



Summoning the Oracle to Slay It:

Mitigating Look-Ahead Bias in Financial Backtesting with Large Language Models

Weixian Waylon Li Mengyu Wang Tiejun Ma

University of Edinburgh

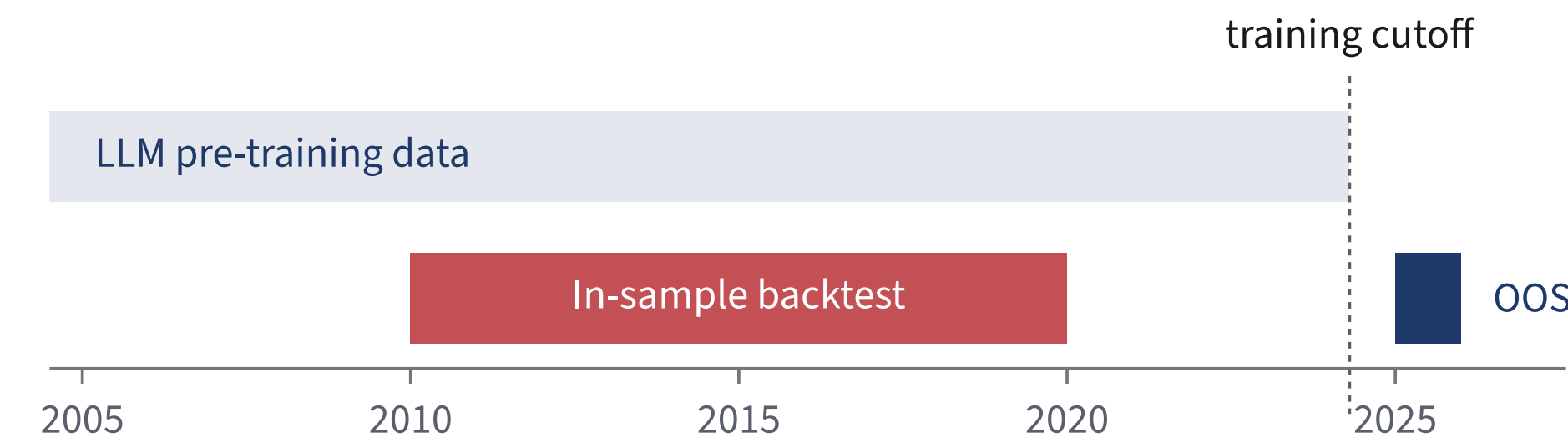
{waylon.li, mengyu.wang, tiejun.ma}@ed.ac.uk

EMNLP2026
BUDAPEST

University logo

Motivation

LLM agents are increasingly backtested on historical market data. If a model's pre-training corpus overlaps the backtest window, the model may already know what happened next.



In-sample windows (2010–2020) lie inside the training data; the clean post-cutoff window (2025) is short.

We call this **parametric look-ahead bias**. It lives in the model weights, so data-pipeline controls (point-in-time data, anonymisation) cannot remove it. Long backtests become misleading, while the bias-free window is too short for reliable conclusions.

Method

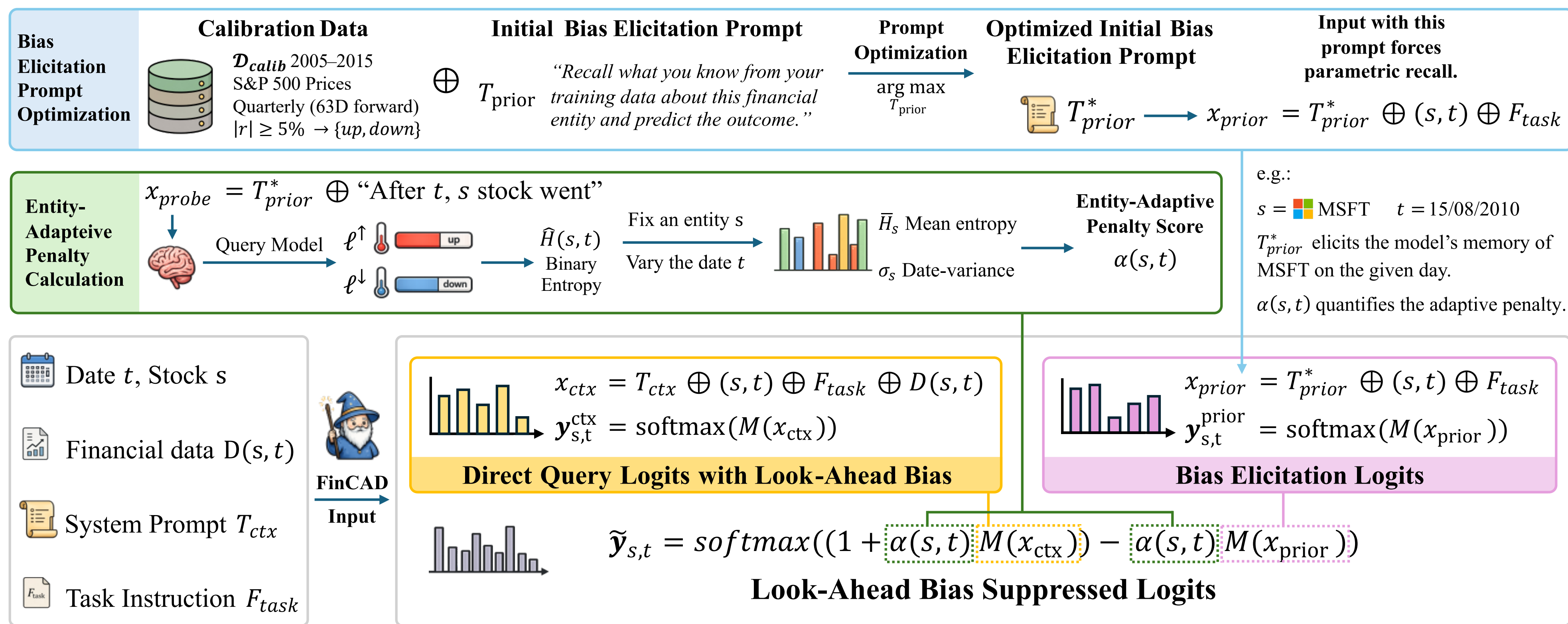


Figure 1. FINCAD pipeline: (1) bias-elicitation prompt optimisation; (2) entity-adaptive penalty calculation; (3) look-ahead-bias-suppressed decoding.

Ranking Alignment

Does the in-sample leaderboard predict post-cutoff performance? 11 LLMs, SPY, Sharpe rankings; mean over all $\binom{11}{7}=330$ model subsets.

Method	Spearman $\bar{\rho}$		Kendall $\bar{\tau}$	
	Sharpe	Sortino	Sharpe	Sortino
Baseline	.779	.642	.673	.491
Anonymisation	.547	.488	.418	.382
Prompt-injection	.482	.320	.345	.236
FINCAD	.846	.694	.709	.527

Table 1. Subset-averaged correlation between each in-sample condition and the post-cutoff ranking.

FINCAD is the only mitigation that improves alignment; input-side fixes reduce it. The gain holds in 10 of 11 leave-one-model-out runs and for every subset size $k = 5, \dots, 11$.

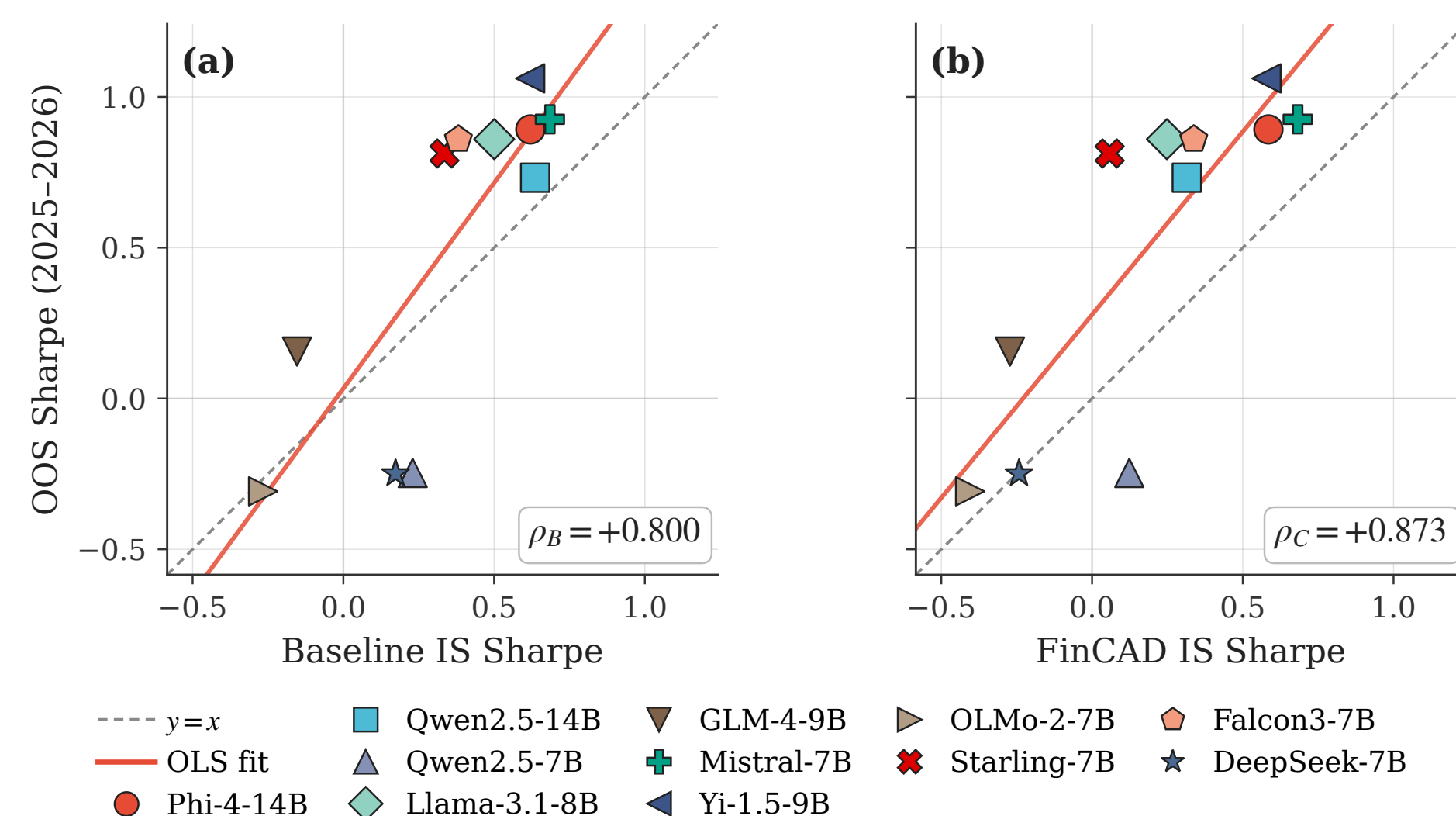


Figure 2. (IS, OOS) Sharpe per LLM, baseline (a) vs. FINCAD (b). Red: OLS fit; dashed: $y=x$.

Approach

Context-Aware Decoding (CAD) subtracts an off-context prior from the context-conditioned logits. FINCAD repurposes it: the prior is *elicited* to carry the model's memory of historical outcomes, then subtracted with an entity- and date-specific strength:

$$\tilde{\mathbf{y}}_{s,t} = \text{softmax}((1 + \alpha(s, t)) \mathcal{M}(x_{\text{ctx}}) - \alpha(s, t) \mathcal{M}(x_{\text{prior}}))$$

s : stock, t : decision date, x_{ctx} : agent input with financial context, x_{prior} : memory-activating prompt without context.

Two design questions follow:

- Which prior?** x_{prior} must activate recall of outcomes, not answer the task.
- How strong?** Memorisation varies across entities and dates, so a fixed α under-corrects some cases and over-penalises others.

Contributions

- FINCAD, the first decoding-time adaptation of CAD for mitigating parametric look-ahead bias, requiring **no retraining**.
- An **adversarial bias-discovery** pipeline that learns a model-specific memory-activating prior, and an **entity- and date-adaptive** rule for $\alpha(s, t)$.
- Evaluation on 5–11 LLMs (7–14B), mega- to small-cap equities and five reasoning benchmarks.

Key results

- In-sample vs. post-cutoff ranking alignment (11 LLMs): Spearman $\bar{\rho}$ **0.779** $\rightarrow$ **0.846**.
- In-sample return corrections up to **−67.1%** (model mean), with near-zero change post-cutoff for the larger models.
- Mean benchmark accuracy positive or within **−1.7** points for four of five models.

(1) Adversarial bias discovery

On S&P 500 quarter-end dates (2005–2015), MIPROv2 optimises the instruction that best recovers the 63-day up/down direction from memory alone:

$$T_{\text{prior}}^* = \arg \max_{T_{\text{prior}}} \sum_i \delta(\hat{d}_i, d_i).$$

Optimised once per model; the proposer never sees tickers, dates or labels.

(2) Entity- & date-adaptive α

The probe “ $T_{\text{prior}}^* \oplus$ After t , s stock went” gives a normalised entropy $\hat{H}(s, t)$; date-variance σ_s separates date-specific memory from stable brand priors:

$$\alpha(s, t) \propto \min\left(1, \frac{\sigma_s}{\sigma_{\text{rel}}}\right) \max\left(0, \frac{\hat{H}_s - \hat{H}(s, t)}{\Delta_{\text{range}}}\right),$$

so α falls to 0 when post-cutoff confidence drops.

(3) Contrastive decoding

Both branches share (s, t) and F_{task} , so the subtraction targets content rather than format. Cost: one extra forward pass per token and one probe per decision; calibration is offline.

In-Sample Correction

Single-stock backtests on five mega-caps; change in ending value, $\Delta = (C - B)/|B|$ (%).

Ticker	Phi-4-14B		Qwen2.5-14B	
	IS	OOS	IS	OOS
NVDA	−37.5	0.0	−78.2	0.0
MSFT	+1.2	0.0	−60.8	+8.0
AAPL	−19.5	0.0	−17.9	+1.1
NFLX	−22.3	0.0	−42.0	0.0
AMZN	−17.2	0.0	−28.2	0.0
Mean	−19.1	0.0	−45.4	+1.8

Table 2. IS: 2010–2020; OOS: 2025. Daily rebalancing, initial \$100K.

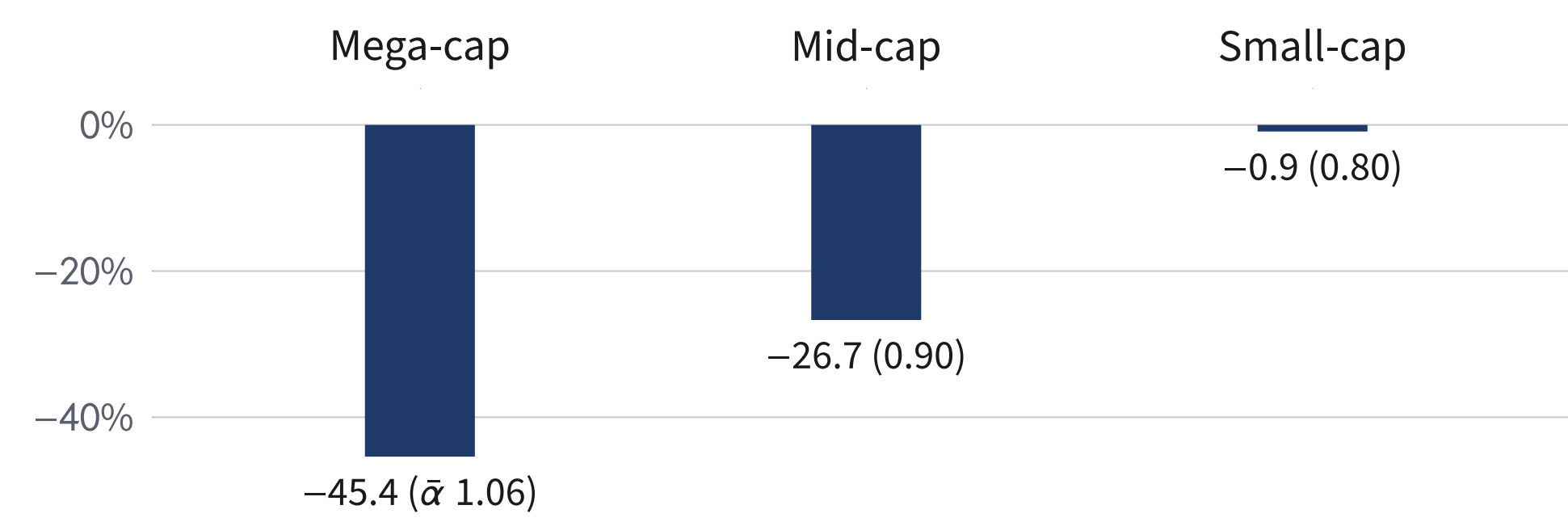


Figure 3. Qwen2.5-14B mean IS return change and mean $\bar{\alpha}_S$ by market-cap tier.

- Corrections are large in-sample and near zero post-cutoff: Phi-4 receives $\alpha = 0$ on every 2025 trading day.
- Less-covered stocks receive smaller α and smaller corrections.

Reasoning Preservation

Accuracy change (points, FINCAD – baseline) at a fixed model-specific α_{bench} .

	Phi	Qwen	Llama	Starl.	DSeek
GSM8K	+0.5	+0.4	−4.9	−6.9	−21.3
MMLU-Pro	−0.2	+2.1	+6.6	+0.2	+2.4
Comp. Math	+0.8	−2.0	−7.8	−1.0	−4.6
FinEval	+7.1	−0.6	+0.2	−1.0	+2.8
HumanEval	+3.0	+0.6	+0.6	0.0	−0.6
Mean	+2.2	+0.1	−1.0	−1.7	−4.3

Table 3. Phi-4-14B, Qwen2.5-14B, Llama-3.1-8B, Starling-7B, DeepSeek-7B-Chat. Greedy decoding.

Four of five models stay positive or within $−1.7$ points; DeepSeek-7B-Chat is the exception, driven by GSM8K.

Conclusion

FINCAD mitigates parametric look-ahead bias at inference time: it makes the model's memory of historical outcomes explicit and subtracts it only where it is detected. It complements input-side protections, and plays a role for LLM backtests similar to the deflated Sharpe ratio for backtest overfitting.

Limitation. FINCAD is a mitigation, not a guarantee that all training-data leakage is removed; formal inference needs more models and post-cutoff periods.



Paper



Code



Project page