FaVOR: LLM-Based Agentic Framework for Factor Mining via Empirical Validation

EMNLP2026 BUDAPEST

FINDINGS · EMNLP 2026

Code & prompts
github.com/damilab/FaVOR

Hyeonjin Kim[1*], Minseok Kim[1*], Seunghyeon Jung[1*], Sujin Pyo[2], Huisu Jang[3], Woojin Lee[1†]

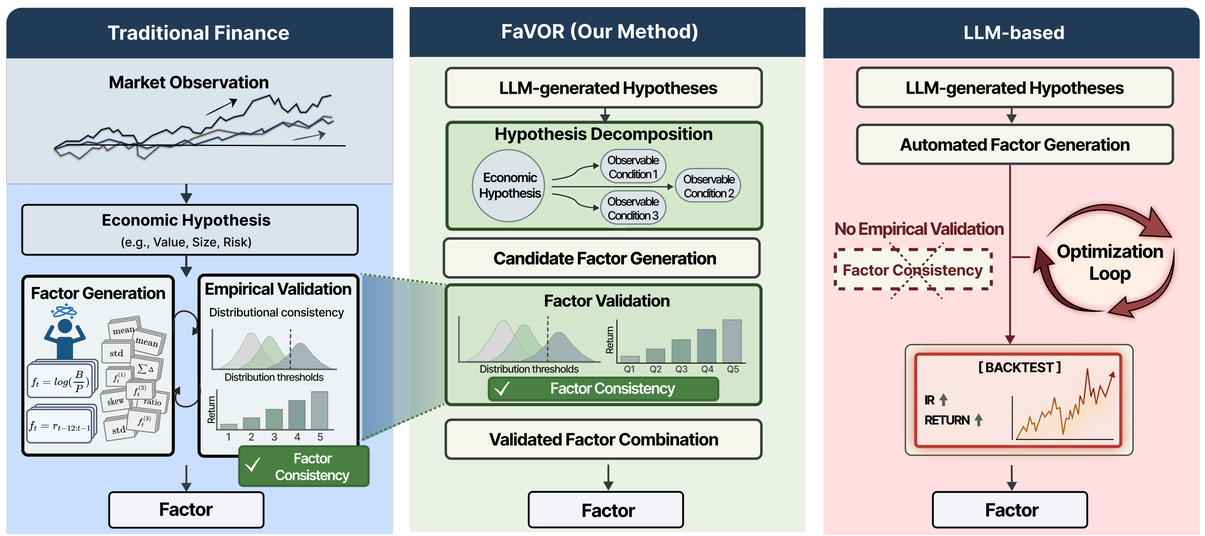
[1]Dept. of Computer Science and AI, Dongguk University-Seoul [2]Seoul National University [3]Soongsil University

[*]Equal contribution [†]Corresponding author

❶ Motivation

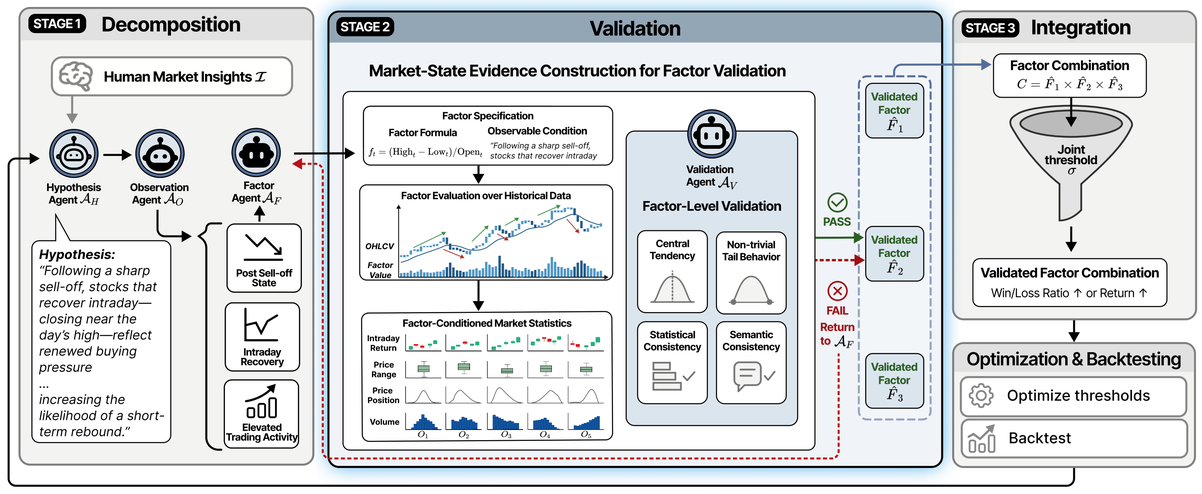
LLM-based multi-agent systems scale factor mining far beyond manual effort, but they **optimize directly for returns** and rarely check whether a generated factor still **expresses the economic hypothesis** that motivated it. Such factors blur real signals and spurious correlations and **break down across regime shifts**.

> **Does the factor measure the market state its hypothesis describes — before we look at returns?**



Traditional design is interpretable but manual; LLM-based mining scales but skips validation. **FaVOR validates every factor before backtesting**, keeping both.

❷ FaVOR framework



Example. The insight "stocks that recover intraday after a sharp sell-off tend to rebound" becomes three conditions — post sell-off state, intraday recovery, elevated trading activity. Each gets candidate formulas, the validation agent tests them one by one, and only the survivors are combined into a trading signal.

❸ Method: agents and their roles

$\mathcal{A}_H$ Hypothesis agent
Reads a human market insight and writes a concrete **economic hypothesis**: which market state occurs, why it happens, and how prices should react.

$\mathcal{A}_O$ Observation agent
Breaks the hypothesis into **independent conditions visible in daily price and volume**, e.g. a post sell-off state, intraday recovery, elevated trading activity.

$\mathcal{A}_F$ Factor agent
Writes several **candidate formulas** for each condition, built only from a fixed operator library (rank, TS_MIN, TS_ZSCORE, …) so every factor stays readable and executable.

$\mathcal{A}_V$ Validation agent
Sorts stocks into five bins by the candidate's value and looks at how the **same-day market state** (intraday return, daily range, close position, volume) changes from the lowest to the highest bin. A factor passes only if that state shifts **consistently and in the direction its condition predicts**:

i the center shifts across bins **ii** the tails change **iii** ≥2 statistics agree **iv** the meaning matches

Future returns are never used; failed factors go back to the factor agent.

σ Integration
Takes one validated factor per condition and trades only when **all of them fire together**. The entry threshold is tightened step by step, and a combination is kept only if stricter entries bring higher returns or win rates. Thresholds are tuned on 2024, then frozen for the 2025 test.

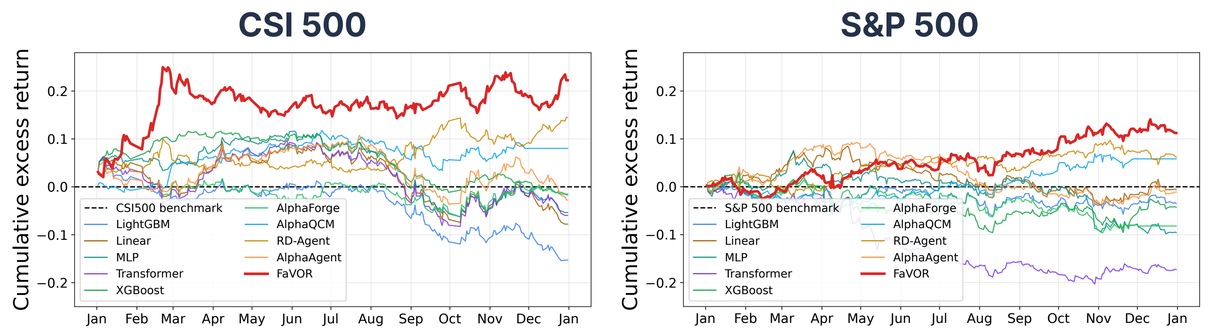
Why an LLM validator? The same statistical pattern can support one rationale while contradicting another, so it must be judged against the hypothesis.

Outer loop. A memory of past hypotheses and their average IR guides the next round. All agents run on GPT-4o by default.

❹ Experiments and main results

CSI 500 and S&P 500 (daily OHLCV, Qlib) · train 2022–23, validation 2024, **test 2025** · excess returns after costs · **bold** best, underline 2nd

	CSI 500				S&P 500			
Method	**AR**	**IR**	**MDD**	**CR**	**AR**	**IR**	**MDD**	**CR**
Linear	−0.0702	−0.5008	−0.1649	−0.0778	−0.0010	−0.0104	−0.1403	−0.0055
XGBoost	−0.0123	−0.1642	−0.1370	−0.0152	−0.0357	−0.3345	−0.2326	−0.0426
LightGBM	−0.1576	−1.3954	−0.2173	−0.1530	−0.0302	−0.3389	−0.2154	−0.0353
MLP	−0.0493	−0.4812	−0.1725	−0.0537	−0.0895	−0.8041	−0.1686	−0.0956
Transformer	−0.0521	−0.3943	−0.1760	−0.0597	−0.1744	−1.5509	−0.1860	−0.1729
AlphaForge	−0.0133	−0.1255	−0.1260	−0.0167	−0.0833	−0.8617	−0.2093	−0.0819
AlphaQCM	0.0869	1.0604	−0.1396	0.0803	0.0619	0.8250	−0.2345	0.0583
RD-Agent	0.1382	1.3360	−0.1427	0.1452	0.0604	0.8028	−0.1704	0.0624
AlphaAgent	−0.0192	−0.1490	−0.1369	−0.0275	−0.0060	−0.0631	−0.1428	−0.0110
FaVOR	**0.2067**	**1.5295**	**−0.0853**	**0.2225**	**0.1062**	**1.1315**	**−0.0443**	**0.1123**



FaVOR is **best on all four metrics in both markets**.

❺ Case study: factor-level validation

> **Hypothesis:** After a sharp sell-off, stocks that exhibit intraday recovery tend to enter early stabilization and rebound short-term.

Candidate factor	Distributional evidence	
O1 Post sell-off state		
`TS_ZSCORE(C,10)`	DIR mean and median decrease (0.14 → −0.11); tail statistics move down together	PASS
`TS_ZSCORE(V,10) * TS_ZSCORE(DELTA(C,1),10)`	upper-tail expansion with kurtosis compression: internally conflicting tail signals	FAIL
O2 Intraday recovery		
`(close - TS_MIN(low,3)) / (high - TS_MIN(low,3) + eps)`	POS mean and median increase monotonically (0.083 → 0.914)	PASS
`-CORR(rank(DELTA(logV)), rank(DIR))`	evidence decreases across bins, contradicting its stated polarity	FAIL
O3 Early stabilization		
`(TS_MIN(low,3) - low) / (high - low + eps)`	MAG mean and median decrease (−21%, −18%); tail evidence compresses	PASS
`1 / (eps + TS_STD(C,20))`	kurtosis drops (−28.4%) while q90 barely moves (+1.8%): inconsistent tails	FAIL

❻ Ablation (CSI 500, 2025)

Setting	AR	IR	MDD
FaVOR	**0.2067**	**1.5295**	**−0.0853**
w/o Stage 2	0.0492	0.1526	−0.2091
w/o Stage 3	−0.3902	−1.2748	−0.4260
w/o Stage 2 & Stage 3	−0.2793	−1.4316	−0.3087

No Stage 2: **MDD more than doubles**. Neither stage: all metrics negative.

❼ Sensitivity to LLM backbone

	CSI 500				S&P 500			
Backbone	**AR**	**IR**	**MDD**	**CR**	**AR**	**IR**	**MDD**	**CR**
GPT-4o (default)	**0.2067**	**1.5295**	−0.0853	**0.2225**	**0.1062**	**1.1315**	**−0.0443**	**0.1123**
GPT-5.4-mini	0.1039	1.3246	**−0.0443**	0.1079	0.0299	0.2553	−0.1186	0.0244
Gemini-2.5-Flash	0.0732	0.4998	−0.1418	0.0657	0.0808	0.7039	−0.0621	0.0807
Claude-Sonnet-4.6	0.0541	0.5364	−0.0878	0.0511	0.0072	0.0817	−0.0898	0.0035
Llama-3.3-70B	0.0449	0.4183	−0.1313	0.0407	0.0146	0.1545	−0.0801	0.0107
Qwen3-235B	0.0948	1.0598	−0.0936	0.0967	0.0174	0.1670	−0.1033	0.0126

AR and IR stay **positive for every backbone**.

Supported in part by NRF grant (RS-2025-00556289) and IITP grants (IITP-2026-RS-2020-II201789, IITP-2026-RS-2023-00254592) funded by the Korea government (MSIT).

EMNLP2026 BUDAPEST