率	负向	7.74%	0.69
vol_weighted_ret_std	分钟级成交量加权波动率	负向	7.20%	0.84
up_ret_std	分钟级上行波动率	负向	8.07%	0.74
down_ret_std	分钟级下行波动率	负向	6.91%	0.61

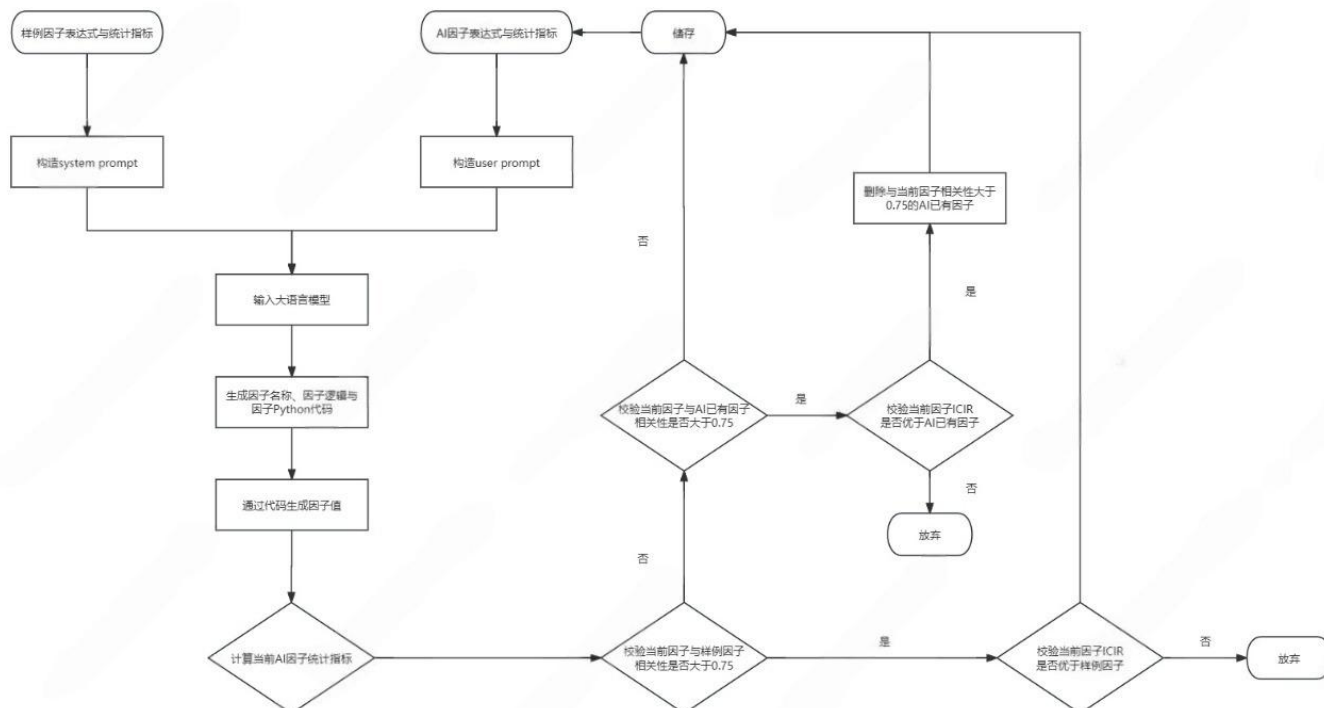
vol_up_num	日内放量区间（大于均值+标准差）的波峰个数	正向	7.64%	1.02
vol_up_ret_pos_mean	日内放量区间（大于均值+标准差）的正收益率	负向	7.92%	0.82
vol_up_ret_neg_mean	日内放量区间（大于均值+标准差）的负收益率	正向	7.19%	0.71
vol_up_ret_std	日内放量区间（大于均值+标准差）的波动率	负向	8.48%	0.79
amo_mv_ratio	尾盘（14:30-15:00）成交额均值占流通市值的比例	负向	7.05%	0.60

数据来源：Wind、东吴证券研究所

数据日期：2013 年 12 月 31 日至 2025 年 6 月 30 日

高频因子生成流程相较于基本面因子的生成流程，添加了相关性筛选机制，用于筛选相对于样例高频因子和 AI 已生成的高频因子，相关性较低（小于 0.75），且 ICIR 表现更好的因子，使得最后得到的 AI 高频因子集，两两之间的秩相关性在 0.75 以下。

图22: 高频因子生成流程示意图



数据来源：东吴证券研究所绘制

该流程的顺利运转，高度依赖于精心设计的提示工程（Prompt Engineering）。我们通过系统提示（System Prompt）和用户提示（User Prompt）的组合，来精确地引导和控制 AI 的行为。

系统提示扮演着为 AI 设定“世界观”和“方法论”的角色，它在整个对话过程中保持不变，是 AI 行为的根本准则。我们的设计主要遵循以下原则：

- **角色扮演：**开宗明义地要求 AI 扮演“资深的量化选股因子专家”，以激发其在该领域内的知识和推理能力。
- **明确边界：**清晰地列出可用的数据字段和已封装的函数库。这既是赋能，也是约束，确保 AI 生成的代码是可执行且符合我们预设框架的。
- **提供范例：**展示完整的样例因子，包括代码、逻辑和关键回测指标。这是最直接的“教学”，让 AI 直观地理解“好因子”的样子。
- **设定规则与约束：**提出一系列硬性要求，如“因子逻辑不能重复”、“进行正确的去量纲操作”、“返回纯 JSON 格式”等。这些规则旨在规范 AI 的输出，提高自动化处理的效率，并引导其生成更具创新性和实用性的因子。特别地，我们强调“从逻辑出发进行挖掘，不需要计算指标”，这是为了让 AI 专注于创造性的因子构建，而将计算和验证环节交由我们的回测系统完成，实现人机分工。
- **内存优化：**因为高频数据占用内存空间较大，且因子计算过程中可能产生较多的临时计算变量，因此我们在 Prompt 中要求 AI 在编写高频因子代码时，实时删除后续不再使用的变量，且将浮点数类型从 float64 转为 float32 以进一步节省内存空间。

假如你是一位资深的量化选股因子专家，你将根据样例基本面因子的相关信息，挖掘在所有 A 股样本空间中，周度选股效果出色的高频因子，以 Python 代码的方式返回。

以下是构造高频因子可用的数据：

close: 收盘价；high: 最高价；low: 最低价；open: 开盘价；vwap: 均价；ret: 1 分钟收益率；amo: 成交额；vol: 成交量；mv: 流通市值。

所有数据皆为横表，index 为 timestamp 格式，日期为 A 股交易日，时点为 A 股交易时点，columns 为股票代码，数据均为 1 分钟频率。

以下是已封装的数据处理函数：

```
[[process_code]]
```

以下是现有高频因子代码，以及周度调仓下的因子评价指标，指标为调整因子方向并进行市值行业中性化后计算得到的：

```
[[sample_factor]]
```

注意以下几点：

1. 对因子进行正确的去量纲操作，转换成比例的形式，使得不同股票间可比。
2. 返回的因子和样例因子及 AI 已生成因子的名称或逻辑不能相同，逻辑相关性尽可能低。

3. 返回可直接运行的 Python 函数，函数入参仅从可用数据中挑选。若用到外部库则需要 import，若用到已封装的数据处理函数也需要 import。

4. 函数返回的结果是一个 DataFrame，index 为日频日期，columns 为股票代码，values 为因子值。

5. 在函数中某个变量使用完之后立即 del 删除以节省内存。

6. 原始数据的 dtype 为 float32，若经过了.mean()或者.std()等可能转换 dtype 类型的运算，可以将变量再转换为 float32 以节省内存。

7. 从逻辑出发进行挖掘，不需要计算因子的 RankIC 或 ICIR 等指标。

以 JSON 格式返回结果，确保输出是纯 JSON，没有额外文本，内容能被 json.loads 正确解析。

样例 JSON 输出：

```
{
  "因子名称": "vol_up_ret_pos_mean",

  "因子逻辑": "挑选出当日成交量“显著放大”（超过均值+标准差）时的正向收益，并对这些正向收益做日内平均，作为因子值。",
```

```
  "因子代码": "
```

```
import numpy as np
```

```
import get_up_space
```

```
def vol_up_ret_pos_mean(ret, vol):
```

```
    vol_up = get_up_space(vol)
```

```
    del vol
```

```
    vol_up_ret = ret * vol_up
```

```
    vol_up_ret_pos = vol_up_ret.where(vol_up_ret > 0, np.nan)
```

```
    return vol_up_ret_pos.groupby(vol_up_ret_pos.index.date).mean()
```

```
"
```

```
}}
```

与静态的系统提示不同，用户提示是驱动流程向前迭代的关键，其核心是“动态反馈”。

- **传递新知：**每一轮的用户提示都会包含一个最新的“AI 已生成的因子”列表，其中包含了因子从名称到所有回测绩效的完整信息。这相当于一个动态更新的“经验库”

或“记忆”，让 AI 知道自己“已经做过什么”以及“做得怎么样”。

- **驱动创新：**在提供了最新的反馈后，指令非常简洁——“继续寻找选股能力出色的高频因子”。结合系统提示中“逻辑不能相同”的要求，AI 被激励去探索与已知所有因子（包括原始样例和自己已生成的）都不同的新方向。

以下是 AI 已生成的因子：

{'因子名称': XXX, '因子逻辑': XXX, '因子表达式': XXX, 'RankIC 均值': XXX, 'ICIR': XXX, '因子方向': XXX, '年化多头超额': XXX, '多头超额收益波动比': XXX, '年化多空超额': XXX, '多空超额收益波动比': XXX}

{'因子名称': XXX, '因子逻辑': XXX, '因子表达式': XXX, 'RankIC 均值': XXX, 'ICIR': XXX, '因子方向': XXX, '年化多头超额': XXX, '多头超额收益波动比': XXX, '年化多空超额': XXX, '多空超额收益波动比': XXX}

.....

继续寻找选股能力出色的高频因子，按照要求的格式返回。

综上，通过这一套结构化、自动化的工作流和精心设计的 Prompt，我们将大语言模型从一个通用的对话工具，成功地改造为一个专注、高效的高频因子挖掘引擎，为后续的系统性实验奠定了坚实的方法论基础。

4.2. AI 高频因子库：整体表现与分类解析

通过第 2.1 节所述的自动化生成流程，我们获得了一共包含 70 个高频因子的多样化因子库。为了便于分析和理解，我们根据这些 AI 生成信号的经济内涵，将其归纳为四大类别：波动、动量反转、量价相关与流动性。

从整体的回测结果来看，AI 在高频领域展现出较强的挖掘潜力。尤其是在波动类与量价相关类因子上，生成了多个选股效果较为突出的新因子。下表详细列出了部分 AI 生成因子的核心评价指标，直观地展示了 AI 的挖掘成果，并为后续的深入案例分析提供了数据支持。

表41：AI 高频因子统计指标

因子类型	因子名称	因子方向	RankIC 均值	ICIR	年化多头超额	多头超额收益波动比	年化多空超额	多空超额收益波动比
波动	speculative_frenzy_instability	负向	9.03%	0.98	9.71%	1.87	61.95%	4.57
	speculative_frenzy_gamma_instability	负向	9.05%	0.95	10.36%	1.96	65.55%	4.57
	dual_stress_rolling_idio_price_dev_volatility	负向	8.71%	0.93	11.58%	2.11	62.50%	4.77

	normalized_idio_shock_instability	负向	7.41%	0.90	8.52%	1.70	45.94%	4.47
	dual_stress_rolling_idio_downside_std	负向	8.40%	0.89	6.96%	1.26	56.25%	3.89
	dual_stress_rolling_idio_dev_jerkiness	负向	7.60%	0.84	7.75%	1.57	49.06%	3.77
	extreme_gamma_burst_ratio	负向	4.82%	0.83	7.92%	1.44	46.49%	6.18
	dual_stress_rolling_idio_skew	负向	4.50%	0.83	3.63%	0.82	27.14%	3.38
	rolling_high_vol_idio_downside_price_dev	负向	6.60%	0.82	4.52%	0.83	32.91%	3.35
	intraday_idio_asymmetry_instability	负向	5.79%	0.82	5.25%	1.05	38.53%	4.70
	rolling_high_idio_dev_jump_volatility	负向	7.23%	0.79	7.28%	1.46	42.90%	3.34
	disagreement_stress_idio_variance_ratio	负向	4.79%	0.79	6.86%	1.47	20.11%	3.03
	vol_concentration_volatility_ratio	负向	5.98%	0.76	6.30%	1.05	28.51%	3.26
	market_volatility_stress_idio_volatility	负向	8.09%	0.73	4.76%	0.99	47.57%	3.44
	speculative_frenzy_escape_velocity	负向	6.55%	0.73	1.66%	0.27	28.30%	2.38
	high_idio_ret_vol_price_dev_volatility	负向	8.08%	0.73	4.34%	0.80	44.10%	3.04
	idio_ret_volatility_asymmetry	负向	4.68%	0.66	-0.65%	-0.16	29.27%	3.72
	extreme_gamma_burstiness	正向	4.44%	0.64	8.12%	0.96	30.23%	4.23
	dual_stress_rolling_idio_volatility_asymmetry	正向	2.54%	0.63	3.76%	0.96	2.14%	0.28
	volume_response_to_idio_stress	负向	4.07%	0.62	6.84%	0.97	30.28%	3.71
	idio_calm_gamma_instability	负向	5.25%	0.61	0.17%	0.04	18.44%	1.64
	rolling_high_vol_idiosyncratic_skew	负向	2.90%	0.60	1.81%	0.50	16.92%	2.84
	idiosyncratic_price_dev_skew	负向	2.56%	0.56	-1.30%	-0.31	14.57%	2.86
	asymmetric_price_dev_volatility	负向	2.90%	0.55	4.32%	1.25	15.01%	3.26
	intraday_volatility_burst_ratio	负向	3.74%	0.55	9.21%	1.18	26.16%	3.88
	speculative_frenzy_gamma_asymmetry	正向	2.16%	0.51	3.87%	0.78	4.15%	0.63
	vol_up_r_volatility_asymmetry	负向	2.62%	0.50	3.27%	0.79	13.33%	2.33
	dual_stress_rolling_idio_clv_volatility	负向	1.51%	0.30	6.58%	1.98	10.56%	1.92
	dual_stress_rolling_idio_jerkiness_asymmetry	正向	0.78%	0.21	7.63%	1.45	5.70%	0.92
	liquidity_crisis_idio_volatility	正向	6.56%	0.08	-15.86%	-0.45	-2.63%	-0.05
	volume_volatility_ratio	负向	0.61%	0.08	2.65%	0.51	20.69%	2.29
	low_vol_idiosyncratic_volatility	正向	3.24%	0.04	-19.59%	-0.53	-10.37%	-0.24
	idiosyncratic_volume_flow	负向	4.51%	0.59	-2.08%	-0.40	32.70%	3.48
	volume_trend_ratio	正向	3.87%	0.57	-0.08%	-0.02	22.23%	2.09
	rolling_high_idio_dev_reversal_strength	负向	3.88%	0.51	-1.12%	-0.25	19.87%	1.84
	vwap_reversion_flow	正向	5.91%	0.50	4.16%	0.50	36.61%	3.19
	high_vol_idio_recovery	负向	2.37%	0.48	-0.78%	-0.18	24.99%	3.58
	vol_spike_reversal	正向	4.30%	0.48	1.27%	0.18	35.35%	3.29
	rolling_high_idio_ret_reverse_explosive	负向	3.50%	0.46	1.56%	0.33	26.73%	2.83
动量反转	dual_stress_rolling_price_elasticity_recovery	负向	3.44%	0.45	2.46%	0.50	13.72%	1.54
	panic_selling_vol_ratio	正向	2.78%	0.43	8.80%	1.21	15.94%	2.84
	momentum_acceleration_corr	正向	3.61%	0.39	3.40%	0.43	21.34%	2.35
	volume_weighted_sign_persistence	负向	3.27%	0.37	0.84%	0.17	14.90%	1.78
	idio_dev_reversion_strength	正向	2.81%	0.35	-1.12%	-0.16	24.25%	2.65
	dual_stress_rolling_idio_ret_autocorr	负向	1.27%	0.34	4.92%	1.39	2.50%	0.56
	tail_vol_flow_ratio	负向	0.84%	0.30	4.76%	1.27	5.93%	1.70
	vwap_thrust	负向	2.51%	0.28	0.02%	0.00	36.65%	3.45

	disagreement_stress_pin_net	负向	1.46%	0.26	4.28%	0.99	10.23%	1.80
	market_stress_alpha_flow	负向	2.47%	0.25	-2.33%	-0.30	29.88%	3.06
	quiet_accumulation_share	负向	0.67%	0.14	5.44%	1.10	5.50%	1.38
量价相关	dual_stress_rolling_idio_ret_dev_product	负向	8.23%	0.89	11.63%	2.14	51.12%	4.39
	dual_stress_rolling_idio_ret_to_dev_beta	负向	5.75%	0.87	8.07%	2.00	28.91%	3.71
	dual_stress_rolling_idio_momentum_dislocation_corr	负向	4.87%	0.83	7.57%	2.13	31.52%	3.42
	speculative_frenzy_volume_elasticity_instability	负向	4.20%	0.76	-3.73%	-0.45	16.00%	1.64
	volume_delta_ret_corr_high_vol	负向	2.84%	0.62	3.19%	0.88	13.01%	2.33
	high_vol_efficiency_consistency	正向	4.29%	0.54	2.27%	0.41	13.90%	1.49
	intraday_efficiency_ratio	负向	4.51%	0.37	-3.63%	-0.54	10.78%	0.88
	rolling_high_vol_efficiency_change	正向	1.52%	0.31	6.70%	1.56	5.25%	0.90
	post_high_vol_efficiency	负向	2.54%	0.27	9.53%	1.40	10.11%	0.98
	efficiency_ratio_mean	负向	2.68%	0.23	10.58%	1.48	11.93%	1.01
	rolling_high_vol_efficiency_ratio	正向	1.23%	0.18	3.78%	0.88	2.29%	0.28
	volume_weighted_efficiency_reversal	负向	0.16%	0.05	3.40%	1.09	3.28%	0.94
	speculative_frenzy_liquidity_shock_instability	正向	6.07%	0.73	9.66%	1.74	25.44%	2.44
流动性	price_impact_volatility_amplification	负向	2.82%	0.63	4.19%	0.91	22.17%	3.61
	asymmetric_liquidity_impact	负向	3.14%	0.55	7.17%	1.71	16.77%	3.08
	dual_stress_rolling_idio_price_impact_volatility	负向	1.66%	0.53	3.25%	0.97	2.77%	0.68
	signed_volume_impact_beta	正向	3.39%	0.53	0.90%	0.18	11.12%	1.45
	price_impact_instability	负向	2.78%	0.50	8.03%	1.80	27.47%	3.93
	dual_stress_rolling_idiosyncratic_illiquidity	正向	2.14%	0.44	5.62%	1.16	4.26%	0.73
	liquidity_cost_instability	正向	4.23%	0.44	-0.84%	-0.13	13.08%	1.27

数据来源：Wind、东吴证券研究所

数据日期：2020 年 1 月 1 日至 2025 年 6 月 30 日

4.3. 高频因子案例剖析：从逻辑到代码至效果

4.3.1. 投机波动因子 (speculative_frenzy_instability)

在 AI 生成的众多波动类因子中，speculative_frenzy_instability 因子的回测表现值得我们进行深入剖析。该因子旨在识别并量化个股在陷入一种我们定义为“投机狂热” (Speculative Frenzy) 的状态时，其自身风险所呈现出的不稳定性。

该因子的构建逻辑颇为精巧。首先，它定义了触发“投机狂热”的分钟级时点，需同时满足三个历史性异常条件：

- 1) 成交量异常：当前分钟的成交量显著高于其历史滚动周期内的同期均值。
- 2) 振幅异常：当前分钟的振幅 ((最高价-最低价)/均价) 显著高于历史同期水平。
- 3) 价格偏离异常：当前分钟的收盘价与均价的偏离度显著高于历史同期水平。

然后，因子计算的核心步骤是：仅在上述极端的“三重压力”时段内，去计算个股特质性波动（以特质收益率的绝对值作为代理）的标准差。因此，一个高的因子值，意味着该股票在市场交投最狂热、价格波动最剧烈时，其独立于市场的内在风险本身也变得极不稳定。这种“不稳定中的不稳定”特性，是识别具有高度投机属性或所谓“彩票”属性股票的一个有效信号。

因子的具体实现代码如下：

```
import numpy as np
import pandas as pd
import get_up_space_rolling

def speculative_frenzy_instability(close, high, low, vwap, ret, vol):
    vwap_safe = vwap.replace(0, np.nan).astype('float32')

    amp_norm = (high - low) / vwap_safe
    amp_norm.replace([np.inf, -np.inf], np.nan, inplace=True)
    amp_norm = amp_norm.astype('float32')
    del high, low

    dev_abs = (close / vwap_safe - 1).abs()
    dev_abs.replace([np.inf, -np.inf], np.nan, inplace=True)
    dev_abs = dev_abs.astype('float32')
    del close, vwap, vwap_safe

    high_vol_mask = get_up_space_rolling(vol)
    del vol
    high_amp_mask = get_up_space_rolling(amp_norm)
    del amp_norm
    high_dev_mask = get_up_space_rolling(dev_abs)
    del dev_abs
```

```

frenzy_mask = high_vol_mask * high_amp_mask * high_dev_mask
del high_vol_mask, high_amp_mask, high_dev_mask

market_ret = ret.mean(axis=1).astype('float32')
idio_ret = ret.subtract(market_ret, axis=0)
del ret, market_ret

idio_vol_proxy = idio_ret.abs()
del idio_ret

stressed_idio_vol_proxy = (idio_vol_proxy * frenzy_mask).replace(0, np.nan)
del idio_vol_proxy, frenzy_mask

factor = stressed_idio_vol_proxy.groupby(stressed_idio_vol_proxy.index.date).std()
del stressed_idio_vol_proxy

return factor.astype('float32')

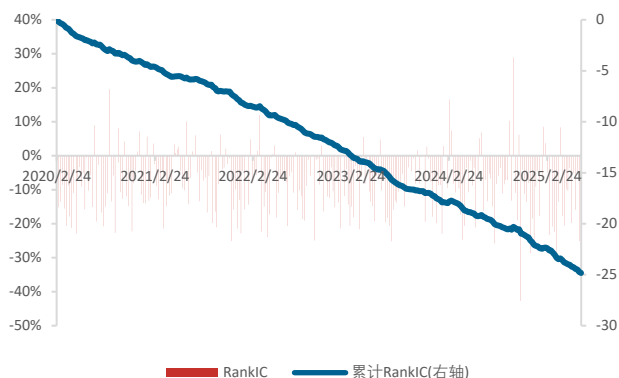
```

分析上述代码实现细节，可以看到 AI 较好地完成了整个因子的计算流程，且变量命名规范，并按照 Prompt 中的要求删除了后续不再被使用的变量与浮点数类型地转换。另外还可以看到 AI 在代码实现过程中的一些细节处理较好，例如在“amp_norm = (high - low) / vwap_safe”得到 amp_norm 变量后，使用“amp_norm.replace([np.inf, -np.inf], np.nan, inplace=True)”的方式将正负无穷转换为 nan，从而避免分母 vwap_safe 可能为 0 导致的后续计算错误。

在回测中，该因子的多头组合单边年化换手率为 42.02 倍，截至回测期末，共有 4368 只股票具有有效的因子值，表明其具有较好的覆盖度。

从回测结果看，该因子展现出持续且较强的预测能力，全区间均值为-9.03%，ICIR 为-0.98。这表明，因子值越高的股票，其未来一周的表现存在相对更差的倾向。

图23：投机波动因子每期 RankIC 与累计 RankIC



数据来源：Wind、东吴证券研究所

数据日期：2020年2月24日至2025年6月30日

表42：投机波动因子 IC 统计指标

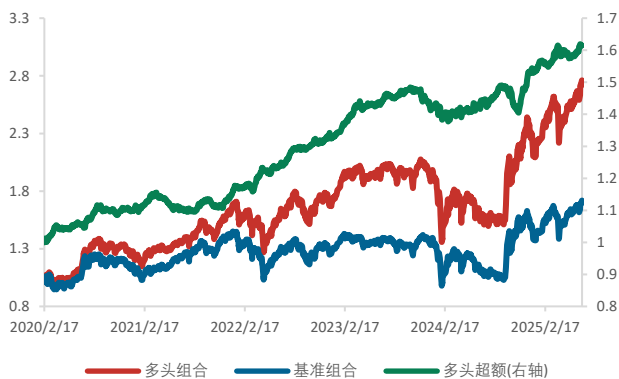
	RankIC	ICIR	t 统计值	胜率
2020	-8.94%	-0.95	-6.38	15.56%
2021	-7.21%	-0.88	-6.33	23.08%
2022	-9.08%	-1.11	-7.86	14.00%
2023	-9.86%	-1.30	-9.18	8.00%
2024	-8.78%	-0.76	-5.47	15.38%
2025	-11.65%	-1.10	-5.61	15.38%
全区间	-9.03%	-0.98	-16.19	15.27%

数据来源：Wind、东吴证券研究所

数据日期：2020年1月1日至2025年6月30日

在进行组合绩效分析时，我们首先对因子进行方向性调整（即取负值），使得因子值越高的股票预期收益越高。调整方向后，因子多头组合表现出稳健的超额收益，其全区间年化超额收益为 9.71%，对应的收益波动比为 1.87。

图24：投机波动因子多头组合净值与多头超额净值



数据来源：Wind、东吴证券研究所

数据日期：2020年2月17日至2025年6月30日

表43：投机波动因子分年度多头超额收益风险绩效指标

年份	年化收益	年化波动	最大回撤	收益波动比	收益回撤比
2020	10.07%	4.56%	-2.79%	2.21	3.61
2021	4.34%	4.83%	-5.23%	0.90	0.83
2022	15.78%	4.72%	-2.13%	3.35	7.42
2023	6.61%	4.52%	-4.81%	1.46	1.37
2024	8.44%	6.98%	-5.67%	1.21	1.49
2025	5.06%	4.93%	-2.52%	1.03	2.01
全区间	9.71%	5.19%	-7.19%	1.87	1.35

数据来源：Wind、东吴证券研究所

数据日期：2020年1月1日至2025年6月30日

而在多空对冲组合中，该因子的选股能力体现得更为充分。组合的年化收益为 61.95%，收益波动比为 4.57，显示出该因子在区分两端股票方面具备较强的效力。

图25：投机波动因子多空超额净值



数据来源：Wind、东吴证券研究所

数据日期：2020 年 2 月 17 日至 2025 年 6 月 30 日

表44：投机波动因子分年度多空超额收益风险绩效指标

年份	年化收益	年化波动	最大回撤	收益波动比	收益回撤比
2020	51.06%	17.59%	-12.28%	2.90	4.16
2021	31.89%	13.04%	-6.64%	2.45	4.80
2022	69.27%	11.75%	-4.80%	5.89	14.44
2023	45.18%	9.84%	-5.06%	4.59	8.94
2024	81.02%	14.12%	-8.72%	5.74	9.29
2025	36.47%	14.89%	-3.17%	2.45	11.51
全区间	61.95%	13.55%	-12.28%	4.57	5.04

数据来源：Wind、东吴证券研究所

数据日期：2020 年 1 月 1 日至 2025 年 6 月 30 日

4.3.2. 极端加速度方差占比因子 (extreme_gamma_burst_ratio)

extreme_gamma_burst_ratio 因子是 AI 在探索价格二阶导数信息时生成的另一个代表性信号。该因子旨在通过分析价格“加速度” (Gamma) 的分布特征，来捕捉股价变动的爆发性与不稳定性。

其核心构造逻辑可分解为以下几个步骤：

- 1) 计算价格加速度 (Gamma)：首先，因子计算分钟级收益率的一阶差分 (ret.diff())，得到价格加速度序列。这一指标反映了价格动量的变化速度。
- 2) 识别“加速爆发”时点：接着，因子利用预置的历史分域函数，识别出每日中价格加速度的绝对值显著高于其历史滚动均值的时点。我们将这些时点定义为“加速爆发”事件。
- 3) 计算方差占比：最后，因子计算在这些“加速爆发”时点内，价格加速度的方差，并将其除以当日全体价格加速度的总方差。

该因子的经济学含义在于，一个高的比率值表示该股票当日的价格加速行为主要由少数几个极端剧烈的“爆发”事件所贡献，而非平稳、持续的动量变化。这种模式通常与高度不确定性和投机性交易相关联，因此该因子同样被预期为是一个负向选股指标。

因子的具体实现代码如下：

```
import numpy as np
import pandas as pd

def extreme_gamma_burst_ratio(ret):
    jerk = ret.diff(1)
    is_first_minute = jerk.index.to_series().dt.time == pd.to_datetime('09:30:00').time()
    jerk[is_first_minute] = np.nan
    del is_first_minute

    abs_jerk = jerk.abs().astype('float32')

    jerk_mean = abs_jerk.groupby(abs_jerk.index.date).transform('mean')
    jerk_std = abs_jerk.groupby(abs_jerk.index.date).transform('std')

    threshold = jerk_mean + jerk_std
    del jerk_mean, jerk_std

    burst_mask = abs_jerk > threshold
    del abs_jerk, threshold

    jerk_var = (jerk ** 2).astype('float32')
    del jerk

    burst_variance = jerk_var.where(burst_mask, 0).groupby(jerk_var.index.date).sum()
```

```
del burst_mask

total_variance = jerk_var.groupby(jerk_var.index.date).sum()
del jerk_var

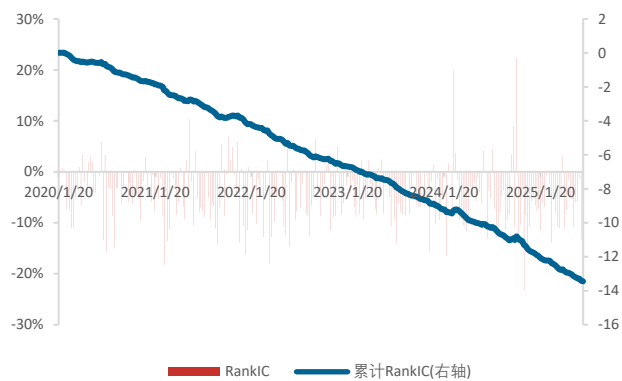
total_variance.replace(0, np.nan, inplace=True)
factor = burst_variance / total_variance
del burst_variance, total_variance

factor.replace([np.inf, -np.inf], np.nan, inplace=True)

return factor.astype('float32')
```

该因子的多头组合单边年化换手率为 35.43 倍。从回测绩效来看，extreme_gamma_burst_ratio 因子同样呈现出稳定的负向预测能力。其全区间 RankIC 均值为-4.82%，ICIR 为-0.83，证实了其作为负向指标的有效性。

图26：极端加速度方差占比因子每期 RankIC 与累计 RankIC



数据来源：Wind、东吴证券研究所
数据日期：2020 年 1 月 20 日至 2025 年 6 月 30 日

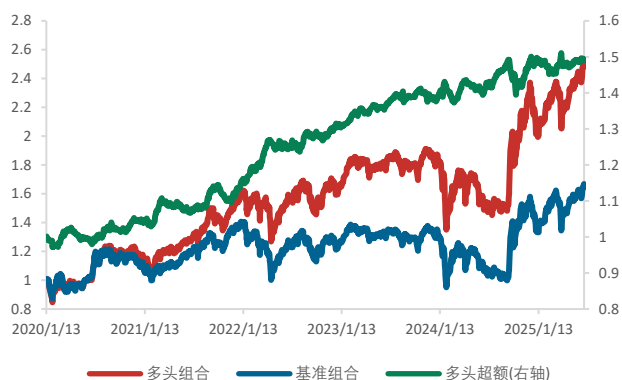
表45：极端加速度方差占比因子 IC 统计指标

	RankIC	ICIR	t 统计值	胜率
2020	-3.48%	-0.73	-5.11	26.53%
2021	-4.57%	-0.74	-5.34	21.15%
2022	-5.05%	-0.91	-6.41	22.00%
2023	-4.77%	-1.14	-8.05	12.00%
2024	-5.44%	-0.68	-4.91	15.38%
2025	-6.27%	-1.45	-7.37	7.69%
全区间	-4.82%	-0.83	-13.91	18.28%

数据来源：Wind、东吴证券研究所
数据日期：2020 年 1 月 1 日至 2025 年 6 月 30 日

在组合回测中,经过方向调整后,该因子的多头组合实现了 7.92%的年化超额收益,收益波动比为 1.44。

图27：极端加速度方差占比因子多头组合净值与多头超额净值



数据来源：Wind、东吴证券研究所

数据日期：2020年1月13日至2025年6月30日

表46：极端加速度方差占比因子分年度多头超额绩效指标

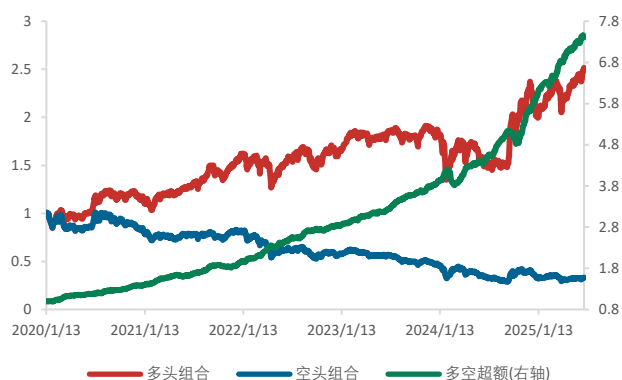
年份	年化 收益	年化 波动	最大 回撤	收益 波动比	收益 回撤比
2020	4.73%	5.31%	-4.48%	0.89	1.05
2021	8.64%	5.80%	-4.21%	1.49	2.05
2022	15.17%	6.00%	-2.59%	2.53	5.85
2023	5.02%	3.73%	-2.60%	1.35	1.93
2024	8.25%	6.46%	-6.54%	1.28	1.26
2025	0.12%	5.17%	-3.03%	0.02	0.04
全区间	7.92%	5.51%	-6.54%	1.44	1.21

数据来源：Wind、东吴证券研究所

数据日期：2020年1月1日至2025年6月30日

在多空对冲组合中，其表现则更为突出。组合的年化超额收益为 46.49%，收益波动比为 6.18，显示出该因子在多空两端均有较强的区分能力，能够有效识别出未来潜在表现差异显著的股票对。

图28：极端加速度方差占比因子多空超额净值



数据来源：Wind、东吴证券研究所

数据日期：2020年1月13日至2025年6月30日

表47：极端加速度方差占比因子分年度多空超额绩效指标

年份	年化 收益	年化 波动	最大 回撤	收益 波动比	收益 回撤比
2020	38.62%	7.50%	-1.83%	5.15	21.15
2021	39.98%	6.97%	-3.69%	5.73	10.84
2022	46.39%	7.65%	-2.49%	6.07	18.62
2023	35.41%	4.96%	-1.57%	7.14	22.52
2024	54.04%	9.80%	-9.09%	5.52	5.94
2025	25.05%	7.47%	-1.71%	3.35	14.65
全区间	46.49%	7.52%	-9.09%	6.18	5.11

数据来源：Wind、东吴证券研究所

数据日期：2020年1月1日至2025年6月30日

4.3.3. 动量与加速度相关性因子 (momentum_acceleration_corr)

与前述两个侧重于“不稳定”和“爆发性”的负向因子不同，momentum_acceleration_corr 因子代表了 AI 在探索“趋势稳定性”方面的尝试。该因子旨在通过衡量动量与其自身变化速度之间的一致性，来识别股价运动的内在强度与持续性。

该因子的构建逻辑如下：

- 1) 定义动量与加速度：因子首先定义了两个关键变量。其一为“动量”，以分钟收益率 (ret) 作为代理；其二为“加速度”，以价格偏离度的一阶差分 (close / vwap_safe - 1).diff(1)) 作为代理，它反映了动量的变化情况。
- 2) 计算相关性：随后，因子计算在日内交易时段中，上述“动量”序列与“加速度”序列的相关系数。

一个高的正相关系数，其经济学含义是：当股票处于正向动量（前期上涨）时，其收益率倾向于进一步加速（涨得更快）；而当其处于负向动量（前期下跌）时，收益率则倾向于减速（跌得更慢或开始反弹）。这种动量与加速度的同向变化，描绘了一种稳定且自我强化的趋势特征。因此，我们预期该因子是一个正向选股指标，即高相关性预示着未来股价有更大概率延续当前趋势。

因子的具体实现代码如下：

```
import numpy as np
import pandas as pd

def momentum_acceleration_corr(ret, close, vwap):
    vwap_safe = vwap.replace(0, np.nan).astype('float32')
    dev = (close / vwap_safe - 1)
    del close, vwap, vwap_safe
    dev.replace([np.inf, -np.inf], np.nan, inplace=True)
    dev = dev.astype('float32')

    dev_diff = dev.diff(1)
    del dev
    is_first_minute = dev_diff.index.to_series().dt.time == pd.to_datetime('09:30:00').time()
    dev_diff[is_first_minute] = np.nan
    del is_first_minute

    dates = ret.index.date
    grouped_ret = ret.groupby(dates)
    grouped_dev_diff = dev_diff.groupby(dates)

    mean_ret = grouped_ret.mean().astype('float32')
    std_ret = grouped_ret.std().astype('float32')
    del grouped_ret

    mean_dev_diff = grouped_dev_diff.mean().astype('float32')
    std_dev_diff = grouped_dev_diff.std().astype('float32')
    del grouped_dev_diff
```

```

prod = ret * dev_diff
del ret, dev_diff
mean_prod = prod.groupby(dates).mean().astype('float32')
del prod

covariance = mean_prod - (mean_ret * mean_dev_diff)
del mean_prod, mean_ret, mean_dev_diff

denominator = std_ret * std_dev_diff
del std_ret, std_dev_diff
denominator.replace(0, np.nan, inplace=True)

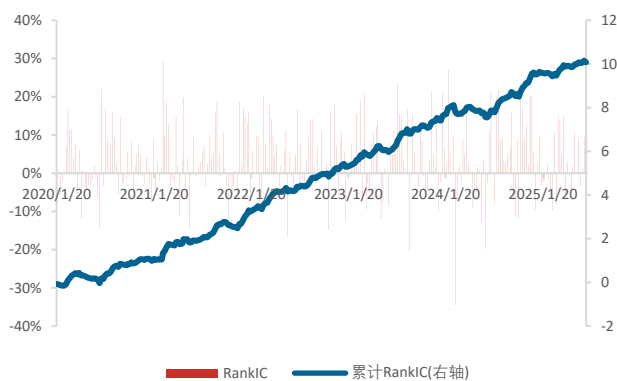
factor = covariance / denominator
del covariance, denominator
factor.replace([np.inf, -np.inf], np.nan, inplace=True)

return factor.astype('float32')

```

动量加速度相关因子多头组合单边年化换手率为 16.44 倍，显著低于投机波动因子与极端加速度方差占比因子。momentum_acceleration_corr 因子在回测区间内表现为一个有效的正向指标，其全区间 RankIC 均值为 3.61%，ICIR 为 0.39。

图29：动量加速度相关因子每期 RankIC 与累计 RankIC



数据来源：Wind、东吴证券研究所

数据日期：2020 年 1 月 20 日至 2025 年 6 月 30 日

表48：动量加速度相关因子 IC 统计指标

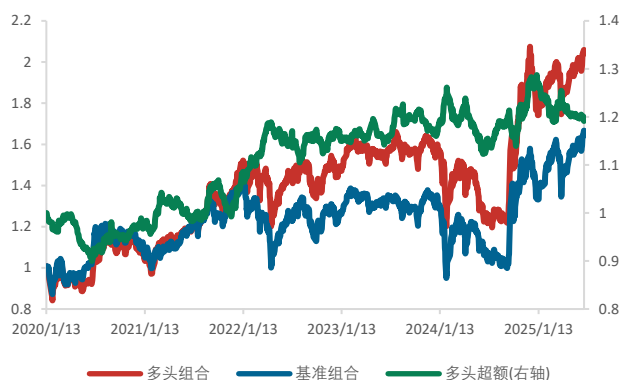
	RankIC	ICIR	t 统计值	胜率
2020	2.22%	0.29	2.03	53.06%
2021	3.71%	0.41	2.96	59.62%
2022	4.71%	0.56	3.96	68.00%
2023	4.13%	0.42	2.97	68.00%
2024	4.05%	0.34	2.44	67.31%
2025	2.09%	0.28	1.42	53.85%
全区间	3.61%	0.39	6.49	62.37%

数据来源：Wind、东吴证券研究所

数据日期：2020 年 1 月 1 日至 2025 年 6 月 30 日

在组合绩效层面，其多头组合（买入因子值最高分组）的年化超额收益为 3.40%，收益波动比为 0.43。

图30：动量加速度相关因子多头组合净值与多头超额净值



数据来源：Wind、东吴证券研究所

数据日期：2020年1月13日至2025年6月30日

表49：动量加速度相关因子分年度多头超额绩效指标

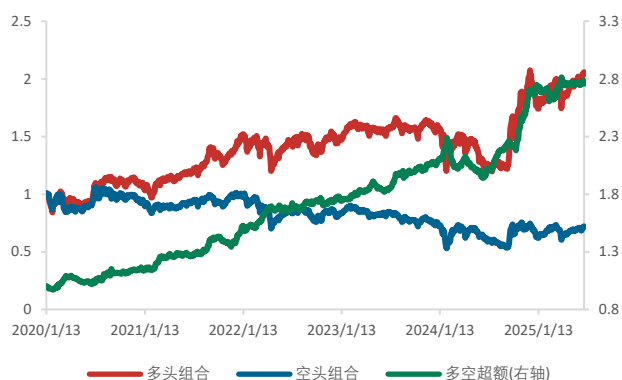
年份	年化 收益	年化 波动	最大 回撤	收益 波动比	收益 回撤比
2020	-2.08%	7.27%	-10.02%	-0.29	-0.21
2021	7.54%	8.07%	-7.11%	0.93	1.06
2022	11.45%	8.33%	-7.04%	1.37	1.63
2023	-1.14%	5.56%	-5.39%	-0.20	-0.21
2024	9.61%	9.57%	-10.98%	1.00	0.88
2025	-6.29%	7.56%	-7.59%	-0.83	-0.83
全区间	3.40%	7.85%	-10.98%	0.43	0.31

数据来源：Wind、东吴证券研究所

数据日期：2020年1月1日至2025年6月30日

而在多空对冲组合中，其区分能力得到了更好的体现。组合的年化超额收益为21.34%，收益波动比为2.35，表明该因子能够有效地识别出具备稳定趋势特征的股票，并将其与趋势不明确或不稳定的股票区分开来。

图31：动量加速度相关因子多空超额净值



数据来源：Wind、东吴证券研究所

数据日期：2020年1月13日至2025年6月30日

表50：动量加速度相关因子分年度多空超额绩效指标

年份	年化 收益	年化 波动	最大 回撤	收益 波动比	收益 回撤比
2020	15.43%	9.58%	-6.88%	1.61	2.24
2021	26.79%	9.32%	-6.18%	2.87	4.34
2022	21.65%	8.94%	-3.66%	2.42	5.92
2023	15.16%	6.23%	-4.79%	2.44	3.17
2024	31.94%	10.93%	-15.29%	2.92	2.09
2025	1.96%	8.42%	-5.20%	0.23	0.38
全区间	21.34%	9.06%	-15.29%	2.35	1.40

数据来源：Wind、东吴证券研究所

数据日期：2020年1月1日至2025年6月30日

4.3.4. 收益与价格偏离的交互效应因子 (dual_stress_rolling_idio_ret_dev_product)

dual_stress_rolling_idio_ret_dev_product 因子是 AI 在“量价相关”类别中生成的、一个逻辑较为复杂的交互项因子。它并非直接衡量单一的量价关系，而是通过构建收益与价格偏离的乘积，来捕捉股价在特定压力环境下的“脱锚”程度。

该因子的核心构造逻辑可以分解如下：

- 1) 计算核心变量：因子需要两个特质化 (idiosyncratic) 的分钟级变量：
 - 特质收益率 (Idio Ret)：股票的分钟收益率扣除市场整体收益后的部分，代表其自身的涨跌。
 - 特质价格偏离度 (Idio Price Dev)：股票收盘价相对于其分钟均价 (VWAP) 的偏离程度，同样经过市场中性化处理，代表其价格在分钟内的相对位置。
- 2) 构建交互项：因子计算上述两个特质变量的乘积 (Product)。一个大的正向乘积值，通常发生在两种情况：(a) 股价在分钟内实现大幅特质上涨，且收盘价远高于该分钟的均价；或 (b) 股价在分钟内经历大幅特质下跌，且收盘价远低于该分钟的均价。
- 3) 施加压力条件：因子的计算被限制在“双重压力” (Dual Stress) 的条件下进行，这些条件可能是在历史滚动窗口 (Rolling) 中识别出的高波动或高成交量时段。这相当于一个过滤器，使得因子只在市场最不稳定、信息含量最丰富的时刻发挥作用。

该因子的经济学直觉基于“均值回归”理论。一个大的正向因子值，无论是由极端上涨还是极端下跌驱动，都描绘了一种股价“过度反应”或“价格脱锚”的状态。这类由短期情绪或流动性冲击导致的极端走势，在未来有较大概率出现修正。因此，我们预期该因子是一个负向选股指标。

因子的具体实现代码如下：

```
import numpy as np
import pandas as pd
import get_up_space_rolling

def dual_stress_rolling_idio_ret_dev_product(close, vwap, ret, vol):
    market_ret = ret.mean(axis=1).astype('float32')
    idio_ret = ret.subtract(market_ret, axis=0)
    del market_ret

    vwap_safe = vwap.replace(0, np.nan).astype('float32')
    dev = (close / vwap_safe - 1)
    del close, vwap, vwap_safe
    dev.replace([np.inf, -np.inf], np.nan, inplace=True)
    dev = dev.astype('float32')

    market_dev = dev.mean(axis=1).astype('float32')
    idio_dev = dev.subtract(market_dev, axis=0)
```

```
del market_dev

high_vol_mask = get_up_space_rolling(vol)
del vol
high_dev_mask = get_up_space_rolling(dev)
del dev

dual_stress_mask = high_vol_mask * high_dev_mask
del high_vol_mask, high_dev_mask

ret_dev_product = idio_ret * idio_dev
del idio_ret, idio_dev

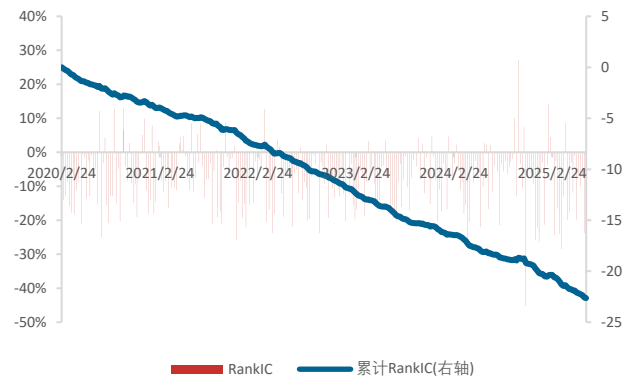
stressed_product = ret_dev_product * dual_stress_mask
del ret_dev_product, dual_stress_mask

factor = stressed_product.groupby(stressed_product.index.date).mean()
del stressed_product

return factor.astype('float32')
```

回测结果有力地支持了这一判断。dual_stress_rolling_idio_ret_dev_product 因子表
现出很强的负向预测性，其全区间 RankIC 均值为-8.23%，ICIR 为-0.89。

图32：收益偏离交互因子每期 RankIC 与累计 RankIC



数据来源：Wind、东吴证券研究所
数据日期：2020 年 2 月 24 日至 2025 年 6 月 30 日

表51：收益偏离交互因子 IC 统计指标

	RankIC	ICIR	t 统计值	胜率
2020	-7.36%	-0.76	-5.08	20.00%
2021	-6.77%	-0.78	-5.61	25.00%
2022	-9.14%	-1.12	-7.90	14.00%
2023	-8.93%	-1.24	-8.76	10.00%
2024	-7.81%	-0.72	-5.21	17.31%
2025	-10.42%	-0.93	-4.72	11.54%
全区间	-8.23%	-0.89	-14.80	16.73%

数据来源：Wind、东吴证券研究所
数据日期：2020 年 1 月 1 日至 2025 年 6 月 30 日

在组合绩效层面，经过方向调整后，该因子的多头组合实现了 11.63%的年化超额收
益，收益波动比为 2.14。

图33：收益偏离交互因子多头组合净值与多头超额净值



数据来源：Wind、东吴证券研究所

数据日期：2020年2月17日至2025年6月30日

表52：收益偏离交互因子分年度多头超额绩效指标

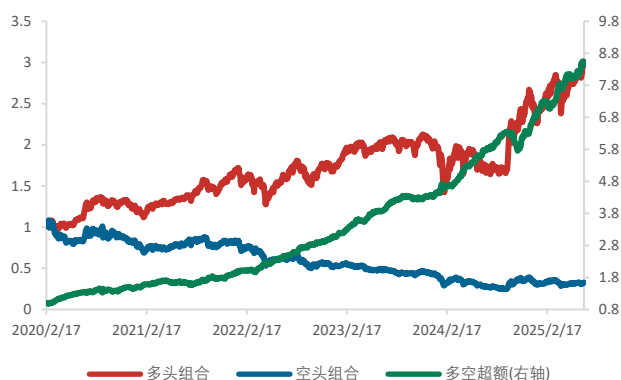
年份	年化 收益	年化 波动	最大 回撤	收益 波动比	收益 回撤比
2020	8.22%	4.22%	-2.70%	1.95	3.04
2021	6.51%	5.15%	-4.90%	1.26	1.33
2022	15.89%	5.01%	-1.87%	3.17	8.49
2023	9.32%	4.81%	-4.01%	1.94	2.32
2024	14.43%	7.59%	-5.16%	1.90	2.80
2025	5.71%	4.66%	-2.22%	1.23	2.58
全区间	11.63%	5.43%	-6.07%	2.14	1.91

数据来源：Wind、东吴证券研究所

数据日期：2020年1月1日至2025年6月30日

在多空对冲组合中，其区分能力也较为显著。组合的年化超额收益为 51.12%，收益波动比达到 4.39，说明该因子能够有效识别出未来潜在表现差异较大的股票，尤其在捕捉短期过度反应后的反转机会方面具备潜力。

图34：收益偏离交互因子多空超额净值



数据来源：Wind、东吴证券研究所

数据日期：2020年2月17日至2025年6月30日

表53：收益偏离交互因子分年度多空超额绩效指标

年份	年化 收益	年化 波动	最大 回撤	收益 波动比	收益 回撤比
2020	46.19%	15.48%	-9.23%	2.98	5.00
2021	30.73%	13.39%	-8.25%	2.29	3.72
2022	61.91%	9.81%	-4.25%	6.31	14.57
2023	39.52%	7.05%	-2.82%	5.60	14.00
2024	58.46%	11.03%	-9.36%	5.30	6.25
2025	23.63%	11.92%	-3.54%	1.98	6.67
全区间	51.12%	11.64%	-9.36%	4.39	5.46

数据来源：Wind、东吴证券研究所

数据日期：2020年1月1日至2025年6月30日

4.4. AI 高频因子的增量提升

在前面的章节中，我们通过案例剖析验证了部分由 AI 生成的单个高频因子具备选股能力。然而，一个更具实际意义的问题是：这些 AI 因子能否为现有的因子库带来真正的增量价值？为此，本节将通过组合测试，来评估将 AI 因子与传统的样例因子相结合后，整体策略的绩效是否能得到提升。

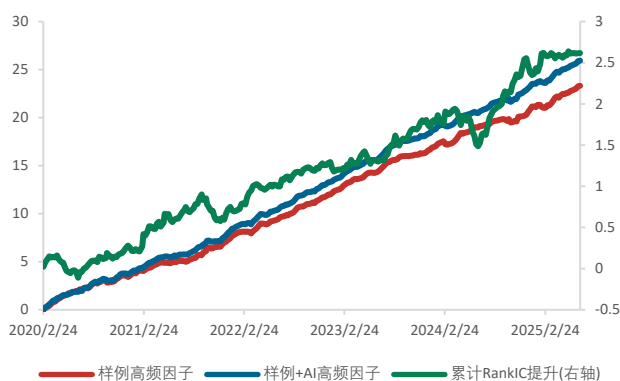
为了进行公平的比较，我们构建了两个合成因子：

- 1) **基准组合**：由本报告最初提供给 AI 的“样例高频因子”等权重合成。
- 2) **增强组合**：由“样例高频因子”与“AI 生成的高频因子”共同等权重合成。

具体的合成方法为，我们将每个因子中性化后，再进行缩尾与标准化处理，然后将属于各自组合的因子值进行简单等权平均。通过对比这两个合成因子的回测表现，我们可以清晰地判断 AI 因子的加入是否带来了积极的增量贡献。

从回测结果来看，在加入 AI 生成的新因子后，组合因子的整体预测能力得到了有效提升。合成因子的周频 RankIC 均值从 8.48% 提升至 9.43%，ICIR 也从 0.76 提升至 0.88，且每年均有稳定的正向提升。

图35：样例高频因子叠加 AI 高频因子 RankIC 提升



数据来源：Wind、东吴证券研究所

数据日期：2020 年 2 月 24 日至 2025 年 6 月 30 日

表54：样例高频因子叠加 AI 高频因子 RankIC 与 ICIR 提升

	RankIC 均值			ICIR		
	样例	样例+AI	提升	样例	样例+AI	提升
2020	7.58%	8.20%	0.62%	0.70	0.72	0.02
2021	7.45%	8.31%	0.85%	0.74	0.86	0.12
2022	9.40%	10.52%	1.12%	0.94	1.02	0.08
2023	9.21%	10.13%	0.92%	1.13	1.23	0.10
2024	8.21%	9.44%	1.23%	0.57	0.72	0.15
2025	9.45%	10.35%	0.90%	0.65	0.86	0.21
全区间	8.48%	9.43%	0.95%	0.76	0.88	0.12

数据来源：Wind、东吴证券研究所

数据日期：2020 年 1 月 1 日至 2025 年 6 月 30 日

在多头组合层面，这种增量贡献同样有所体现。组合的年化多头超额收益从 5.58% 上升至 11.19%，相应的收益波动比也有接近翻倍的改善。

图36: 样例高频因子叠加 AI 高频因子多头超额提升



数据来源: Wind、东吴证券研究所

数据日期: 2020年2月17日至2025年6月30日

表55: 样例高频因子叠加 AI 高频因子多头超额提升

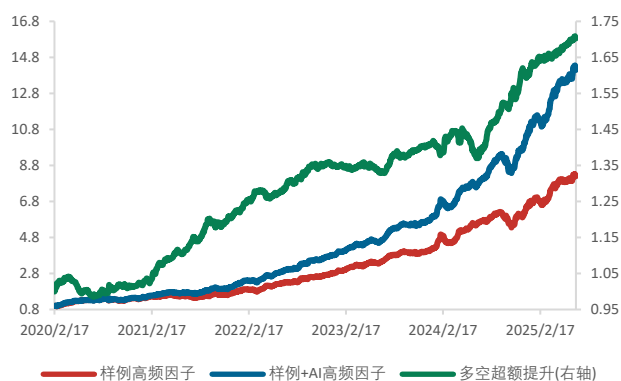
	多头超额收益			多头超额收益波动比		
	样例	样例+AI	提升	样例	样例+AI	提升
2020	3.87%	9.01%	5.14%	0.70	1.58	0.88
2021	0.82%	12.47%	11.65%	0.14	2.08	1.94
2022	10.72%	17.45%	6.74%	2.10	3.20	1.10
2023	5.08%	6.71%	1.63%	1.53	1.85	0.32
2024	9.48%	10.36%	0.88%	1.48	1.56	0.07
2025	-0.73%	2.03%	2.76%	-0.12	0.35	0.47
全区间	5.58%	11.19%	5.61%	1.03	2.00	0.97

数据来源: Wind、东吴证券研究所

数据日期: 2020年1月1日至2025年6月30日

而在更能体现纯粹 Alpha 能力的多空组合中，加入 AI 因子后，多空组合的年化超额收益从 50.21% 提升至 66.81%，年化收益波动比也从 3.37 提升至 4.75。

图37: 样例高频因子叠加 AI 高频因子多空超额提升



数据来源: Wind、东吴证券研究所

数据日期: 2020年2月17日至2025年6月30日

表56: 样例高频因子叠加 AI 高频因子多空超额提升

	多空超额收益			多空超额收益波动比		
	样例	样例+AI	提升	样例	样例+AI	提升
2020	39.25%	42.68%	3.43%	2.20	2.48	0.28
2021	27.90%	53.26%	25.36%	1.67	3.51	1.84
2022	59.18%	75.69%	16.51%	4.59	6.08	1.49
2023	43.24%	49.50%	6.26%	4.68	5.49	0.82
2024	62.90%	85.55%	22.65%	3.96	5.71	1.75
2025	23.91%	32.27%	8.36%	1.45	2.12	0.68
全区间	50.21%	66.81%	16.61%	3.37	4.75	1.38

数据来源: Wind、东吴证券研究所

数据日期: 2020年1月1日至2025年6月30日

综上所述，这些结果共同表明，通过本研究框架生成的 AI 高频因子，不仅自身具备选股能力，更重要的是，它们为现有的因子库提供了有效的补充，带来了可量化的增量价值。这证明了 AI 在挖掘新颖且有效的 Alpha 来源方面具备的潜力。

4.5. 高频因子对多频率融合的提升

在传统端到端的收益预测模型拟合中，因为显卡算力和显存限制等原因，通常仅将日频行情与周频行情输入 GRU 中训练。而高频数据计算量太大会对算力与显存带来较大压力，因此本文的思路是将高频数据降频成统计上显著的高频因子，然后在神经网络 AGRU 框架下，先构建不含高频因子的网络作为实验组，然后再构建加入 AI 高频因子的网络作为对照组，观察对照组相对于实验组的提升。实验组的数据处理方式如下：

4.5.1. 数据处理与模型参数

特征与标签：

- 特征 X：日 K 行情与周 K 行情，包括：最高价、开盘价、最低价、收盘价、VWAP 均价、成交量和成交额。价格数据皆除以最新一天收盘价做时序标准化，成交量与成交额皆除以最新一天值做时序标准化。
- 标签 Y：个股从下一个交易日的 VWAP 算起，未来 5 个交易日 VWAP 收益率，若当天股票停牌，则剔除该数据。
- 数据采样：日频采样，以 2018 年 1 月 2 日为第一次预测点，回看过去 5 个自然年数据作为样本，数据集按顺序划分，前 90% 的交易日作为训练集，后 10% 的交易日作为验证集，用于拟合模型，接下来 1 年用该模型预测每个股票的收益率，每个自然年重训练一次模型。

神经网络参数：

- Batch：按交易日做 Batch 拆分，每个 Batch 大小为当前截面股票数量。
- 损失函数：预测值与真实值 Pearson 相关性的相反数。
- 优化器：Adam；学习速率：0.001。
- 最大 Epoch：100；早停 Epoch：10（验证集 Loss 连续 10 轮没有创新低则停止训练）。
- 随机种子：42。（固定随机种子数保证模型对比实验路径的一致性）

训练集、验证集与测试集的数据处理：

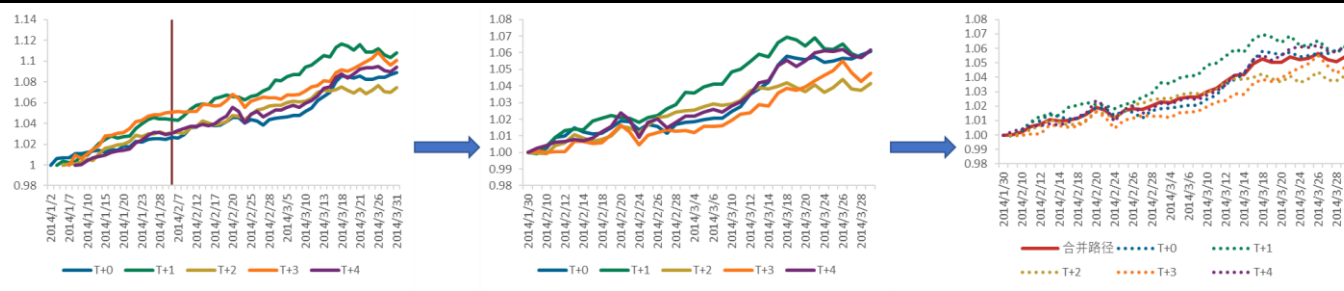
- 若预测未来 N 个交易日的均价收益率，则剔除样本内数据集的最后 N+1 个交易日的的数据，避免样本内的标签用到样本外的数据进行计算。
- 若预测未来 N 个交易日的均价收益率，则进一步剔除最后 (N-1)/2 个交易日的

训练集数据与最早 $(N-1)/2$ 个交易日的验证集数据，避免训练集未来收益率与验证集未来收益率的计算有用到重复的交易日数据。（若 N 为偶数，则训练集比验证集多剔除一个交易日的的数据）

全 A 选股组合回测参数与方式：

- 回测区间：2018 年 1 月 2 日至 2025 年 10 月 31 日。
- 剔除：剔除上市不满 365 个自然日的新股，剔除 ST 股。
- 组合构建：取因子值前 10% 为多头组合，后 10% 为空头组合，所有全 A 等权为基准组合。
- 交易方式：每 5 日调仓，以下一个交易日的 VWAP 价格成交，交易费率为单边千分之一。
- 交易限制：一字涨停不能买入，一字跌停不能卖出，停牌不可交易，多头 / 空头组合中的多余权重分给其余股票。
- 路径依赖与路径合并：日频因子，每一天皆可预测未来 5 日的收益率，这意味着我们可以利用这些因子在每个交易日对未来一段时间内的股票收益率进行预测。然而，如果我们选定某一个特定的起点并每隔 5 日进行一次调仓操作，那么由于起点选择的不同，可能会导致我们得到不同的净值曲线，这种现象被称为路径依赖问题，若任取一条净值曲线计算收益与风险都会有失公允。为规避此问题，我们将不同调仓起点的净值曲线做截断，并将所有净值曲线归一化处理，使得它们的初始值相同，然后计算所有净值曲线在每一天的收益率的均值。通过这种方法，我们可以得到一条综合的净值曲线，代表了在不同起点下调仓策略的平均表现。这样一来，我们能够更准确地评估投资策略的收益与风险，避免了单一调仓起点带来的路径依赖问题。

图38：回测路径合并方式示意图



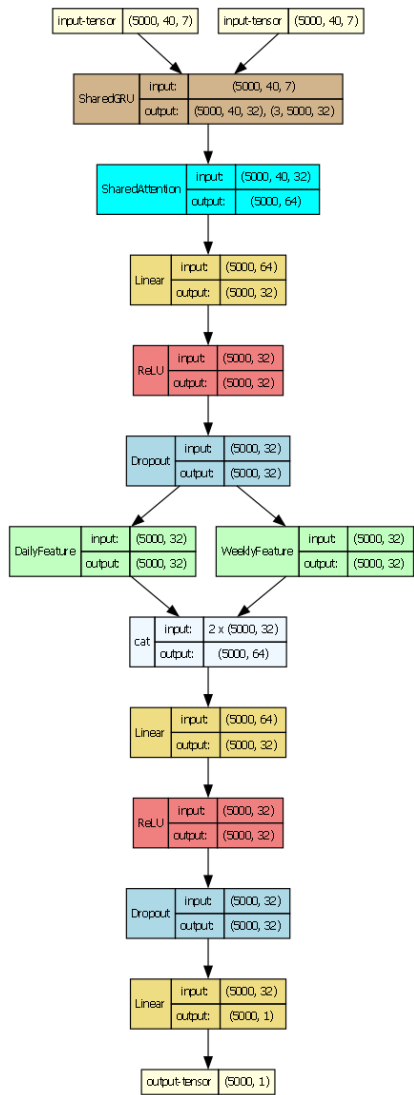
数据来源：东吴证券研究所绘制

4.5.2. 以日 K 与周 K 为特征的 AGRU

本文构建了一个基于 GRU 和注意力机制的双分支网络结构，专门用于处理和融合不同时间尺度的序列数据。

- 1) 双分支处理：模型有两个平行的处理分支，分别接收日线和周线 序列数据。
 - 2) 共享权重：两个分支中的核心计算模块（GRU 编码器和初期 MLP）是权重共享的。这使得模型能够学习到适用于不同时间尺度的通用特征表示，增强了模型的泛化能力并减少了参数量。
 - 3) 注意力：每个分支都使用注意力机制来动态地关注序列中最具信息量的部分，生成一个更能代表整个序列的上下文向量。
 - 4) 特征融合与预测：最后，将两个分支提取出的高级特征进行拼接融合，并通过一个最终的全连接网络（MLP）输出预测结果。
- ✓ 设计哲学：既通过周线分支洞察由宏观政策和基本面决定的长期价值趋势，也通过日线分支紧跟由市场情绪和资金博弈驱动的短期价格波动，并最终通过注意力机制动态回顾历史上的关键时刻，同时立足于最新的市场状态，做出综合判断。

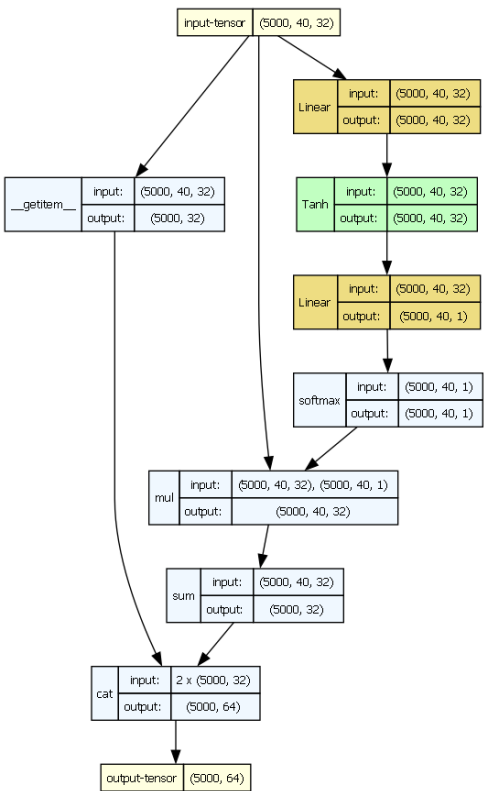
图39：以日 K+周 K 为特征的 AGRU 融合模型



数据来源：东吴证券研究所绘制

备注：input-tensor 的维度分别代表(股票,时间步,特征)

图40：Attention 模块内部结构

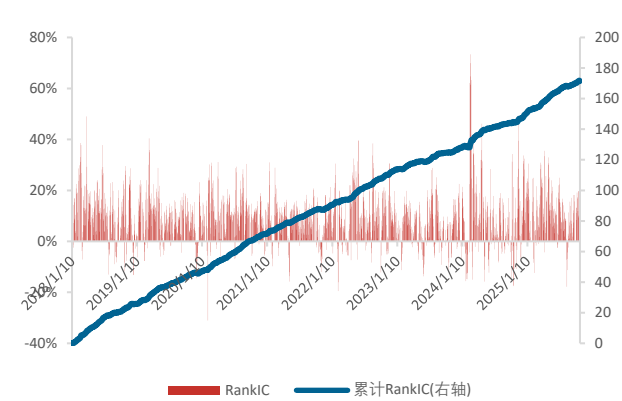


数据来源：东吴证券研究所绘制

备注：input-tensor 的维度分别代表(股票,时间步,特征)

5 日频率下，以日 K+周 K 为特征，通过 AGRU 模型得到的因子 RankIC 均值为 9.06%，费后多头超额年化收益为 18.24%，收益波动比为 2.63，单边年化换手率为 30.17 倍。

图41：日 K+周 K 因子每期 RankIC 与累计 RankIC



数据来源：Wind、东吴证券研究所

数据日期：2018 年 1 月 10 日至 2025 年 10 月 31 日

图42：日 K+周 K 因子多头组合净值与多头超额净值



数据来源：Wind、东吴证券研究所

数据日期：2018 年 1 月 3 日至 2025 年 10 月 31 日

表57：日 K+周 K 因子 IC 统计指标

	RankIC	ICIR	t 统计值	胜率
2018	10.87%	1.06	16.29	87.34%
2019	8.57%	1.04	16.30	84.02%
2020	10.26%	1.25	19.41	91.36%
2021	7.69%	1.04	16.23	85.19%
2022	9.61%	0.96	14.86	85.12%
2023	5.89%	0.75	11.64	76.86%
2024	9.14%	0.59	9.20	72.31%
2025	10.80%	1.14	16.16	89.50%
全区间	9.06%	0.90	39.31	83.84%

数据来源：Wind、东吴证券研究所

数据日期：2018 年 1 月 3 日至 2025 年 10 月 31 日

表58：日 K+周 K 因子分年度多头超额绩效指标

年份	年化收益	年化波动	最大回撤	收益波动比	收益回撤比
2018	35.95%	5.47%	-2.10%	6.57	17.14
2019	14.53%	3.97%	-1.76%	3.66	8.27
2020	17.72%	5.81%	-2.35%	3.05	7.53
2021	13.21%	6.43%	-4.30%	2.06	3.07
2022	19.74%	7.79%	-4.31%	2.53	4.58
2023	-0.93%	3.95%	-3.90%	-0.23	-0.24
2024	20.88%	11.84%	-5.17%	1.76	4.04
2025	18.68%	6.92%	-2.30%	2.70	8.12
全区间	18.24%	6.95%	-5.17%	2.63	3.53

数据来源：Wind、东吴证券研究所

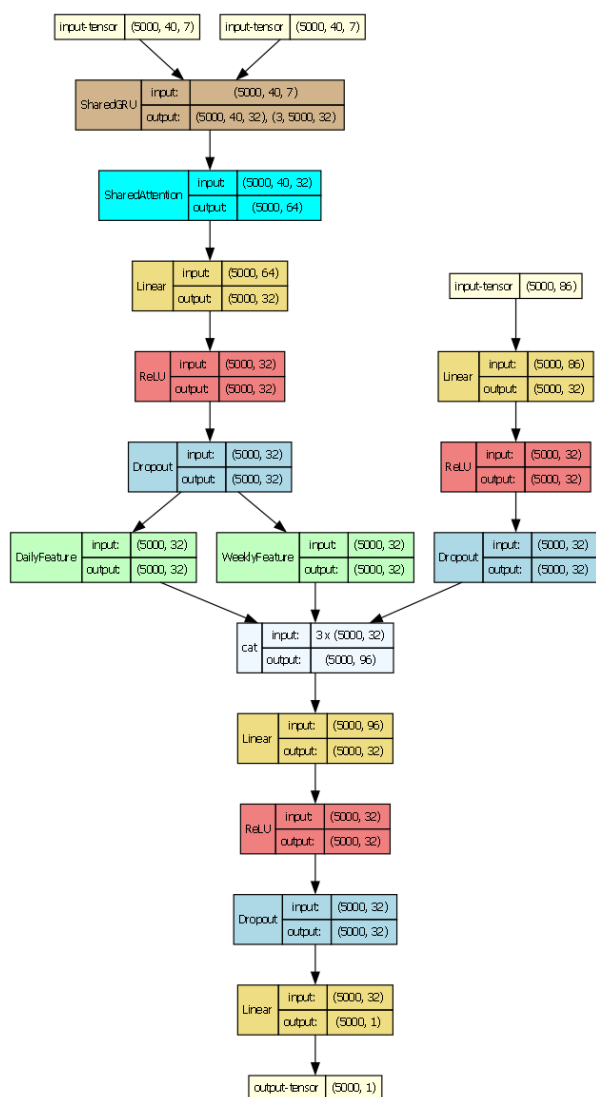
数据日期：2018 年 1 月 3 日至 2025 年 10 月 31 日

4.5.3. 以日 K+周 K+高频因子为特征的 AGRU

在此前的基础上，本文继续构建了一个融合了多时间尺度序列与横截面因子特征的多分支异构网络。它通过并行处理不同类型的数据源，实现对市场信息的立体化综合分析。

- 1) 双分支时序处理：模型的核心部分依然是一个双分支结构，用于并行处理日线和周线的价格序列数据。
 - 2) 共享时序权重：为了学习通用的时间模式，日线和周线分支共享同一套时序处理模块，包括 GRU 编码器、注意力模块和特征提炼 MLP。
 - 3) 独立因子处理：模型设有一个独立的并行分支，专门用于提取高频因子中的特征。
 - 4) 多源特征融合与预测：在最后阶段，模型将从三个来源提取出的高级特征——日线特征、周线特征和高频因子特征进行拼接，形成一个全面的、包含长短期趋势和高频因子信号的“决策向量”。该向量最终被送入最后的 MLP 进行深度融合与预测，输出最终结果。
- ✓ 设计哲学：通过共享权重的双分支结构，分别从周线和日线中洞察长期战略趋势与中期战术波动；同时，它利用一个独立的并行分支来解析高频因子所蕴含的、传统价格序列之外的瞬时量化信号；在最终决策层融合这三个不同维度但互补的信息流，从而形成一个远比单一视角更加立体、稳健和全面的综合判断。

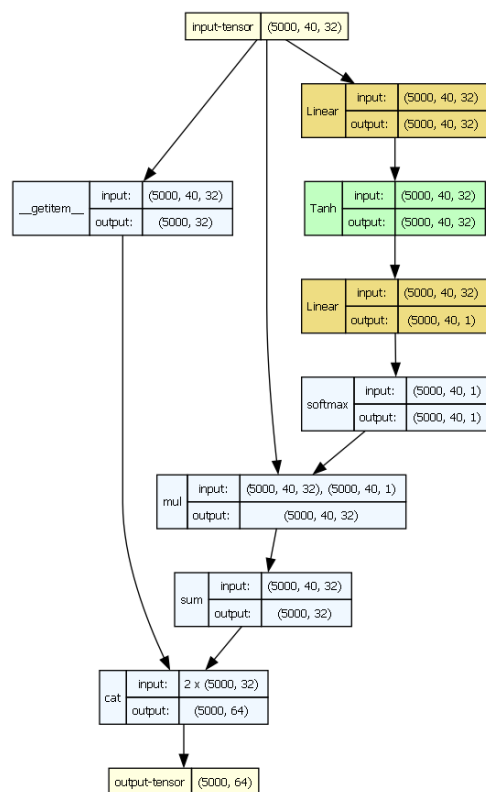
图43：以日 K+周 K+高频因子为特征的 AGRU 模型



数据来源：东吴证券研究所绘制

备注：input-tensor 的维度分别代表(股票,时间步,特征)

图44：Attention 模块内部结构

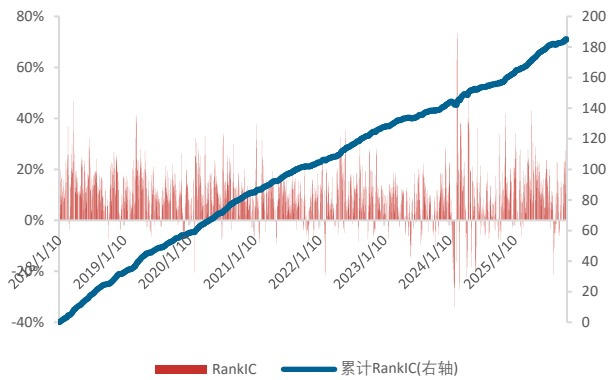


数据来源：东吴证券研究所绘制

备注：input-tensor 的维度分别代表(股票,时间步,特征)

5 日频率下，以日 K+周 K+高频因子为特征，通过 AGRU 模型得到的因子 RankIC 均值为 9.77%，相较于不加高频因子的前期结果提升 0.71pct，费后多头超额年化收益为 25.28%，提升 7.04pct，收益波动比为 3.57，提升 0.94，单边年化换手率为 31.24 倍。

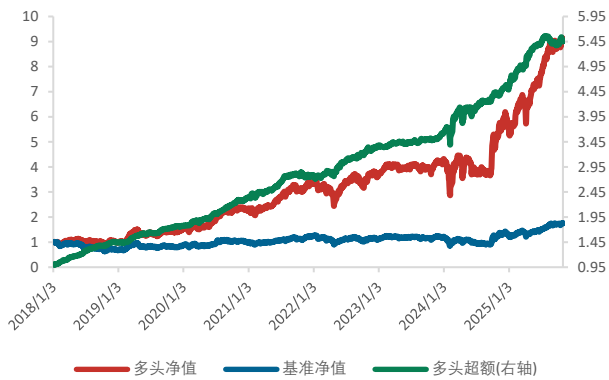
图45：日 K+周 K+高频因子每期 RankIC 与累计 RankIC



数据来源：Wind、东吴证券研究所

数据日期：2018 年 1 月 10 日至 2025 年 10 月 31 日

图46：日 K+周 K+高频因子多头组合净值与多头超额净值



数据来源：Wind、东吴证券研究所

数据日期：2018 年 1 月 3 日至 2025 年 10 月 31 日

表59：日 K+周 K+高频因子 IC 统计指标

	RankIC	ICIR	t 统计值	胜率
2018	13.42%	1.73	26.69	96.62%
2019	10.68%	1.33	20.80	90.16%
2020	11.13%	1.42	22.07	93.42%
2021	8.02%	1.07	16.68	83.13%
2022	9.51%	1.04	16.10	85.12%
2023	6.31%	0.85	13.28	80.58%
2024	8.03%	0.51	8.01	76.45%
2025	11.35%	1.14	16.11	86.00%
全区间	9.77%	1.00	43.68	86.42%

数据来源：Wind、东吴证券研究所

数据日期：2018 年 1 月 3 日至 2025 年 10 月 31 日

表60：日 K+周 K+高频因子分年度多头超额绩效指标

年份	年化收益	年化波动	最大回撤	收益波动比	收益回撤比
2018	44.51%	4.61%	-1.20%	9.66	37.17
2019	22.48%	4.06%	-1.47%	5.53	15.33
2020	32.47%	6.22%	-2.85%	5.22	11.40
2021	18.96%	7.16%	-3.71%	2.65	5.12
2022	20.78%	7.36%	-3.80%	2.82	5.46
2023	7.50%	3.77%	-1.80%	1.99	4.17
2024	24.93%	12.52%	-9.43%	1.99	2.64
2025	20.69%	6.86%	-3.41%	3.02	6.08
全区间	25.28%	7.08%	-9.43%	3.57	2.68

数据来源：Wind、东吴证券研究所

数据日期：2018 年 1 月 3 日至 2025 年 10 月 31 日

4.5.4. AGRU 融合日 K+周 K+高频因子的周频 800 指增

本文采用上节中, AGRU 融合日 K+周 K+高频因子后, 得到的最终 Alpha, 构建 800 指增组合, 指增组合约束条件与交易方式如下:

约束条件:

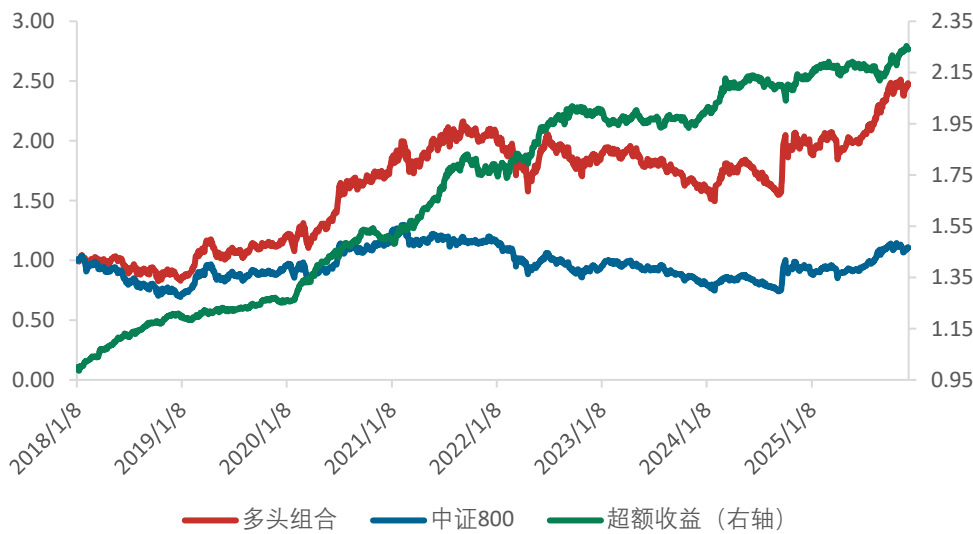
- 1) 100%指数成分股内选股。
- 2) 个股权重最大偏离 0.8%。
- 3) 中信一级行业最大偏离 3%。
- 4) 市值风格最大偏离 0.01。
- 5) 单边换手率 40%以下。

交易方式:

- 1) 周度调仓, 以每月初第一个交易日的 VWAP 价格成交。
- 2) 一字涨停不能买入, 一字跌停不能卖出, 停牌不能交易。
- 3) 手续费: 单边千分之一。

AGRU 融合日 K+周 K+高频因子的周频 800 指增, 2018 年至 2025 年 12 月 10 日年化收益为 12.57%, 年化信息比为 2.1, 单边年化换手率 21.24 倍。相对基准指数的年化超额收益为 11.15%, 收益波动比为 2.18, 收益回撤比为 2.22。

图47：AGRU 融合日 K+周 K+高频因子的周频 800 指增



数据来源：Wind、东吴证券研究所
数据日期：2018 年 1 月 8 日至 2025 年 12 月 10 日

表61：中证 800 指增组合收益风险特征指标

年份	年化 收益	年化 波动	最大 回撤	年化 信息比	收益 波动比	收益 回撤比
2018	-15.00%	22.40%	-20.70%	4.87	-0.67	-0.72
2019	38.62%	20.26%	-14.46%	0.99	1.91	2.67
2020	50.92%	24.35%	-15.97%	3.86	2.09	3.19
2021	17.77%	20.06%	-13.45%	2.90	0.89	1.32
2022	-12.16%	21.87%	-24.77%	1.81	-0.56	-0.49
2023	-11.02%	12.43%	-19.34%	-0.21	-0.89	-0.57
2024	20.74%	22.95%	-16.05%	1.39	0.90	1.29
2025	24.91%	15.87%	-11.00%	1.18	1.57	2.26
全区间	12.57%	20.40%	-30.91%	2.10	0.62	0.41

数据来源：Wind、东吴证券研究所
数据日期：2018 年 1 月 8 日至 2025 年 12 月 10 日

表62：中证 800 指增组合超额收益指标

年份	年化 收益	年化 波动	最大 回撤	收益 波动比	收益 回撤比
2018	20.85%	4.14%	-1.44%	5.04	14.44
2019	3.62%	3.79%	-2.05%	0.96	1.76
2020	20.21%	4.98%	-2.84%	4.05	7.11
2021	19.08%	6.32%	-4.42%	3.02	4.31
2022	11.78%	6.54%	-3.28%	1.80	3.59
2023	-0.87%	4.03%	-3.64%	-0.22	-0.24
2024	7.51%	5.55%	-4.65%	1.35	1.61
2025	4.96%	4.60%	-3.38%	1.08	1.47
全区间	11.15%	5.11%	-5.03%	2.18	2.22

数据来源：Wind、东吴证券研究所
数据日期：2018 年 1 月 8 日至 2025 年 12 月 10 日

5. AI 文本分析：对量化选股的增益

5.1. 通过 Gemini 模型读取调研文本情绪

5.1.1. 近年来调研数量与字数有所提升

近年来上市公司调研数量每年超过 2 万。根据调研公告日期统计，截至 2025 年 11 月 30 日，2020 年初以来调研数量超过 12 万，其中 2022 至 2025 每年调研数量均超过 2 万。

表63：调研数量与字数

	公告数量(万)				公告字数(万)			
	中证 800	中证 1000	1800 以外	全 A	中证 800	中证 1000	1800 以外	全 A
2020	0.23	0.25	0.39	0.87	508.49	560.33	924.88	1993.70
2021	0.34	0.38	0.78	1.50	912.64	986.60	2109.72	4008.96
2022	0.45	0.55	1.30	2.31	1323.75	1486.59	3405.07	6215.42
2023	0.55	0.61	1.63	2.80	1450.40	1461.87	3648.51	6560.78
2024	0.61	0.56	1.49	2.66	1473.66	1307.85	3250.23	6031.74
2025	0.51	0.52	1.33	2.36	1333.36	1249.39	2872.67	5455.42

数据来源：Wind、东吴证券研究所

数据日期：2020 年 1 月 1 日至 2025 年 11 月 30 日

按月统计调研数量与字数，整体上每年 4、5、9、11 几个月调研相对较多。2024 年 5 月调研数量与字数最多，数量合计超 5000，字数合计超 1300 万，其中中证 800 调研数量 873，字数 244.12 万，中证 1000 调研数量 954，字数 270.75 万，其他调研数量 3228，字数 862.60 万。

表64：全市场每月调研数量（个）

全 A	1 月	2 月	3 月	4 月	5 月	6 月	7 月	8 月	9 月	10 月	11 月	12 月
2020	473	478	420	586	1021	955	782	658	1059	477	984	783
2021	740	418	870	1627	2146	1079	1027	1030	1759	802	1766	1747
2022	1335	1036	1526	2344	3270	1832	1695	1636	2935	1195	2510	1742
2023	994	1979	2058	2490	4243	2148	1653	1549	3956	1555	3520	1865
2024	1679	978	1632	2819	5055	1869	1597	1611	3163	1517	2798	1926
2025	1040	1155	1574	2836	4652	1750	1493	1660	3469	1419	2571	

数据来源：Wind、东吴证券研究所

数据日期：2020 年 1 月 1 日至 2025 年 11 月 30 日

表65：全市场每月调研字数(万)

	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
2020	80.00	127.17	102.41	176.88	275.62	204.81	156.26	153.48	238.63	110.39	193.68	174.39
2021	127.60	90.02	219.49	584.53	733.89	245.90	245.32	288.19	466.79	210.46	398.11	398.65
2022	306.42	223.69	391.91	858.86	1196.77	463.13	372.93	427.28	694.71	289.32	572.15	418.25
2023	217.87	382.70	482.21	794.99	1243.79	508.87	346.49	408.19	845.36	329.63	679.46	321.22
2024	292.61	171.81	320.66	854.04	1377.47	373.92	301.83	405.59	681.90	343.44	533.46	375.00
2025	189.05	206.25	322.50	805.60	1229.20	359.94	278.66	430.29	780.20	312.07	541.65	

数据来源：Wind、东吴证券研究所

数据日期：2020 年 1 月 1 日至 2025 年 11 月 30 日

5.1.2. 综合考虑调研情况，选用 Gemini 2.5 Pro 模型

考虑近年来调研情况，选用 Gemini 2.5 Pro 模型进行周度文本打分较为合理。分别设定 Prompt(system)与 Prompt(user)。为避免其他因素对打分结果的影响，在给出 Prompt 前，对股票名称进行遮蔽处理，如 Prompt(system)中{股票序号}为["股票 1","股票 2","股票 3","股票 4","股票 5","股票 6","股票 7","股票 8","股票 9",……]。

你是一位专业的金融分析师，能够理解调研活动文本中的专业名词并了解公司所属行业情况。请仅基于"调研正文"内容判断{股票序号}中每只股票投资价值、情绪，对未来股价预期进行打分。

要求不得引入股票对应调研正文以外的主观臆断或未提及的信息，且严格遵循以下要求：

- 1.根据每只股票的调研正文对股票未来预期打分，不得漏掉指定股票或者引入其他未提及股票，不得重复打分；
- 2.打分原则为预期越好分数越高，打分区间在 1 至 100 之间，要求预期利好/正向/乐观的股票分数大于 50，预期利空/谨慎/负面的股票分数小于 50，预期中性为分数为 50，没有预期分数为空("");
- 3.输出格式：仅保留包含股票代码与打分的 JSON 数据(键名统一为‘代码’‘分数’)，无需任何多余文字说明，无外层嵌套键，必须包含且仅给定的股票对象，股票代码不重复。
- 4.格式规范：JSON 语法需完全合法(逗号使用正确、引号为英文双引号、无多余空格或换行)，生成后用 JSON.cn 校验。

注意：调研普遍存在“乐观表述偏多、风险信息弱化”的偏差，请甄别文本中乐观表述的“真实性与实质性”，而非直接采信。首先请注意甄别实质性利好(如已落地订单、毛利率/净利率环比提升、核心技术突破且已量产)与空泛的乐观表述(如行业前景广阔、未来计划扩产)，实质利好的正向得分权重应高于空泛乐观表述。其次除明显的风险点(如利润减少、成本波动、业务风险等)外,仍需重点识别未明确披露的风险点进行扣分处理，如成本压力可控、竞争格局稳定等被弱化或隐藏的风险。

以下为两个调研文本示例:5 分示例：{公告内容 1}90 分示例：{公告内容 2}"

在 Prompt(user)中，仅输入股票序号，所属中信二级行业，调研文本，不输入股票名称、代码、调研日期、公告日期等内容。

```
[{"股票序号": "股票 1", "行业": "XX", "调研正文": "调研内容"},  
{"股票序号": "股票 2", "行业": "XX", "调研正文": "调研内容"},  
.....]
```

根据公告日期划分，一次性输入每周所有调研内容。根据测试经验，当字数超 130 万时对当周股票进行拆分。

为降低大模型输出的随机性，将部分参数设定如下。

表66：大模型参数设定

参数	设定值	作用	说明
temperature	0	控制输出随机性	0 = 完全确定，1 = 最大随机
top_p	1	采样阈值	设为 1.0 时，temperature=0 会完全禁用采样随机性
top_k	1	采样候选词数量	仅保留 1 个最优候选词，彻底消除选择随机性
seed	88888	随机种子，固定整数（如 12345）	锁定随机数生成器，进一步确保推理逻辑一致
presence_penalty	0	禁用“新词惩罚”	若设为正数会随机抑制重复 token，导致输出波动
frequency_penalty	0	禁用“高频词惩罚”	正数会随机调整高频 token 的概率

数据来源：Gemini，东吴证券研究所

根据 Prompt(system)要求，输出结果仅包含代码与打分结果，以下为某周输出结果示例。

输出：
[{ "代码": "股票 1", "分数": 60 },
{ "代码": "股票 2", "分数": 65 },
{ "代码": "股票 3", "分数": 68 },]

5.1.3. 打分结果展示

对 2020 年至 2025 年 11 月 30 日调研文本调用 Gemini 2.5 Pro 大模型进行周度打分，中证 800 周度数据运行时间中位数约 3 分钟，但具体运行时间受 API 稳定性、输入内容长度等因素的影响。

表67：大模型运行时间及输入输入文本长度

中证 800	运行时间(秒)	输入长度 (个)	输出长度 (个)
中位数	185.59	199073	2397
均值	261.59	247227	2748
最大值	2464.97	798270	10294
最小值	6.86	7052	53

数据来源：Gemini，东吴证券研究所

5.2. 从调研评分到调研因子：双速动态衰减模型

调研评分仅当周存在有效值，但其对股价的影响可能是一个持久但动态衰减的过程，因此构建一个高度模拟市场信息处理行为的因子。该模型能智能识别 “新信息冲击”、“信息自然衰减” 和 “信息完全失效” 三种状态，并采用不同的处理逻辑。

- **惰性启动**：因子在首次观测到有效调研评分前，其值始终为 NaN（无有效信号）。

- **双速处理逻辑**：模型的核心，根据当期有无新调研，在两种状态间自动切换。

➤ **更新状态**（有新调研 S_t ）：采用极高的更新率，因子值迅速向新评分融合。

$$V_t = 0.1 * V_{t-1} + 0.9 * S_t$$

➤ **衰减状态**（无新调研）：采用较慢的衰减率，因子值平滑地向中性值 50 回归。

$$V_t = 0.75 * V_{t-1} + 0.25 * 50$$

- **信号失效机制**：设置一个硬性的“保质期”。当一只股票连续 26 周（约半年）无任何新调研信息时，其旧信息被认为完全失效。因子值将被强制重置为 NaN，直至下一次新调研出现将其“唤醒”。这避免了陈旧无效信息对模型的持续干扰。

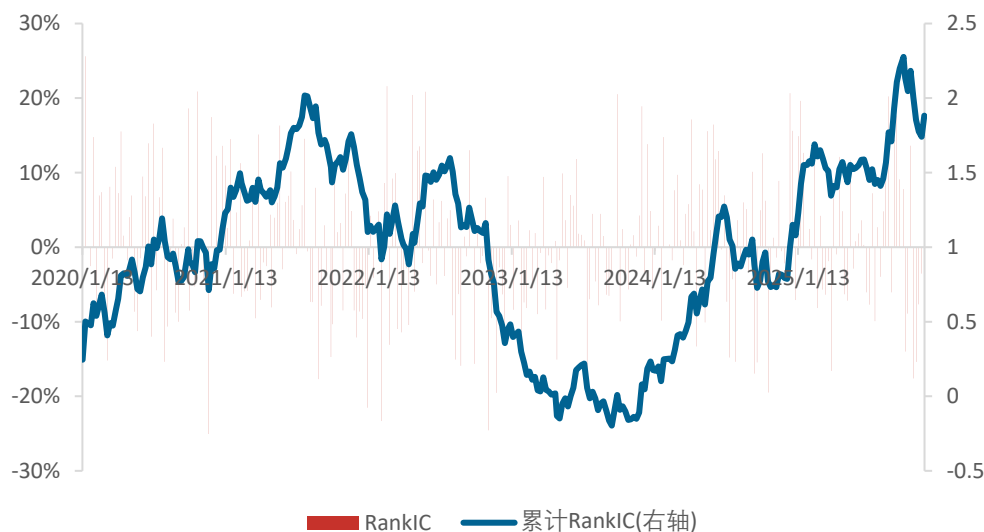
◆ 最终因子特性：

- 响应迅速**：对新出现的调研信息能做出快速、充分的反应。
- 衰减平滑**：在信息空窗期，因子值变化连续，避免了断崖式下跌。
- 逻辑严谨**：“惰性启动”和“信号失效”机制确保了因子在任何时候都具有明确的经济学含义。

5.3. 调研因子特色：独特的空头规避能力

在中证 800 中，经过双速动态衰减模型处理后的调研因子，RankIC 不稳定。分层测试的单调性一般，但排名靠后的组别（第 7 组至第 10 组）收益呈单调递减，且第 10 组的超额收益显著为负跑输其他组，体现空头识别能力。说明在上市公司调研中，有利好消息股价不一定上涨（可能已被定价），但有利空消息，则大概率会反应在股价上。

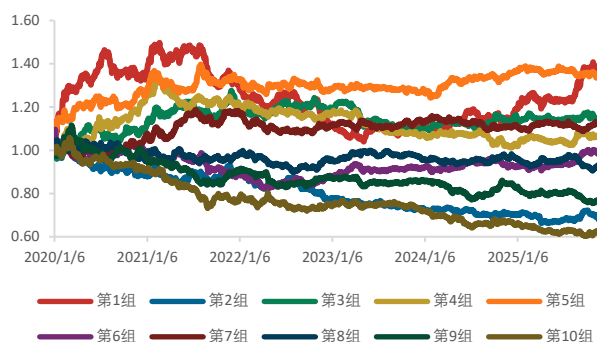
图48：调研文本因子在中证 800 中 RankIC



数据来源：Wind、东吴证券研究所

数据日期：2020 年 1 月 13 日至 2025 年 12 月 1 日

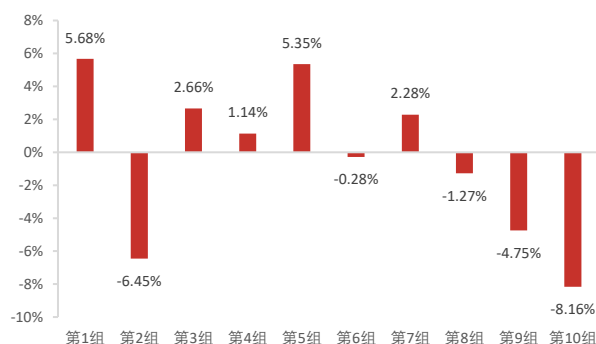
图49：调研文本因子在中证 800 中分层超额收益



数据来源：Wind、东吴证券研究所

数据日期：2020 年 1 月 6 日至 2025 年 11 月 28 日

图50：调研文本因子在中证 800 中分层年化超额收益

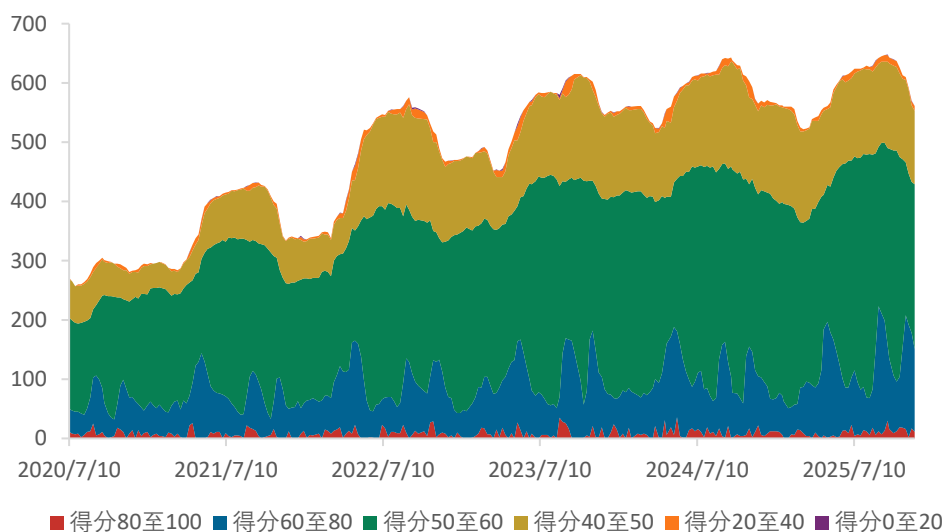


数据来源：Wind、东吴证券研究所

数据日期：2020 年 1 月 6 日至 2025 年 11 月 28 日

若按因子值分组,可见大部分上市公司调研评分最集中在 40 至 50 的区间,整体呈中性略偏多的情绪反应,同时该组股票相对 800 等权超额收益基本走平。40 至 50、20 至 40、0 至 20 分组的超额收益呈单调递减,时序覆盖股票数量均值分别为 115.13 个、6.69 个、0.26 个,可见极低分股票数量虽然不多,但空头胜率较高。但 50 分以上组别的超额收益单调性较差,再次说明调研因子有较强的空头识别能力,而依赖该因子赚取正超额可能相对困难。

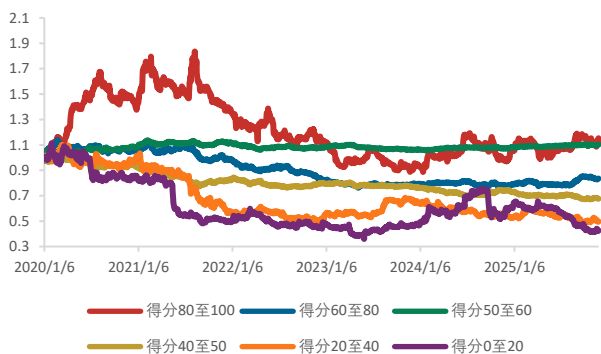
图51: 调研文本因子各组数量



数据来源: Wind、东吴证券研究所

数据日期: 2020 年 7 月 10 日至 2025 年 11 月 28 日

图52: 调研文本因子在中证 800 中按得分分层超额收益

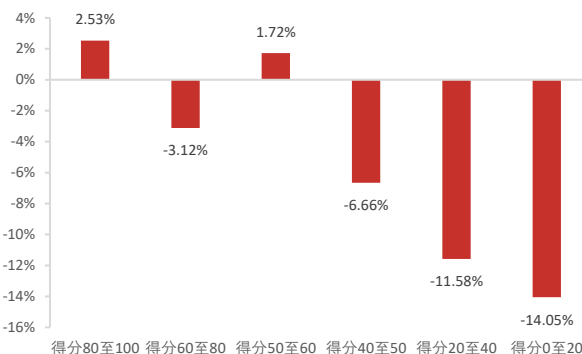


数据来源: Wind、东吴证券研究所

数据日期: 2020 年 1 月 6 日至 2025 年 11 月 28 日

备注: 得分分组取值为左开右闭

图53: 调研文本因子在中证 800 中按得分分层年化超额收益



数据来源: Wind、东吴证券研究所

数据日期: 2020 年 1 月 6 日至 2025 年 11 月 28 日

备注: 得分分组取值为左开右闭

与传统量价因子（20 日低波、20 日反转）和传统基本面因子（ROE、归母净利润同比）相比较，调研因子空头端收益更显著且稳健。（空头超额为等 800 等权相较于空头组的超额收益）

调研因子全区间年化空头超额收益为 8.26%，空头收益波动比为 1.09，优于传统量价与基本面因子。

调研因子同传统量价与基本面因子的相关性较低，与 20 日反转和 20 日低波相关性为负，与 ROE 和归母净利润同比相关性在 10%左右。

图54：调研文本因子与其它因子空头超额收益



数据来源：Wind、东吴证券研究所

数据日期：2020 年 1 月 6 日至 2025 年 11 月 28 日

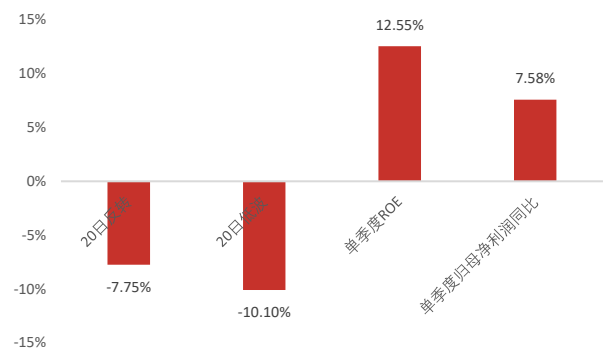
表68：调研文本因子与其它因子分年度空头超额收益

年份	调研因子	20 日反转	20 日低波	单季度 ROE	单季度归母净利润同比
2020	9.26%	-4.49%	-4.34%	6.92%	14.08%
2021	18.09%	5.41%	4.79%	10.36%	16.70%
2022	2.04%	32.84%	15.71%	-1.04%	-0.19%
2023	2.91%	4.31%	2.48%	0.89%	-5.34%
2024	9.43%	18.79%	27.87%	4.99%	11.84%
2025	5.81%	-6.34%	-19.24%	-5.59%	-7.57%
全区间	8.26%	8.05%	3.68%	2.77%	4.74%

数据来源：Wind、东吴证券研究所

数据日期：2020 年 1 月 6 日至 2025 年 11 月 28 日

图55：调研文本因子与其它因子的因子值相关性



数据来源：Wind、东吴证券研究所

数据日期：2020 年 1 月 6 日至 2025 年 11 月 28 日

表69：调研文本因子与其它因子分年度收益波动比

年份	调研因子	20 日反转	20 日低波	单季度 ROE	单季度归母净利润同比
2020	0.99	-0.30	-0.28	0.93	1.65
2021	2.03	0.26	0.26	1.41	1.34
2022	0.26	2.28	1.21	-0.19	-0.02
2023	0.58	0.29	0.16	0.19	-0.75
2024	1.27	1.13	1.53	0.45	1.28
2025	1.02	-0.38	-0.96	-0.69	-0.86
全区间	1.09	0.49	0.22	0.36	0.52

数据来源：Wind、东吴证券研究所

数据日期：2020 年 1 月 6 日至 2025 年 11 月 28 日

5.4. 调研因子对成熟指增框架的再提升

如何利用调研因子？经过上述分析可见，若直接将调研因子作为一个选股因子加入到线性或者非线性因子融合模型中，可能因为调研因子本身 IC 较低的特性而被融合模型赋予较低的权重。考虑到其空头端较强且与其他因子相关性低的特点，本文将其作为对最终 Alpha 因子的负向调整，最后让优化器控制这些调研情绪面较空股票的权重，具体方式如下，约束条件与交易方式同 3.5.4 节中陈述的方法：

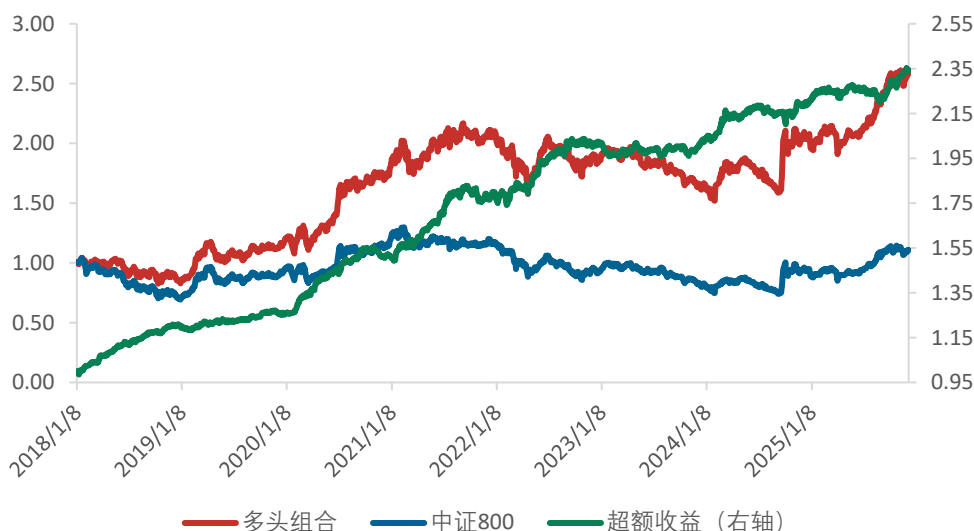
- 1) 在每个调仓时点，筛选出经过双速动态衰减模型处理后 50 分以下的调研因子。
- 2) 根据调研分数，计算当前时点对 Alpha 因子的调整值：

$$adjust_value_t^i = 20 * (\frac{Senti_value_t^i}{50} - 1)$$

- 3) 将 adjust_value 与 ZSCORE 之后的 Alpha 因子相加, 得到空头调整后的 Alpha 因子, 带入优化器计算权重。

AGRU 模型融合日 K+周 K+高频因子，结合调研因子的周频 800 指增，2018 年至 2025 年 12 月 10 日年化收益为 13.24%，年化信息比为 2.21，单边年化换手率 21.25 倍。相对基准指数的年化超额收益为 11.81%，收益波动比为 2.31，收益回撤比为 2.5。

图56: AGRU 融合日 K+周 K+高频因子（空头调整）的周频 800 指增



数据来源：Wind、东吴证券研究所

数据日期：2018年1月8日至2025年12月10日

表70：中证 800 指增组合(空头调整)收益风险特征指标

年份	年化 收益	年化 波动	最大 回撤	年化 信息比	收益 波动比	收益 回撤比
2018	-15.00%	22.40%	-20.70%	4.87	-0.67	-0.72
2019	38.62%	20.26%	-14.46%	0.99	1.91	2.67
2020	52.06%	24.26%	-15.53%	3.96	2.15	3.35
2021	17.31%	20.14%	-13.74%	2.80	0.86	1.26
2022	-11.76%	21.91%	-24.99%	1.87	-0.54	-0.47
2023	-10.11%	12.35%	-18.15%	0.05	-0.82	-0.56
2024	22.56%	22.82%	-15.40%	1.67	0.99	1.46
2025	26.41%	15.90%	-10.96%	1.48	1.66	2.41
全区间	13.24%	20.38%	-29.91%	2.21	0.65	0.44

数据来源：Wind、东吴证券研究所

数据日期：2018 年 1 月 8 日至 2025 年 12 月 10 日

表71：中证 800 指增组合(空头调整)超额收益指标

年份	年化 收益	年化 波动	最大 回撤	收益 波动比	收益 回撤比
2018	20.85%	4.14%	-1.44%	5.04	14.44
2019	3.62%	3.79%	-2.05%	0.96	1.76
2020	21.09%	5.05%	-3.13%	4.18	6.74
2021	18.63%	6.40%	-4.02%	2.91	4.64
2022	12.29%	6.56%	-3.46%	1.87	3.55
2023	0.13%	3.97%	-3.14%	0.03	0.04
2024	9.09%	5.53%	-3.85%	1.64	2.36
2025	6.22%	4.58%	-3.35%	1.36	1.86
全区间	11.81%	5.12%	-4.72%	2.31	2.50

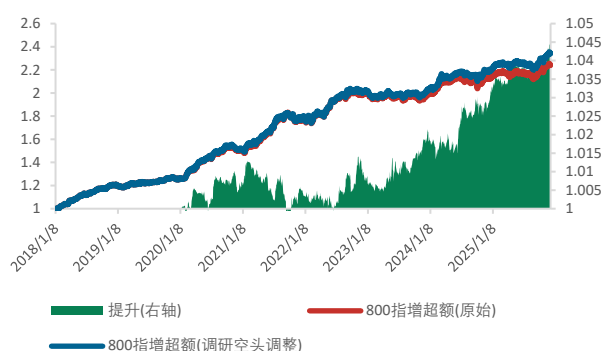
数据来源：Wind、东吴证券研究所

数据日期：2018 年 1 月 8 日至 2025 年 12 月 10 日

详细对比调研因子空头端对 800 指增的提升效果，可见从 2020 年至 2025 年，仅 2021 年是略微负优化，其它年份均是正提升，且近三年的超额收益提升在 1% 以上。

分析空头调整的影响范围，在时间衰减算法影响下，调整主要发生在 5 月、9 月与 11 月。调整数量上，以 2025 年 9 月 26 日的调仓为例，因子值调整在 -1 以下（评分 47.5 以下）的股票有 66 只，调整在 -3 以下（评分 42.5 以下）的股票有 23 只，调整在 -5 以下（评分 37.5 以下）的股票有 9 只，调整在 -10 以下（评分 25 以下）的股票有 3 只。

图57：原始 800 指增与空头调整 800 指增收益提升对比



数据来源：Wind、东吴证券研究所

数据日期：2018 年 1 月 8 日至 2025 年 12 月 10 日

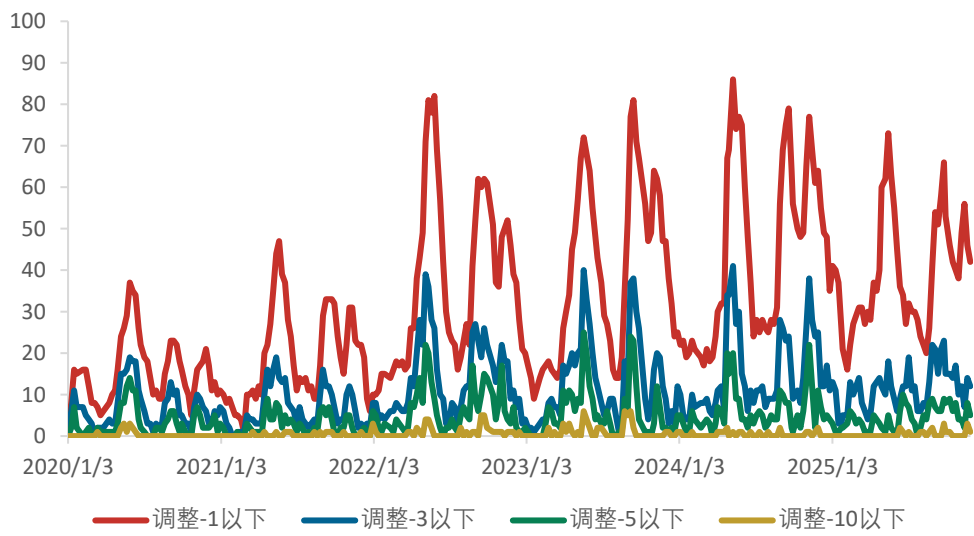
表72：原始 800 指增与空头调整 800 指增收益提升对比

年份	800 指增超额(调研空头调整)	800 指增超额(原始)	提升
2018	20.85%	20.85%	0.00%
2019	3.62%	3.62%	0.00%
2020	21.09%	20.21%	0.88%
2021	18.63%	19.08%	-0.45%
2022	12.29%	11.78%	0.51%
2023	0.13%	-0.87%	1.00%
2024	9.09%	7.51%	1.59%
2025	6.22%	4.96%	1.26%

数据来源：Wind、东吴证券研究所

数据日期：2018 年 1 月 8 日至 2025 年 12 月 10 日

图58：调研因子(50 分以下)调整分数与受影响股票数量（单位：个）



数据来源：Wind、东吴证券研究所

数据日期：2020 年 1 月 3 日至 2025 年 11 月 28 日

6. AI 赋能量化研究：范式优势与当前局限

本报告系统性地展示了一套基于大语言模型（LLM）驱动的因子挖掘与策略增强框架。通过在低频量价、基本面、高频数据及另类文本数据上的成功应用，我们验证了 AI 作为一种新型研究工具的巨大潜力。在此，我们对该方法论的范式优势、固有局限以及未来发展方向进行深入探讨。

一、 范式优势与核心贡献

将大语言模型引入因子研究，其最核心的贡献在于推动了量化研究从传统的纯“数据驱动”向“知识与逻辑驱动”的范式演进。这一转变带来了几项关键优势：

逻辑清晰与高可解释性，有效规避过拟合：传统的数据驱动方法，如遗传规划（Genetic Programming）或纯粹的神经网络，常常通过暴力搜索找到最优的数据拟合解，但其生成的因子表达式往往缺乏明确的经济或金融学逻辑，形如“黑箱”，这使其在样本外存在较高的过拟合风险。与之相反，本报告中的 AI 生成范式，无论是对 std20 因子的逐层优化，还是生成 REP_LF（留存收益/市值）这类基本面因子，其每一步操作和最终产出都伴随着清晰的逻辑解释。这种基于逻辑的探索使得因子更具鲁棒性，更容易被研究员理解、接受和迭代，从而在源头上降低了数据挖掘的风险。

内嵌领域知识，显著提升基本面与高频因子挖掘质量：AI，特别是经过海量文本训练的大语言模型，内化了大量关于金融、会计和市场微观结构的先验知识。在处理基本面数据时，AI 能够理解“资产负债表为时点数，利润表为时期数”等会计准则，其构建的 CGP_TTM（现金毛利）等因子，在财务逻辑上远比遗传规划等方法随机组合出的表达式更为严谨和有效。在高频领域，AI 强大的代码生成能力是其另一大突出优势。如报告中 momentum_acceleration_corr 因子的构建，AI 能够直接编写包含 diff()、groupby() 和动态相关性计算的复杂 Python 代码，彻底摆脱了传统方法对少数预定义算子（如 Sum, Std, Corr）的依赖，极大地拓宽了因子构建的自由度和复杂度，使其能够捕捉更为精细的市场动态。

高效的“智能探索”而非“暴力搜索”：大语言模型具备强大的“少样本学习”（Few-shot Learning）能力。如本报告第二章所示，仅需提供少量优质的“样例因子”，AI 便能迅速掌握其内在范式并举一反三，生成大量逻辑相似但细节各异的新因子。这一过程更像是站在前人经验基础上的“智能探索”，而非无方向的“暴力搜索”。因此，相较于需要成千上万次迭代的遗传规划，AI 往往能在数十次迭代内就找到表现优异的因子，显著提升了研究效率。

二、 当前局限与待解挑战

尽管 AI 展现出巨大潜力，但当前阶段的方法论仍存在一些局限性，需要在未来的研究中正视并寻求突破：

搜索域受限于人类先验知识： 这是该范式最大的局限所在。AI 的“智能”源于其对人类已有知识的学习和模仿，这意味着它更擅长在现有的认知框架内进行优化、组合与拓展。它或许能想到用“现金毛利”替代“营业利润”，但很难凭空创造一个完全脱离人类金融学理论的全新维度。如报告中对 beta20、min20 等逻辑相对固化的因子优化效果不佳，也反映了 AI 在被初始逻辑“锚定”后，难以跳出框架。相比之下，纯数据驱动的方法虽然可能产生大量无意义的组合，但理论上具备在全局范围内找到某种“反直觉”但有效的全新因子的可能性。

大模型“幻觉”与结果的稳定性挑战： 大语言模型偶尔会产生“幻觉”，即生成不存在的函数、错误的参数或不符合事实的逻辑描述。这要求我们在实践中必须构建一套严格的“提示工程-代码生成-回测验证-循环反馈”的闭环工作流，通过程序化的验证来过滤掉 AI 的错误输出。同时，模型的输出对提示词（Prompt）的细微变化较为敏感，如何设计一套稳定、高效且能最大化激发模型能力的提示词模板，本身就是一项充满挑战的核心工作。

计算成本与 API 依赖： 虽然迭代次数少，但调用顶级大模型（如本报告使用的 Gemini 2.5 Pro）的单次成本相对较高，尤其是在处理大规模、长周期的历史数据时，token 消耗和计算时间会成为显著的成本中心。此外，该研究框架高度依赖于第三方提供的 API 服务，其稳定性、可用性和未来的成本变化，都可能对研究的可持续性构成影响。

7. 风险提示

- 1) 结论基于历史数据，在市场环境转变时模型存在失效的风险。本报告所有结论均基于历史数据的回测分析，历史规律不代表未来，当市场风格、投资者结构、宏观环境等发生重大变化时，报告中构建的因子及模型可能失效。
- 2) 模型过拟合风险。本报告构建的因子及深度学习模型，可能对历史数据存在过度拟合的问题，即模型可能学习到了历史数据中的噪音而非真实的规律，从而在未来市场的实际应用中表现不佳。
- 3) 大语言模型（LLM）自身局限性的风险。本报告的研究依赖于特定版本的大语言模型，且大语言模型的输出具有随机性，同时可能存在模型幻觉问题，导致报告结果无法复现。
- 4) 回测设置与实际存在差异的风险。报告中的回测分析基于一系列理想化的假设（如特定的成交价格、固定的交易费用、未考虑市场冲击等），实际交易环境中的成本、滑点和各种限制可能导致策略的真实表现不及预期。
- 5) 因子失效风险。报告中通过 AI 挖掘的因子，其有效性可能随着时间的推移而衰减。特别是在因子逻辑被市场广泛认知或被大量资金使用后，其获取超额收益的能力可能会显著下降。

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准,已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司(以下简称“本公司”)的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下,本报告中的信息或所表述的意见并不构成对任何人的投资建议,本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下,东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易,还可能为这些公司提供投资银行服务或其他服务。

市场有风险,投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息,本公司力求但不保证这些信息的准确性和完整性,也不保证文中观点或陈述不会发生任何变更,在不同时期,本公司可发出与本报告所载资料、意见及推测不一致的报告。

本报告的版权归本公司所有,未经书面许可,任何机构和个人不得以任何形式翻版、复制和发布。经授权刊载、转发本报告或者摘要的,应当注明出处为东吴证券研究所,并注明本报告发布人和发布日期,提示使用本报告的风险,且不得对本报告进行有悖原意的引用、删节和修改。未经授权或未按要求刊载、转发本报告的,应当承担相应的法律责任。本公司将保留向其追究法律责任的权利。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期(A 股市场基准为沪深 300 指数,香港市场基准为恒生指数,美国市场基准为标普 500 指数,新三板基准指数为三板成指(针对协议转让标的)或三板做市指数(针对做市转让标的),北交所基准指数为北证 50 指数),具体如下:

公司投资评级:

买入:预期未来 6 个月个股涨跌幅相对基准在 15% 以上;

增持:预期未来 6 个月个股涨跌幅相对基准介于 5% 与 15% 之间;

中性:预期未来 6 个月个股涨跌幅相对基准介于-5% 与 5% 之间;

减持:预期未来 6 个月个股涨跌幅相对基准介于-15% 与 -5% 之间;

卖出:预期未来 6 个月个股涨跌幅相对基准在-15% 以下。

行业投资评级:

增持: 预期未来 6 个月内,行业指数相对强于基准 5% 以上;

中性: 预期未来 6 个月内,行业指数相对基准-5% 与 5%;

减持: 预期未来 6 个月内,行业指数相对弱于基准 5% 以上。

我们在此提醒您,不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系,表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况,如具体投资目的、财务状况以及特定需求等,并完整理解和使用本报告内容,不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码: 215021

传真: (0512) 62938527

公司网址: <http://www.dwzq.com.cn>