From Knowing to Doing: A Memory-Controlled Benchmark for LLM Trading Agents on Stock Markets

Anonymous ACL submission

Abstract

Evaluating whether large language model (LLM) agents can profit in capital markets is increasingly framed as end-to-end trading: place an agent in a historical market, let it trade, and measure portfolio returns. This setup is vulnerable to two evaluation failures. First, long backtests often overlap with the knowledge cutoffs of frontier LLMs, allowing memorized tickers, dates, prices, and market narratives to substitute for investment reasoning. Second, raw returns are a noisy proxy for stock-selection ability, since positive performance may come from market beta, style exposure, or favorable regimes rather than genuine alpha.

We introduce KTD-FIN (**K**nowing-**T**o-**D**oing **Fin**ancial Benchmark), an end-to-end stock-market trading benchmark that addresses both issues. KTD-FIN uses a data-side masking protocol to anonymize key identifiers and calendar information consistently across prompts and tools, separating historical market memory from investment decision-making. It also incorporates a Barra-style performance attribution framework that decomposes portfolio returns into market, style, and stock-selection alpha components.

Across ten frontier LLM agents evaluated on the Chinese A-share CSI300 over a 2024–2026 trading window, masking substantially changes agent rationales, pushing them away from company-specific narratives and toward anonymized factor-based reasoning. Attribution analysis further shows that LLM agents' cumulative returns under leakage-controlled evaluation are largely explained by passive market and style exposure, with limited evidence of persistent stock-selection alpha. These findings suggest that financial LLM benchmarks should evaluate not only whether an agent makes money, but also whether the source of returns reflects transferable investment skill. We release KTD-FIN as a reproducible template for leakage-controlled and attribution-aware evaluation of LLM trading agents.

1 Introduction

A growing number of recent systems treat the LLM not as a question-answerer about finance but as a trading agent: something that watches the market, decides what to look at next, and commits capital (Yu et al., 2023; Zhang et al., 2024; Li et al., 2024a). The benchmarks built around this shift fall into two groups. One keeps the model outside the market and tests how well it can read filings, answer questions, or extract structured information (Islam et al., 2023; Guo et al., 2023; Xie et al., 2023, 2024). The other puts the model inside a historical simulator, lets it trade, and reports portfolio return. As this end-to-end framing becomes the standard way to ask whether an LLM can "profit" in capital markets, two evaluation failures emerge. First, long backtests routinely overlap with the knowledge cutoffs of frontier models, so memorized tickers, prices, dates, and macro narratives can substitute for the data the benchmark thinks it is showing. Second, even when leakage is controlled, cumulative return is a noisy proxy for stock-selection skill: positive performance may simply reflect market beta, persistent style exposure, or a favorable regime rather than anything the agent has decided.

Take the first failure concretely. Real tickers and dates are passed to the model verbatim, and pretraining corpora are full of price commentary, brand descriptions, and macro coverage (Li et al., 2025; Ma, 2026). In a sanity check on our own environment—where the agent sees only OHLCV-derived numbers—a representative model writes "SH601857—*defensive blue-chip with strong downside protection, increase position*" when tickers are visible, but declines to trade in all 25 of 25 cells when those tickers are anonymized and the same numbers are fed. The signal driving the first behaviour is the ticker, not the data. Existing fixes are partial: DeepFund (Li et al., 2025) restricts evaluation to post-cutoff dates, which shrinks the win-

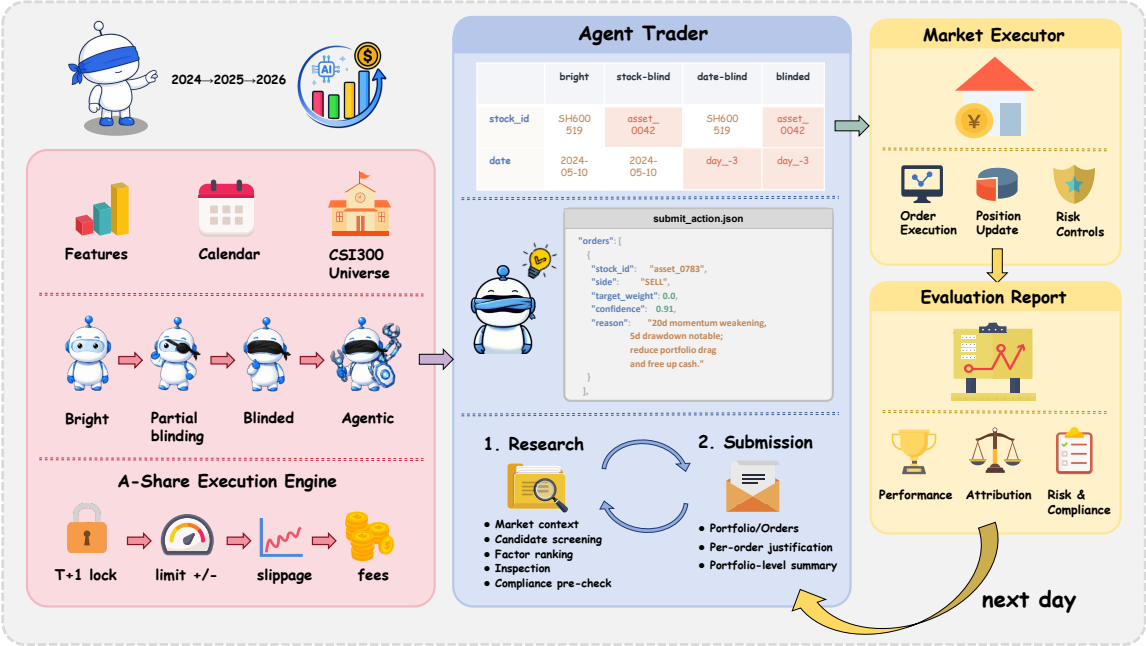

Figure 1: Overview of KTD-FIN. An LLM agent receives CSI300 features under one of four data-side masking conditions (*bright*, *stock-blind*, *date-blind*, *blinded*), executes a *research*-then-*submission* loop, and submits a JSON action that is run by a Qlib-based stock-market executor with realistic constraints.

dow with every model release and prevents shared windows across models; OracleProto (Ma, 2026) shows in a static forecasting setting that prompt-level "do not use external memory" instructions fail, motivating data-level mitigation, which we extend from their static setting to sequential trading.

The second failure is subtler. A leakage-controlled backtest still collapses to a single number—cumulative return—that lumps market drift, persistent style tilts, and genuine stock-specific skill into one figure. Empirical asset pricing handles this by decomposing returns into market, style, and a stock-specific residual; without that decomposition, a financial LLM benchmark cannot distinguish "the agent made money" from "the agent rode a factor any similarly-exposed portfolio would have ridden."

KTD-FIN (**K**nowing-**T**o-**D**oing **Fin**ancial Benchmark; Figure 1) is a Qlib-based stock-market trading benchmark—instantiated on the CSI300 universe of the Chinese A-share market—designed to address both failures inside the protocol. For leakage, a four-level data-side *masking protocol*—*bright*, *stock-blind*, *date-blind*, *blinded*—anonymizes tickers, calendar dates, and tool-return timestamps in coordination, so an agent in *blinded* never sees a real identifier even transiently through a tool result; an independent ten-attacker probe (Section 4.3) certifies the mask, with top-1 ticker recovery at most 3.0% and a strict joint (ticker top-5 *and* date within ± 7 trading days) rate at most 1.5%. For the return-versus-skill conflation, KTD-FIN reports a Barra-style daily cross-sectional attribution that decomposes each agent's realized return into market, style, and stock-selection alpha components. Three decision modes—*memory-only*, *fixed-candidate*, and *open-research*—vary memory exposure and information access; reporting is consolidated into a ten-dimensional metric panel covering returns, behavior, and reliability.

The headline evaluation runs ten frontier LLMs in the cleanest *blinded* $\times$ *open-research* setting over a single contiguous 2024-01-01 to 2026-04-10 window, against eighteen Alpha9-trained machine-learning baselines (trained on 2015–2022, validated on 2023) and CSI300 buy-and-hold. Two findings emerge that map directly onto the two failures we set out to address. (i) Masking changes the *shape* of trading, not just the score: under *bright*, agent rationales are brand-driven ("defensive blue-chip," "liquor leader") and overweight memorized index heavyweights; under *blinded*, the same model switches to factor-rank reasoning over anonymized IDs. (ii) Cumulative return is not the same as stock-selection skill: the Barra attribution shows that LLM agents' headline returns are largely explained by passive exposure to market

and style factors, with only one model at near-zero stock-selection alpha and the rest measurably negative.

Contributions. (1) KTD-FIN, an end-to-end stock-market trading benchmark that handles both failure modes inside the protocol: leakage via a four-level data-side masking protocol (certified by a ten-attacker de-anonymization probe) and return-versus-skill via a Barra-style cross-sectional attribution. (2) A headline ten-LLM evaluation under *blinded* × *open-research* against eighteen Alpha9-trained baselines and CSI300 buy-and-hold, scored on a ten-dimensional metric panel and re-scored by Barra-attributed selection alpha. (3) Public release of environment, scripts, masking and probe protocols, and the full evaluation dataset.

2 Related Work

Benchmarking LLMs in Financial Domain. With the rapid growth of LLMs in finance, a substantial body of benchmarks has been developed to evaluate model capabilities in financial contexts (Islam et al., 2023; Guo et al., 2023; Xie et al., 2023, 2024). Financial LLM benchmarks have evolved from document-understanding frameworks such as FinanceBench (Islam et al., 2023), PIXIU/FLARE (Xie et al., 2023), FinBen (Xie et al., 2024), FinQA (Chen et al., 2021), ConvFinQA (Chen et al., 2022), TAT-QA (Zhu et al., 2021), and the Chinese-language CFLUE (Zhu et al., 2024) and FinEval (Guo et al., 2023), to trading-agent evaluations such as FinMem (Yu et al., 2023), FinAgent (Zhang et al., 2024), InvestorBench (Li et al., 2024a), and the multi-agent designs of TradingGPT (Li et al., 2023), FinCon (Yu et al., 2024), and CryptoTrade (Li et al., 2024b), with open-source LLM and platform releases by FinGPT (Yang et al., 2023) and FinRobot (Yang et al., 2024). These benchmarks share a focus on financial knowledge and trading decisions, but typically pass real tickers and trading calendars to the model, which conflates a model's reasoning with priors memorized during pretraining and leaves no controlled way to separate the two. Our work introduces a four-level data-side masking protocol that turns memory exposure into a controllable axis without depending on each model's training cutoff.

Contamination-aware Benchmarking. There has been growing interest in contamination-controlled evaluation. General-domain efforts include LiveCodeBench (Jain et al., 2024) for rolling post-cutoff problem collection and Min-K% Prob (Shi et al., 2024) for membership inference on pretraining data. In the financial domain, DeepFund (Li et al., 2025) evaluates each model only on dates after its own training cutoff, and OracleProto (Ma, 2026) controls exposure at the data layer for static forecasting and shows that prompt-level restrictions cannot substitute for it. These approaches either shrink the available evaluation window with every new model release or remain limited to single-step forecasting tasks. Our work is the first sequential-decision financial benchmark to control pretraining-memory exposure inside the protocol, with the masking independently certified by a ten-attacker de-anonymization probe.

Agent and infrastructure foundations. The agent loop in KTD-FIN builds on ReAct (Yao et al., 2023), with related tool-use methods in Toolformer (Schick et al., 2023), ToolLLM (Qin et al., 2024), and HuggingGPT (Shen et al., 2023), and general agent evaluation in AgentBench (Liu et al., 2024) and τ-bench (Yao et al., 2024). The execution environment follows Qlib conventions (Yang et al., 2020); FinRL-Meta (Liu et al., 2022) and TradeMaster (Sun et al., 2023) provide adjacent RL-based trading platforms. Our gradient-boosted factor baseline uses LightGBM (Ke et al., 2017).

3 Benchmark Design

KTD-FIN evaluates LLMs as stock-market trading agents under tight control of pretraining-memory exposure (Figure 1). It has four interlocking components—a four-level masking protocol (§3.3), three decision modes (§3.2), a Qlib-based execution engine (§3.1), and a ten-dimensional metric panel (§3.6)—each described below.

3.1 Execution engine

The task is daily portfolio construction over the CSI300 constituent universe. At each trading step the agent observes account state, current positions, last-step execution results, benchmark-relative performance, and summary constraints, and must finally emit a JSON action. Observations are *price-only*: OHLCV histories and derived technical features. The benchmark contains no news, analyst reports, regulatory filings, or fundamentals—any reasoning that names brand, sector, or fundamentals must therefore reflect pretraining priors rather than observation.

The execution engine simulates A-share trading at realistic granularity: long-only positions starting from CNY 1,000,000; T+1 sell availability; CNY 5 minimum cost per order; 5 bps buy cost and 15 bps sell cost. Orders execute at the next-day open price. Daily price-limit filtering excludes limit-up stocks from BUY candidates and limit-down stocks from SELL candidates, using sector-differentiated thresholds that match the A-share board rules (main board $\pm 10\%$, ChiNext/STAR $\pm 20\%$, BSE $\pm 30\%$; full board-code mapping in Appendix H). The simulator logs all violations and execution outcomes so invalid LLM actions remain visible rather than being silently corrected.

3.2 Three decision modes

KTD-FIN decomposes LLM trading competence along three orthogonal capability dimensions, each instantiated as an independent decision mode (Table 1).

Mode	Capability probed	Non-prior information
memory-only	prior-memory dependence	stock IDs only
fixed-candidate	clean-signal selection	curated factor table
open-research	end-to-end agentic reasoning	six retrieval tools

Table 1: Three decision modes, one per capability dimension. All modes share identical execution constraints, masking, and account state.

The memory-only mode carries only stock identifiers and account state, so any non-trivial trade must come from pretraining priors over the listed IDs; fixed-candidate hands the agent a curated OHLCV+factor table over a small factor-top-k union (the typical prior-work setting, probing structured-signal selection); open-research exposes a six-tool retrieval interface that the agent invokes freely (Section 3.4), the fully autonomous agent setting.

3.3 Four-level masking protocol

The masking protocol is the data-side instrument by which KTD-FIN controls pretraining-memory exposure. It operates at four levels: *bright* (real ticker, real date), *stock-blind* (anonymized ticker, real date), *date-blind* (real ticker, relative day index), and *blinded* (both anonymized). Figure 1 shows a concrete alias example. The two intermediate levels jointly form the partial blinding stratum that isolates ticker-specific versus calendar-specific priors. Aliases are stable within an episode and shuffled across episodes to prevent within-run identity reconstruction.

The benchmark harness preserves end-to-end masking consistency. An episode-level alias map translates real tickers and dates to aliases when constructing the prompt; tool arguments are unmasked before querying the data layer and remasked before being returned to the model. Therefore an agent in blinded mode never observes a real ticker or a real calendar date, even transiently through a tool result. All numeric features the model receives—returns, volatilities, drawdowns—are computed on the original time series so the underlying signal is preserved; only the identifying handles change.

Mask validity. An independent ten-attacker deanonymization probe (Section 4.3) finds joint ticker+date recovery at most 1.5%, confirming that residual leakage through board, sector, or volatility-profile features is negligible; in what follows we treat blinded as a near-clean reasoning condition.

3.4 Tool-mediated open research

The open-research decision mode operates as a ReAct-style (Yao et al., 2023) loop with a strict two-phase structure. In a research phase, the model issues calls—as many as it deems necessary—to six read-only tools covering market context, candidate screening, per-stock snapshots, pairwise comparison, portfolio state, and pre-trade risk checks (full interface in Appendix B, Table 6). The initial prompt contains only account state and constraints, so the model must actively retrieve any market information through tools. In a submission phase, the model is forced via `tool_choice` to call `submit_action` and emit a JSON order list, with up to three retries on schema or constraint violations and a fallback to no-trade if all retries fail. Failure modes during the research phase (parse errors) inject structured feedback rather than aborting the run. The harness is provider-neutral: provider-specific tool-call wire formats (OpenAI-compatible, Anthropic-native) are normalized inside per-provider adapters.

3.5 Action schema

The final action is a JSON object with an `orders` list and an `overall_reason` string. Each order contains `stock_id`, `side` (BUY or SELL),

confidence in $[0, 1]$, a short reason, and exactly one of target_weight or shares. The executor rejects orders that violate universe membership, position limits, T+1, or formatting constraints; rejected orders are logged but do not abort the step. The confidence field feeds the calibration metrics; the reason field is used in our motivation analysis (Section 1).

3.6 Ten-dimensional metric panel

A core thesis of KTD-FIN is that LLM trading agents must be judged not only on what they earn but on how they earn it. We report a ten-dimensional metric panel grouped into three families: returns and risk (total return, Sharpe, max drawdown, information ratio); behavior (annualized turnover, Herfindahl–Hirschman index, cash ratio, abstention rate); and reliability and calibration (parse-failure rate, expected calibration error). Two further quantities—tool-validity rate and the Brier score on per-order confidence—are recorded for diagnostic use but not part of the headline panel. Full definitions are in Appendix G.

4 Experiments

4.1 Setup

Evaluation windows. We evaluate on ten regime-stratified short windows of ∼43 trading days each (W1–W10, Table 2) plus one contiguous long window from 2024-01-01 to 2026-04-10 (548 trading days). Only W10 falls strictly post-cutoff for the anchor model; the other nine rely on the masking protocol (Section 3.3) for contamination control. The long window drives the headline cross-model comparison; the short windows drive per-regime analysis on the anchor.

Code	Dates	CSI300	Regime
W1	2024-04-01–06-05	-0.02%	near-flat
W2	2024-05-27–07-25	-6.50%	monotone bear
W3	2024-07-25–09-25	$+0.07\%$	near-flat
W4	2024-09-09–11-15	$+24.30\%$	explosive bull
W5	2024-10-21–12-13	-0.05%	near-flat
W6	2024-12-06–2025-02-14	-0.86%	mild decline
W7	2025-03-03–05-09	-1.09%	V-shape
W8	2025-07-14–09-19	$+12.05\%$	smooth uptrend
W9	2025-10-13–12-19	-0.56%	near-flat
W10†	2026-02-02–04-10	$+0.66\%$	mild rise

Table 2: Ten regime-stratified evaluation windows. †W10 is strictly post-cutoff for the anchor model.

Models. The panel separates depth from breadth. The anchor Step-3.5-Flash sweeps the full mask × mode grid (4 masks × 3 decision modes × W1–W10 × 5 seeds, with informative-cell pruning) to probe mask, mode, and seed effects. Ten frontier LLMs then run the headline comparison under *blinded* × *open-research* on the long window—one model each from Step, OpenAI, Anthropic, Google, ByteDance, DeepSeek, Alibaba, MiniMax, Zhipu, and Moonshot (full identifiers in Table 4). All ten share identical prompts, tool interfaces, masking, and execution rules (Section 3), so cross-model gaps reflect model differences alone. Each is run with multiple seeds; cells report the median.

Baselines. We benchmark the ten LLM agents against eighteen trained Qlib factor models on the long window. The baselines share the nine VIF-screened Alpha9 features (Table 7), are trained on 2008–2022 with 2023 as validation, and execute through the same trading layer as the LLM agents. They span linear, gradient-boosted trees, tabular networks, RNN/Transformer time-series models, and specialized time-series architectures (full model list and Qlib default hyperparameters in Appendix H). CSI300 buy-and-hold is reported as the market reference.

Protocol. Multi-seed cells report median total return with $[Q_1, Q_3]$ as IQR; cross-condition contrasts use Wilcoxon signed-rank tests on paired (window, seed) medians at $\alpha = 0.05$. Mask validity is verified independently (Section 4.3). The harness is deterministic given a seed, so any model is re-evaluable on identical conditions. LLM agents run via vendor APIs; the eighteen Qlib models train on one H800 (5 min to $\sim$ 4 h). One LLM on the long window with three seeds takes 6–12 h wall-clock.

4.2 Anchor depth scan

On the anchor Step-3.5-Flash we sweep the four masks under each of the three decision modes (Table 1) and additionally collect order-level rationales under *bright* to characterize how the model talks when real identifiers are visible. Table 3 summarizes the resulting mode × mask profile on the anchor depth set. Two patterns emerge: pretraining memory of real tickers alone is sufficient to drive active trading even when no observation channel is present—under *memory-only*, *bright* produces non-trivial orders while *blinded* produces voluntary cash holding at exactly 0.00%; and granting

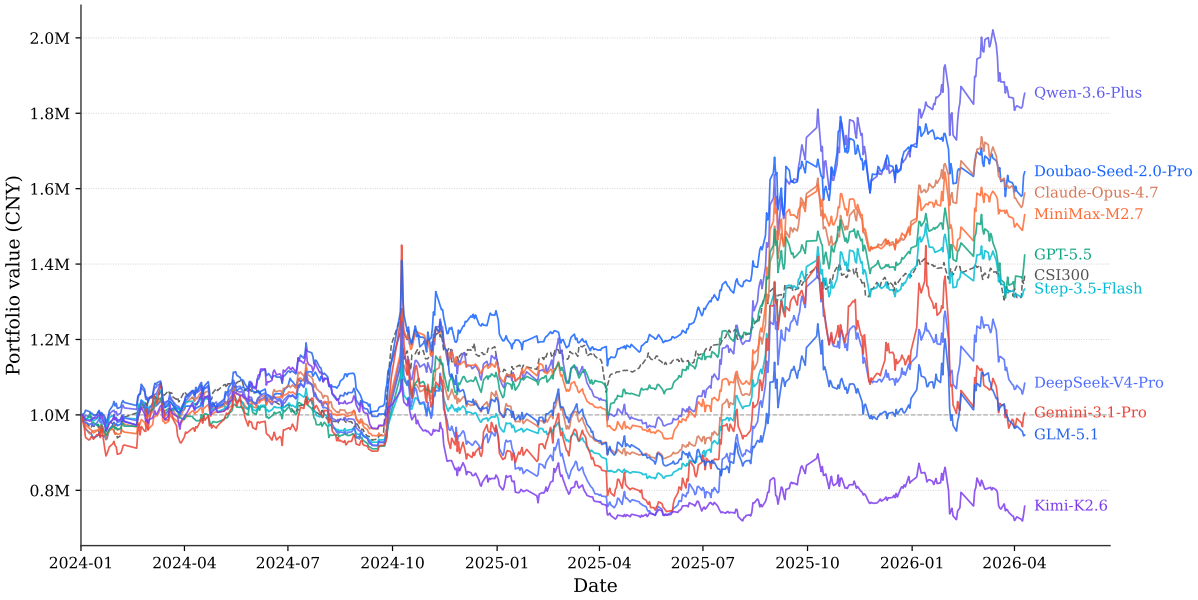

Figure 2: Seed-averaged daily portfolio equity curves of the ten LLM agents and CSI300 buy-and-hold (dashed grey) under *blinded* × *open-research* on the long 2024-01-01 to 2026-04-10 window, starting from CNY 1,000,000.

the agent tool access (*open-research*) consistently lifts capability above the no-tool *fixed-candidate* baseline at both mask levels.

Mode	bright	blinded
open-research	$-\mathbf{2.11}$	$-\mathbf{1.83}$
fixed-candidate	-3.61	-3.76
memory-only	-0.16	0.00

Table 3: Median total return (%) on the anchor depth set × 5 seeds by mode and mask. Bold: best per column.

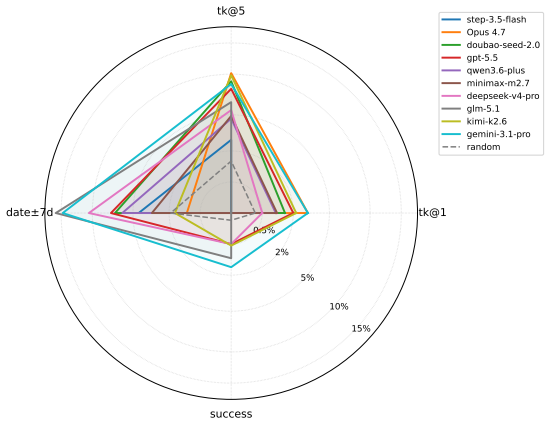

Figure 3: Per-attacker recovery under *blinded* on tk@1, tk@5, date ±7d, and joint success (sqrt-scaled). Grey dashed: random baseline.

4.3 De-anonymization probe

To certify the masking protocol independently of the trading benchmark, we run a de-anonymization probe in which the exact masked payload that the agent sees in *blinded* mode is fed to ten frontier LLMs acting as attackers. Each attacker is asked to recover the ticker (top-5 list), date, sector, and board for 200 (stock, date) probes stratified across five time windows and four boards; we score top-1 / top-5 ticker accuracy, date-within-±7 trading-day rate, and a strict joint success rate (ticker top-5 *and* date within ±7 trading days). Figure 3 shows the per-attacker recovery profile; the strongest attacker recovers only 10.2% ticker top-5 and joint success peaks at 1.5%, both within a small additive margin of the random baseline—confirming that *blinded* effectively cuts the model off from its pretraining memory of the underlying identifiers.

4.4 Cross-model leaderboard

Under *blinded* × *open-research* on the long 2024-01-01 to 2026-04-10 window, Figure 2 shows the seed-averaged daily equity curves of the ten LLM agents against CSI300; Table 4 reports the ten-dimensional metric panel for the LLMs alongside the eighteen supervised Qlib factor models, with CSI300 buy-and-hold inserted into each panel at its return-implied position.

4.5 Barra-style return attribution

Raw portfolio returns can flatter an agent that merely loads on a winning style: if high-volatility names rally during the window, a volatility-tilted

Model	Total ret.	Excess	Sharpe	MDD	IR	Turn.	HHI	ECE	α-rk
LLM agents, blinded × open-research, long window 2024-01-01 to 2026-04-10									
`qwen3.6-plus`	+85.29%	+48.37%	1.13	−24.49%	+0.48	31.88	0.005	0.367	2
`gpt-5.5`	+61.26%	+24.34%	0.98	−24.63%	+0.29	14.38	0.037	0.272	6
`doubao-seed-2.0`	+61.01%	+24.09%	1.15	−15.62%	+0.29	14.18	0.004	0.425	3
`claude-opus-4-7`	+58.80%	+21.88%	0.98	−26.69%	+0.26	31.75	0.012	0.246	**1**
`minimax-m2.7`	+56.31%	+19.39%	0.89	−29.43%	+0.25	43.68	0.029	0.303	4
CSI300 buy-and-hold	+36.92%	0.00%	—	—	—	0.00	—	—	—
`step-3.5-flash`	+21.86%	−15.06%	0.49	−28.99%	−0.13	24.66	0.012	0.357	5
`gemini-3.1-pro`	+8.25%	−28.67%	0.29	−49.67%	−0.11	101.23	0.066	0.387	7
`deepseek-v4-pro`	+1.39%	−35.53%	0.19	−42.35%	−0.29	73.49	0.026	0.352	8
`glm-5.1`	−3.55%	−40.47%	0.10	−44.88%	−0.40	100.25	0.032	0.322	9
`kimi-k2.6`	−24.23%	−61.15%	−0.34	−45.49%	−0.77	85.26	0.090	0.306	10
Trained Qlib factor models (Alpha9, 2008–2022 train, 2023 val); 8 representatives, full 18 in Appendix H									
`SFM`	+86.58%	+49.66%	2.02	−7.41%	+0.57	22.47	—	—	—
`DoubleEnsemble`	+73.66%	+36.74%	1.41	−18.24%	+0.41	72.84	—	—	—
`CatBoost`	+72.97%	+36.05%	1.43	−16.74%	+0.40	29.96	—	—	—
`LightGBM`	+66.22%	+29.30%	1.29	−15.59%	+0.33	46.18	—	—	—
`Transformer`	+60.35%	+23.43%	1.61	−11.29%	+0.27	11.22	—	—	—
`LSTM`	+43.92%	+7.00%	1.12	−11.76%	+0.06	78.03	—	—	—
CSI300 buy-and-hold	+36.92%	0.00%	—	—	—	0.00	—	—	—
`MLP`	+34.08%	−2.84%	0.69	−21.75%	−0.01	18.48	—	—	—
`Linear`	+29.60%	−7.32%	0.74	−13.88%	−0.12	15.61	—	—	—

Table 4: Headline leaderboard, sorted by total return; the shaded CSI300 row sits at its rank position in each panel. α-rk ranks LLMs by Barra-attributed selection alpha; bold marks the only agent with $\alpha \geq 0$.

portfolio scores high without any firm-level skill. We therefore decompose realized returns into market, style, and a residual that we treat as a closer proxy for stock-selection ability, while still reporting standard return/risk metrics in the headline table.

For each trading day t we estimate a weighted least-squares cross-sectional regression over the N_t investable stocks:

$$r_{i,t} = f_{0,t} + \sum_{k=1}^{K} x_{i,k,t-1}\lambda_{k,t} + \varepsilon_{i,t}, \qquad (1)$$

where $r_{i,t}$ is stock i's close-to-close return, $f_{0,t}$ is a unit-exposure common-return component for the universe, $x_{i,k,t-1}$ is the winsorized and cross-sectionally standardized exposure of stock i to style factor k (computed from information strictly before day t), $\lambda_{k,t}$ is the estimated style-factor return, and $\varepsilon_{i,t}$ is the stock-specific residual. Aggregating with the agent's pre-return weights $w^p_{i,t-1}$ decomposes portfolio return as $R^p_t = f_{0,t} + \sum_k(\sum_i w^p_{i,t-1}x_{i,k,t-1})\lambda_{k,t} + \alpha^p_t$, where the residual term $\alpha^p_t = \sum_i w^p_{i,t-1}\varepsilon_{i,t}$ is the factor-adjusted selection alpha.

We use nine cross-sectional style factors spanning momentum, volatility, liquidity, reversal, skewness, and volume–price interaction, screened on a pre-evaluation calibration window by variance inflation factor $\mathrm{VIF}_k = 1/(1-R^2_k)$ to reduce multicollinearity without using evaluation-period information.

Model	Com.	Sty.	Sel. α	Port
Claude Opus 4.7	+35.0	+16.6	+0.2	+51.8
Qwen3.6-Plus	+41.8	+29.2	−0.7	+70.3
Doubao-Seed-2.0	+42.5	+15.5	−4.7	+53.3
MiniMax-M2.7	+29.6	+26.4	−7.4	+48.6
Step-3.5-Flash	+29.5	+17.5	−13.7	+33.3
GPT-5.5	+40.0	+11.8	−14.9	+36.9
Gemini-3.1-Pro	+38.3	+23.6	−38.5	+23.4
DeepSeek-V4-Pro	+38.5	+23.1	−43.0	+18.6
GLM-5.1	+39.5	+27.5	−61.9	+5.1
Kimi-K2.6	+32.1	+22.0	−77.8	−23.7

Table 5: Seed-averaged Barra attribution (%) on the 2024–2026 window, sorted by selection alpha.

We apply Eq. (1) day-by-day to every completed trajectory and report seed-averaged cumulative contributions in Table 5. Since the intercept in Eq. (1) is a unit-exposure common return component rather than an externally observed market-index return, we denote it as Common. By construction, Common + Style + Selection Alpha = Port. Per-LLM interpretation is deferred to Section 5.2.

5 Discussion

5.1 Masking reduces memory leakage

Order rationales recorded by the anchor model on matched runs make the contamination channel directly visible. Under *bright*, justifications interleave factor inputs with content the input does not carry—brand identity, sector taxonomy, and, most diagnostically, specific historical trading days (e.g. *"the Oct*-18 *limit-up move"* for SMIC, a date unrecoverable from the price-only channel). Under *blinded*, the same model receives identical numbers under anonymized aliases (`asset_0006`, `asset_0013`), and rationales reduce to pure factor ranks (*"*20-*day* $+27.4\%$*, low vol* 0.0255*"*); the narrative layer is removed while the factor layer is preserved.

This qualitative shift coincides with a measurable change in trading behavior. In *memory-only*—where the prompt carries no factor data and the identifier is the only candidate signal—the anchor returns -0.16% under *bright* and exactly 0.00% under *blinded* (Table 3): the ticker-visible model trades on identity priors and loses; the anonymized model holds cash. The disappearing trades are precisely those the rationale analysis attributes to pretraining content, and prompt-level restriction alone does not reproduce the ablation—data-side masking does work that no in-prompt instruction has been shown to replicate. Section 4.3 certifies the mask via an independent ten-attacker probe.

5.2 Cumulative return overstates skill

The Barra attribution in Table 5 concentrates almost all cross-agent variance in factor-adjusted selection alpha. Market contributions span a 13pp range and style contributions a 17pp range—every agent, including the worst performers, harvests $+11\%$ to $+29\%$ from systematic factor tilts—while selection alpha spans $+0.2\%$ to -77.8%, a 78pp range that dwarfs the other two combined. The implied ranking diverges sharply from the cumulative-return one: the total-return top four are Qwen3.6-Plus, GPT-5.5, Doubao-Seed-2.0, and Claude Opus 4.7, but on selection alpha only Claude Opus 4.7 ($+0.2\%$) is at par and nine of ten agents post negative selection alpha. The bottom four (Kimi-K2.6, GLM-5.1, DeepSeek-V4-Pro, Gemini-3.1-Pro) share the same broad factor tilts as the top group (Section 5.3); their failure is not a stylistic miss but a within-factor selection failure—holding factor loading constant, their individual picks underperform what the loading alone would have delivered. A benchmark that uses cumulative return as its sole criterion therefore systematically over-credits agents that converge on the prevailing factor portfolio and under-credits within-factor selection skill.

5.3 Factor Exposure Preferences

For every model we compute the portfolio-weighted, z-scored exposure $\sum_i \omega_{i,t}\, x_{i,k,t-1}$ to each style factor in 4.5, average it across invested days in 2024-01-02–2026-04-10, then average within each cohort. The full nine-factor breakdown is in Appendix F (Table 8); the two cohorts agree on three universal signals—long-horizon momentum, the 52-week-high anchor, and an aversion to illiquid names—and occupy near-mirror regimes on the remaining five. ML models form a classical *low-volatility momentum book*: net short realized volatility, amount variability, and intraday momentum, with 15 of 18 sharing the same sign across all nine factors. LLMs flip those tilts, with intraday momentum ($+1.06\sigma$) as their largest single loading and positive exposures on realized volatility and volume–return correlation. In one line: ML books hold quiet trending large caps; LLM books load on high-volatility, recently moving names.

6 Conclusion

KTD-FIN closes two gaps in how end-to-end LLM trading is evaluated. On the input side, a four-level mask anonymizes tickers, calendar dates, and tool-return timestamps consistently across prompts and tools, with an external attacker audit certifying that the channel is closed. On the output side, a Barra-style cross-sectional regression decomposes each agent's daily return into market, style, and stock-selection alpha, separating skill from passive factor harvesting. Across ten frontier LLM agents on a 2024–2026 window against eighteen Alpha9-trained baselines, masking changes the shape of trading—rationales drop from brand narratives to factor ranks—and the cumulative returns a leaderboard would report come from passive market and style exposure rather than from stock-selection alpha. The same pair—a data-side mask that closes the memory channel and an attribution layer that exposes factor luck—transfers directly to news-driven and fundamentals-based LLM evaluation, and we release KTD-FIN as a template for that broader benchmark family.

Limitations

The current implementation focuses on the CSI300 A-share universe with a price-only observation channel (open, high, low, close, volume, and derived technical factors); it does not cover US or Hong Kong equities, nor multimodal inputs such as archived news, filings, or macroeconomic series. These dimensions could affect agent behaviour but are out of scope for the current framework; we plan to extend market coverage and add a timestamp-aligned multimodal track in future iterations. The full mask $\times$ mode $\times$ window grid is populated only on the anchor model Step-3.5-Flash, while the other nine frontier LLMs are evaluated solely under the core *blinded* $\times$ *open-research* condition; running the full grid on all ten models is compute-bound rather than methodologically blocked, as each LLM on the long window with three seeds takes 24–72 hours of vendor-API wall-clock. We have additionally conducted fragmentary local runs on further models (e.g., GPT, Opus and a Doubao variant), and a multi-model depth replication would yield more granular comparisons across the design axes.

References

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R. Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, Xiaolong Liang, Xiaoming Huang, Bing Zhu, Zhongyu Wei, Yun Chen, Weining Shen, and Liwen Zhang. 2023. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *Preprint*, arXiv:2308.09975.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nicolas Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *Preprint*, arXiv:2311.11944.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. *Preprint*, arXiv:2403.07974.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NIPS)*.

Changlun Li, Yao Shi, Chen Wang, Lingling Zhao, Bo Zheng, Minyi Qin, Jiayi Sun, Wentao Zhang, Xiang Liu, and Tie-Yan Liu. 2025. Time travel is cheating: Going live with deepfund for real-time fund investment benchmarking. *Preprint*, arXiv:2505.11065.

Haohang Li, Yupeng Cao, Yangyang Yu, Xiaomin Zhou, and Xiao-Yang Liu. 2024a. Investorbench: A benchmark for financial decision-making tasks with llm-based agent. *Preprint*, arXiv:2412.18174.

Yang Li, Yangyang Yu, Haohang Li, Zhi Chen, and Khaldoun Khashanah. 2023. TradingGPT: Multi-agent system with layered memory and distinct characters for enhanced financial trading performance. *Preprint*, arXiv:2309.03736.

Yuan Li, Bingqiao Luo, Qian Wang, Nuo Chen, Xu Liu, and Bingsheng He. 2024b. CryptoTrade: A reflective LLM-based agent to guide zero-shot cryptocurrency trading. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations (ICLR)*.

Xiao-Yang Liu, Ziyi Xia, Jingyang Rui, Jiechao Gao, Hongyang Yang, Ming Zhu, Christina Dan Wang, Zhaoran Wang, and Jian Guo. 2022. FinRL-Meta: Market environments and benchmarks for data-driven financial reinforcement learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS) Datasets and Benchmarks Track*.

Yiding Ma. 2026. Oracleproto: A reproducible framework for benchmarking llm native forecasting via knowledge cutoff and temporal masking. *Preprint*, arXiv:2605.03762. `https://github.com/MaYiding/OracleProto`.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations (ICLR)*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *International Conference on Learning Representations (ICLR)*.

Shuo Sun, Molei Qin, Wentao Zhang, Haochong Xia, Chuqiao Zong, Jie Ying, Yonggang Xie, Lingxuan Zhao, Xinrun Wang, and Bo An. 2023. TradeMaster: A holistic quantitative trading platform empowered by reinforcement learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS) Datasets and Benchmarks Track*.

Qianqian Xie, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. 2024. The finben: An holistic financial benchmark for large language models. *Preprint*, arXiv:2402.12659.

Qianqian Xie, Weiguang Han, Zhengyu Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *Preprint*, arXiv:2306.05443.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-source financial large language models. *Preprint*, arXiv:2306.06031. FinLLM Workshop at IJCAI 2023.

Hongyang Yang, Boyu Zhang, Neng Wang, Cheng Guo, Xiaoli Zhang, Likun Lin, Junlin Wang, Tianyu Zhou, Mao Guan, Runjia Zhang, and Christina Dan Wang. 2024. FinRobot: An open-source AI agent platform for financial applications using large language models. *Preprint*, arXiv:2405.14767.

Xiao Yang, Weiqing Liu, Dong Zhou, Jiang Bian, and Tie-Yan Liu. 2020. Qlib: An ai-oriented quantitative investment platform. In *Proceedings of the First ACM International Conference on AI in Finance*.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *Preprint*, arXiv:2406.12045.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

Yangyang Yu, Haohang Li, Zhi Chen, Rui Jiang, Yi Yang, and Xiao-Yang Liu. 2023. Finmem: A performance-enhanced llm trading agent with layered memory and character design. *Preprint*, arXiv:2311.13743.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Rong Liu, Zhenyu Cui, Zhaozhuo Xu, Denghui Zhang, Khaldoun Khashanah, Linyi Yang, Yi Yang, Yinheng Xiao, Jianyun Wang, Yangyang Yang, Mengyou Wang, Daniel Wei, and Xiao-Yang Liu. 2024. FinCon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiayi Sun, Minyi Qin, Xinyi Li, Xue Zhao, Jun Cai, Bo Zheng, Lingling Zhao, Xiang Liu, Bo An, and Tie-Yan Liu. 2024. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. *Preprint*, arXiv:2402.18485.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP)*.

Jie Zhu, Junhui Li, Yalong Wen, and Lifan Guo. 2024. Benchmarking large language models on CFLUE – a Chinese financial language understanding evaluation dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5673–5693.

A Rationale case study across the four mask conditions

The motivating observation in Section 1 is that the same model, under price-only observation, writes qualitatively different rationales depending on what identifying handles are visible. The four masking conditions in Section 3.3 predict a specific decomposition: *stock-blind* should erase ticker-anchored priors (brand, sector, business model) while leaving any calendar-anchored prior intact; *date-blind* should erase calendar-anchored priors while leaving ticker-anchored ones intact; *blinded* should erase both; *bright* should permit both. Boxes below reproduce verbatim rationales for the same anchor model under each condition on equivalent numeric inputs.

The four boxes localize the gap predicted at the top of the section. Under *bright* the rationales name brand, sector, and macro identities (KWEICHOW MOUTAI, YANGTZE POWER, CCB, semiconductor industry recovery) that OHLCV cannot supply;

BRIGHT real ticker IDs visible (4 verbatim rationales)

[SH600519 BUY] "**Kweichow Moutai**, *liquor leader*; low volatility (0.0234), drawdown well controlled (-7.3%), 5-day return $+15.2\%$ ranks among the top—suitable as a steady allocation."

[SH600900 BUY] "**Yangtze Power**, low volatility 0.01, 5-day $+3.1\%$, 20-day $+6.8\%$; in a weak market, add a new 5% position as a *defensive utility* core holding."

[SH601939 BUY] "**CCB** 20-day momentum $+11\%$ tops the screen, low volatility, shallow drawdown—fits a *large-cap low-vol high-momentum* style."

[SH688981 BUY] "**Semiconductor manufacturing leader**, beneficiary of an *industry-recovery expectation*; 20-day return $+108\%$, the Oct-18 limit-up move shows strong drive, volatility 8.31% relatively contained—suitable for allocation."

STOCK-BLIND anonymized ticker, real date (3 verbatim rationales)

[asset_0004 SELL] "asset_0004 current weight 10.38% exceeds the 10% single-name cap, must reduce 160 shares to comply; the name has 20-day return $+23.7\%$ and a strong showing, but *valuation is now elevated*—take partial profit, retain the core."

[asset_0017 BUY] "20-day return tops the screen at $+33.6\%$, long-run trend is clear; despite weaker short-term momentum the *fundamental support remains strong*."

[asset_0023 BUY] "5-day return $+12.9\%$ ranks #1, short-term momentum strong; despite a sharper 20-day drawdown, the *current valuation is reasonable*—allocate moderately to catch the trend."

DATE-BLIND real ticker, anonymized date (3 verbatim rationales)

[SH600111 BUY] "**SH600111** has low volatility and *strong defensive properties*; the *rare-earth sector* carries allocation value, helps diversify portfolio risk."

[SH601857 BUY] "**SH601857** 20-day return $+21.4\%$, low volatility 0.0211, drawdown controlled at -5.45%; *low valuation, high safety margin*."

[SH600362 BUY] "**SH600362** short-term momentum strong and volatility low; 20-day return $+27.6\%$ with controlled drawdown—replace some *high-valuation positions*."

BLINDED anonymized aliases (4 verbatim rationales)

[asset_0006 BUY] "asset_0006 has 20-day return $+27.4\%$, low volatility 0.0255, and small drawdown -3.8%; risk/return ratio excellent, suitable to add up to the 10% cap."

[asset_0013 BUY] "asset_0013 5-day return $+14.5\%$ ranks #2, 20-day $+27.6\%$ ranks #1, low volatility and a maximum drawdown of only -3.87%; risk/return ratio superior, build a new 8% position."

[asset_0028 SELL] "asset_0028 20-day return only $+14.4\%$ ranks low and the unrealised loss is largest; risk/return unfavourable, close the position to release capital."

[asset_0011 SELL] "asset_0011 has the largest unrealised loss and is not in the top-15 by 20-day return; underperforms the market average, replace with a stronger candidate."

under *blinded* the same model on identical numbers produces only ranks, momentum, volatility, drawdown, and unrealised PnL. The intermediate two boxes show why: *date-blind* retains brand and sector priors in full force (`SH600111` "rare-earth sector ... defensive"; `SH601857` "low valuation, high safety margin"), while *stock-blind* reduces them to generic valuation filler with no anchoring referent. The ticker handle, not the calendar handle, carries the dominant pretraining prior.

B Tool interface specification

The open-research decision mode exposes six read-only tools (Table 6). Tool arguments are un-masked by the harness before querying the data layer; the returned payload is re-masked before being handed back to the model, so an agent in *blinded* mode never observes a real ticker or real calendar date even transiently. All tools return JSON dictionaries with explicit type schemas.

Tool	Key args	Returned fields
`get_market_context`	—	indices, calendar, sector tone
`screen_candidates`	factors, k	ranked aliases + factor values
`get_stock_snapshot`	alias, lookback	OHLCV bars, factors, prev close
`compare_candidates`	aliases, dims	side-by-side metric grid
`portfolio_state`	—	holdings, cash, T+1 locks
`risk_check`	target weights	violations, projected limits

Table 6: Six read-only tools under *open-research*. The agent may invoke any tool any number of times within a step.

C Prompts and a worked masked trace

One system prompt is given per decision mode in Section 3.2, and a single user-message template is shared across modes.

System prompt: *open-research* (headline configuration)

```
You are an autonomous Chinese A-share trading agent. Within the
current backtest trading day, conduct research, judgement, order
submission, and self-constraint based on available data; emit a
single valid JSON trading decision.

## Goals
- Maximise the overall benchmark score within the hard constraints.
- Decide on your own whether to trade, how many names, concentrated
  or diversified, and whether to use `target_weight` or `shares`.
- If no identifiable edge exists, holding cash or maintaining the
  status quo is an acceptable decision.

## Degrees of freedom
- No fixed workflow is required; decide which tools to call, in what
  order and frequency, and when to stop research and submit.

## Information boundary
- Decisions may rely only on the current message, system context,
  and tool returns; do not assume any externally available
  information.
- All tools have time semantics strictly earlier than the current
  backtest day: market-data tools see the previous trading day and
  earlier; news / report / announcement / search tools have their
  returns filtered as of the current day, but you should still
  verify the date fields in returns to avoid indirectly referencing
  future information.
- The current day's open, close, and intraday data are not visible.

## Hard constraints
1. Long-only; no shorting.
2. Only stocks in today's tradable universe may be operated on.
3. A-share T+1: a stock bought today can only be sold the next day.
4. Each order must contain `stock_id`, `side`, `confidence`, and
   `reason`, with `reason` of at least 10 characters.
5. If no action is intended, return an empty `orders` list.

## Autonomous decision space
- Neither concentration nor the number of holdings has a hard upper
  bound; you may concentrate heavily (e.g., 1-3 names) or diversify
  broadly. Weigh win rate, risk budget, and candidate quality on
  your own.
- State your portfolio-construction rationale in `overall_reason`.

## Limit-board rules (engine layer)
- Price limits are board-differentiated: main / SME (6xxxxx / 00xxxx)
  +/-9.5%; ChiNext (30xxxx) and STAR (688xxx) +/-19.5%; Beijing
  Stock Exchange (8xxxxx / 4xxxxx) +/-29.5%.
- A next-day open at the upper limit rejects BUY; lower limit
  rejects SELL; one-sided days (open == high == low at a limit
  price) are unfillable.
- Tools expose `board / limit_pct / limit_up / limit_down` and
  limit-hit warnings per stock.

## Execution details
- Orders execute sequentially in the order given; on days with both
  buys and sells, list SELL before BUY to avoid running out of cash.
- `target_weight` is the final desired weight; do not re-submit BUY
  for a position whose current weight is already near or above it.
- When opening longs, retain a 1%-3% cash buffer (fees, lot rounding,
  spread).
- SELL only against holdings with `shares_available > 0`.
- When cash or sellable shares are uncertain, prefer to under-trade,
```

```
  lower the target, or do nothing.

## Decision requirements
- Emit an executable trading decision, not an analytical report.
- Order intentions must be clear, non-conflicting, and free of
  meaningless instructions.
- `confidence` in [0, 1]; `overall_reason` >= 20 characters.

## Output format
Emit a single JSON object only, starting with `{` and ending with
`}`; no markdown code-block wrappers, no extra text, no comments,
no tool-call summaries. See Appendix for the full action schema.
```

System prompt: *fixed-candidate*

```
You are a professional Chinese A-share quantitative analyst. The
current experimental mode is fixed-candidate: a fixed candidate
pool is provided; you cannot call tools and cannot extend the
candidate set on your own.

Important notes
- Stock identifiers appear in the form provided in the input
  (e.g., asset_0042); use the input identifiers verbatim.
- The current step identifier may be a real date or a relative
  index. Do not use external memory, news, real-stock common
  knowledge, or any data not provided.
- BUY only from the fixed candidate pool; SELL only against current
  holdings; an empty `orders` list is acceptable.
- Long-only; A-share T+1 (sell the next day at the earliest);
  concentration and number of holdings have no hard upper bound,
  decide on your own.
- Each order must contain `stock_id`, `side`, `confidence`, and
  `reason`; `target_weight` and `shares` are mutually exclusive.

The output must be a single pure JSON object (no markdown, no
extra text).
```

System prompt: *memory-only*

```
You are a professional Chinese A-share quantitative analyst. The
current experimental mode is memory-only: the system provides only
today's tradable universe (identifiers), your account state, and
constraints. No price, volume, return, volatility, factor, or
indicator data is provided, and no tool calls are permitted.

Important notes
- The input contains only stock identifiers. An identifier may be
  a real ticker (e.g., SH600519) or an anonymous alias (e.g.,
  asset_0042); use the input identifiers verbatim.
- The current step identifier may be a real date or a relative
  index.
- Your decision may rely only on the identifiers, account state,
  and constraints in the input. If you have a view on a given
  identifier, you may act on it; if you have no basis for a
  decision, you may return an empty `orders` list.
- BUY only from today's tradable universe (identifiers present in
  the input); SELL only against current holdings; no action is
  always allowed.
- Long-only; A-share T+1; concentration and number of holdings have
  no hard upper bound.
- Each order must contain `stock_id`, `side`, `confidence`, and
  `reason`; `target_weight` and `shares` are mutually exclusive.

The output must be a single pure JSON object (no markdown, no
extra text).
```

User-message template

A single user-message template is shared across decision modes; the `{...}` placeholders are filled at runtime by the harness. Under *blinded*, `date_label` renders as a relative day index (e.g., `day_+12`) and `positions_table` / `universe_preview` / `features_table` contain only `asset_NNNN` aliases.

```
## Backtest step: {date_label}

## Constraint summary
- Concentration and number of holdings have no hard upper bound;
  decide on your own.
- Short selling: {allow_short}
- T+1 settlement: yes

## Current holdings and account
- Total value: {total_value} CNY
- Cash: {cash} CNY ({cash_weight} of NAV)
- {num_positions} positions:
{positions_table}

## Today's tradable universe ({universe_size} names)
{universe_preview}

## Market summary over the last {obs_window} observations
## ({features_count} names; data cutoff T-1 close)
{features_table}

## Previous-step execution outcome
{prev_execution_summary}

## Portfolio performance
- Cumulative return since the start of the backtest:
  {portfolio_ret_cum}
- CSI300 cumulative return over the same window:
  {benchmark_ret_cum}

Emit today's trading decision based on the above (pure JSON, no
markdown).
```

Worked masked trace: one step under *blinded* × *open-research*

Box C reproduces the agent's view of a single step under *blinded* × *open-research* (anchor model Step-3.5-Flash, window W1, step 0). The agent issues four tool calls (`get_market_context`, `screen_candidates`, `compare_candidates`, `risk_check`) and one submission; we show three of the calls and the final action. Every identifier the agent sees is an alias and every date is a relative day index; the executor un-masks aliases back to real tickers post-hoc for order execution, but the agent itself never observes a real identifier in its inputs or outputs.

The trace is the empirical realization of the masking-consistency claim in Section 3.3: aliases and relative day indices propagate end-to-end through every tool argument, tool return, and the agent's submitted action. Field-level masking is also visible: `as_of_date` and `data_cutoff` are relative indices, candidate `stock_ids` are aliases, and the `reason` strings reference only numeric features (ranks, returns, volatility) rather than brand or sector language.

BLINDED × OPEN-RESEARCH **anchor model, W1 step 0** **agent's view**

[1] tool_call get_market_context()

[1] tool_result {as_of_date: day_+0, data_cutoff: day_-3, contains_current_day_market_data: false, market: csi300, universe_size: 300, ...}

[2] tool_call screen_candidates(sort_by=ret_20d, top_k=15)

[2] tool_result {as_of_date: day_+0, data_cutoff: day_-3, candidates: [
{stock_id: **asset_0001**, prev_close: 3.034, ret_1d: 0.0185, ret_5d: 0.1123, ret_20d: 0.3355, vol_20d: 0.0302, drawdown_20d: -0.055, rank: 1},
{stock_id: **asset_0002**, prev_close: 0.567, ret_1d: 0.0996, ret_5d: 0.2213, ret_20d: 0.3104, vol_20d: 0.0272, drawdown_20d: -0.045, rank: 2}, ...}

[3] tool_call risk_check(draft_orders=[
{stock_id: **asset_0001**, side: BUY, target_weight: 0.10, confidence: 0.85, reason: "20-day return ranks #1 (+33.6%); short-term momentum strong; core allocation"},
{stock_id: **asset_0002**, side: BUY, target_weight: 0.10, confidence: 0.82, reason: "1-day +9.96% and 5-day +22.1% both rank #1; explosive short-term strength"}, ...])

[3] tool_result {valid: true, violations: [], projected_weights: {asset_0001: 0.10, asset_0002: 0.10, ...}}

[final] submit_action {orders: [
{stock_id: **asset_0001**, side: BUY, target_weight: 0.10, confidence: 0.85, reason: "..."},
{stock_id: **asset_0002**, side: BUY, target_weight: 0.10, confidence: 0.82, reason: "..."}, ...],
overall_reason: "High-momentum, low-volatility names dominate; build five-name core positions and retain cash buffer."}

D Action schema

The submission phase emits one JSON object per step, schema-validated by the harness:

```
{
  "orders": [
    { "stock_id":      "asset_0042",
      "side":          "BUY" | "SELL",
      "target_weight": 0.05,
      "confidence":    0.75,
      "reason":        "..." }
  ],
  "overall_reason": "..."
}
```

The `confidence` field feeds the calibration metrics (Appendix G); the `reason` field feeds the motivation analysis (Section 1) and the case study in Appendix A.

E Style factors

We select nine classic factors with low mutual correlation via VIF, as is shown in Table 7

Factor	Description
MOM_12_1	12-minus-1 month momentum
RV_60	60-day realized volatility
ILLIQ	Amihud illiquidity
REV_ON	Overnight return
MOM_ID	Intraday momentum
SKEW	Negative return skewness
CORR_PV	Volume–return correlation
HIGH_52W	52-week high anchor
CV_VOL	Amount coefficient of variation

Table 7: Nine style factors used in the cross-sectional attribution model.

F Cohort-level factor exposures

Table 8 reports the portfolio-weighted, z-scored exposures from Section 5.3.

Factor	ML	LLM	ML − LLM
MOM_12_1	+0.34	+0.40	−0.07
RV_60	−0.66	+0.23	**−0.89**
ILLIQ	−0.34	−0.32	−0.02
REV_ON	+0.12	+0.28	−0.17
MOM_ID	−0.12	+1.06	**−1.17**
SKEW	+0.42	−0.08	**+0.50**
CORR_PV	−0.26	+0.33	**−0.59**
HIGH_52W	+0.80	+0.77	+0.03
CV_VOL	−0.60	+0.11	**−0.71**

Table 8: Mean portfolio-weighted z-scored factor exposures by cohort; positive values indicate net long exposure, negative values net short. Bold entries mark $|\text{gap}| > 0.4\sigma$.

G Ten-dimensional metric definitions

Returns and risk. *Total return* is the geometric return over the evaluation period from end-of-day NAV. *Sharpe ratio* is annualized with $r_f = 0$ and a 252-day convention. *Maximum drawdown* is the largest peak-to-trough decline in NAV. *Information ratio* is the annualized mean of daily excess return over CSI300 divided by its annualized standard deviation.

Behavior. *Annualized turnover* is the sum of absolute weight changes per step, scaled to a 252-day year. *HHI* is the average across steps of $\sum_i w_i^2$ over current holding weights. *Cash utilization* is the average fraction of NAV held as cash. *Abstention rate* is the fraction of steps producing zero orders after retries.

Reliability and calibration. *Parse-failure rate* is the fraction of action emissions failing JSON schema validation after three retries. *Tool-validity rate* is the fraction of tool calls whose arguments parse and dispatch successfully. *ECE* bins per-order `confidence` into 10 equal-width bins and reports the weighted mean absolute gap between bin confidence and bin accuracy. *Brier score* treats per-order confidence as the probability that the order's realized next-period return is positive.

H Qlib factor model configuration

The eighteen Qlib factor baselines use the published Qlib defaults for the CSI300 universe (hyperparameters unchanged from the official model-zoo configuration files); only the ranking signal differs across models. The portfolio constructor is identical across models: top-20 by predicted score with a cost-aware rebalancing threshold of 0.06. Table 9 reports the complete leaderboard for all eighteen baselines; the main paper (Table 4) shows eight representatives.

A-share price-limit thresholds (used by both LLM agents and Qlib baselines). The execution engine applies a small buffer below each board's official daily limit to avoid fills at unreachable prices: 9.5% for main-board issues (Shanghai `6xxxxx` and Shenzhen `00xxxx`, daily limit ±10%), 19.5% for ChiNext (`30xxxx`) and STAR Market (`688xxx`, both ±20%), and 29.5% for the Beijing Stock Exchange (`8xxxxx`, ±30%).

Model	Total ret.	Excess	Sharpe	MDD	IR	Turn.
SFM	+86.58%	+49.66%	2.02	−7.41%	+0.57	22.47
DoubleEnsemble	+73.66%	+36.74%	1.41	−18.24%	+0.41	72.84
CatBoost	+72.97%	+36.05%	1.43	−16.74%	+0.40	29.96
XGBoost	+66.41%	+29.49%	1.12	−17.75%	+0.31	101.66
LightGBM	+66.22%	+29.30%	1.29	−15.59%	+0.33	46.18
ADARNN	+66.13%	+29.21%	1.66	−8.52%	+0.35	16.19
KRNN	+62.09%	+25.17%	1.63	−8.88%	+0.31	20.80
Transformer	+60.35%	+23.43%	1.61	−11.29%	+0.27	11.22
Sandwich	+56.57%	+19.65%	1.49	−8.90%	+0.24	29.78
ADD	+54.03%	+17.11%	1.47	−9.92%	+0.20	22.52
Localformer	+53.02%	+16.10%	1.41	−10.59%	+0.19	13.32
TCN	+52.33%	+15.41%	1.31	−10.27%	+0.18	30.17
TabNet	+52.25%	+15.33%	1.38	−10.61%	+0.18	26.13
LSTM	+43.92%	+7.00%	1.12	−11.76%	+0.06	78.03
CSI300 buy-and-hold	+36.92%	0.00%	—	—	—	0.00
MLP	+34.08%	−2.84%	0.69	−21.75%	−0.01	18.48
GRU	+34.05%	−2.87%	0.92	−13.50%	−0.07	57.79
Linear	+29.60%	−7.32%	0.74	−13.88%	−0.12	15.61
ALSTM	+22.67%	−14.25%	0.64	−15.15%	−0.22	95.31

Table 9: Complete leaderboard for all eighteen trained Qlib factor models on the long 2024-01-01 to 2026-04-10 window, sorted by total return. CSI300 buy-and-hold inserted at its rank position.

I Harness configuration

Configuration values used in the headline ten-LLM evaluation (*blinded* × *open-research*, long window 2024-01-01 to 2026-04-10).

Group	Parameter	Value
Execution	benchmark	CSI300
	initial cash (CNY)	1,000,000
	buy cost (bps)	5
	sell cost (bps)	15
	min cost (CNY)	5
	deal price	next-day open
	T+1 settlement	yes
	long-only	yes
Agent	max candidates / step	100
	max tool calls / step	99
	max positions held	300
	max single weight	1.00
	schema retries	2
	retry with feedback	yes
	min reason length	10
	fallback action	hold
LLM call	temperature	0.0
	max tokens	32,768
	timeout (s)	600
	inter-call gap (s)	3

Table 10: Headline-experiment harness configuration.

J Mask sensitivity on additional frontier LLMs

We replicate the *bright* vs. *blinded* contrast under *open-research* on three additional frontier LLMs over ten regime-stratified windows with three seeds per cell; each cell reports the seed- and window-median (Table 11).

Model	Mask	Ret. (%)	Abst.	Turn.
gpt-5.5	bright	−0.15	0.18	26.7
	blinded	−1.43	0.12	29.9
doubao-seed-2.0	bright	−2.97	0.03	31.3
	blinded	−4.07	0.00	33.7
claude-opus-4-7	bright	−0.40	0.20	21.4
	blinded	−1.44	0.02	73.5

Table 11: Mask sensitivity across three frontier LLMs under *open-research* on ten regime-stratified windows with three seeds per cell. `Ret.` is median total return; `Abst.` is median abstention rate; `Turn.` is median annualized turnover.

K De-anonymization probe details

Protocol. For each (stock, date) probe the harness calls `get_market_context`, `screen_candidates`, and `get_stock_snapshot` with bars through the alias map with *blinded* on, then saves the masked payload alongside the gold (ticker, date, sector, board). Cells receive Wilson 95% confidence intervals.

Full-attacker breakdown. Table 12 reports per-attacker recovery rates under `full`. Each attacker is given the same 200 randomly sampled probes; rows report rates over successful API responses only, excluding calls that failed to return a parseable answer. Top-1 ticker accuracy never exceeds 3.0% against a random baseline of 0.3%, and joint success (ticker

in top-5 *and* date within ± 7 trading days) never exceeds 1.5%. Board accuracy is intentionally high because the board field is not anonymized—the agent needs it for limit-price compliance—and serves as a sanity check.

Attacker	tk@1	tk@5	board	date ±7d	date ±30d	date ±90d	success
`step-3.5-flash`	0.0%	2.8%	100.0%	4.4%	13.3%	26.7%	0.0%
`claude-opus-4-7`	3.0%	10.2%	100.0%	1.0%	10.2%	22.8%	0.0%
`doubao-seed-2.0`	1.5%	9.0%	100.0%	7.0%	14.0%	23.5%	0.0%
`gpt-5.5`	2.0%	8.0%	99.5%	7.5%	13.0%	20.5%	0.5%
`qwen3.6-plus`	1.0%	4.6%	100.0%	6.2%	12.3%	28.7%	0.0%
`minimax-m2.7`	1.1%	4.9%	99.5%	3.3%	12.0%	21.7%	0.0%
`deepseek-v4-pro`	0.5%	5.5%	100.0%	10.5%	18.0%	34.0%	0.5%
`glm-5.1`	0.0%	6.4%	100.0%	16.0%	24.5%	45.7%	1.1%
`kimi-k2.6`	2.2%	9.8%	100.0%	1.6%	11.5%	29.0%	0.5%
`gemini-3.1-pro`	3.0%	8.6%	100.0%	14.7%	21.8%	36.5%	1.5%
random baseline	0.3%	1.4%	25.0%	1.9%	8.2%	24.7%	0.0%

Table 12: De-anonymization attacker probe under the `full` masked payload (the exact view the agent receives in *blinded* mode). Each attacker sees 200 randomly drawn probes; rates are over successful responses only. Top-1 and joint success are the headline numbers (Section 4.3); board accuracy is intentionally high (board field unmasked) and serves as a sanity check.