When Market Data Identifies Itself: Linkage Attacks and Agentic Re-identification in Financial LLM Evaluation

Anonymous Author(s)

Abstract

Financial-LLM benchmarks increasingly hide tickers and dates behind aliases, so that a model cannot exploit memorized outcomes for a specific stock in a specific period. We show that hiding the labels is not enough: the numbers that remain can identify the security and period on their own. We audit three published blinding schemes with an adversary that links released numbers back to public market data, with all attacks and metrics frozen on development data before a one-shot evaluation on disjoint tickers and dates. The attacker recovers both ticker and date for 99.0% of blinded price packets (KTD-style) and 94.9% of reconstructed BlindTrade exposure traces, and identifies 21.1% of firms from ratio-valued fundamentals against a 0.76% sector-conditioned null. All five predeclared tests survive Holm correction ($p_{\text{Holm}} = .0025$). Rebasing every price series to a common starting value does not defeat the attack, and recovery stays at 100/100 in a frozen scaling study even with 3.51 million candidate ticker-date windows to search. A sealed mitigation study shows the trade-off directly: six continuous financial summaries remain 99.5% linkable, while train-fitted bins cut ticker recovery to 0.5%. But the bins come at a cost: the binned data no longer passes our predeclared checks that the underlying financial signal survives, and a replication on a later period fails the same checks. Finally, in frozen agent studies, coding agents from two model families build working open-set linkage attacks from generic Python and raw market history when explicitly asked; agents given a neutral task discover identities only occasionally, and a randomized experiment finds no evidence that knowing the true identity improves an agent's forecasts. Numeric blinding can therefore remain linkable and exploitable by a motivated adversary. Whether agents exploit it unprompted, and whether doing so yields any financial advantage, are questions our data do not settle.

CCS Concepts

• **Applied computing** → **Economics**; • **Computing methodologies** → *Natural language processing*; • **Security and privacy** → *Data anonymization and sanitization.*

Keywords

large language models, financial agents, evaluation, memorization, record linkage, anonymization, trustworthy AI

1 Introduction

Historical evaluation creates a distinctive problem for financial language models. A model may have encountered the evaluated company, period, and realized outcome during pretraining. Recent work consequently masks tickers, company names, or calendar dates before asking an LLM to forecast or trade [6, 8, 9, 14]. If performance survives, the experiment is often interpreted as evidence that the model used the supplied financial signal rather than memorized identity-specific information.

That interpretation silently assumes that the remaining signal is not itself an identifier. A sequence of returns, technical indicators, or accounting ratios is high-dimensional and publicly searchable. Removing the header from such a packet may be analogous to removing usernames from a ratings table: the values that make the data useful can also link it back to the public record [10, 12]. In that case the model receives neither a name nor an anonymous representation. It receives a fingerprint.

Existing financial studies mostly validate masking either by comparing blinded and unblinded model behavior or by asking an LLM to guess the hidden identity. These are useful behavioral probes, but neither tests whether the packet is linkable by a tool-using system with access to public market data. KTD-Fin, for example, reports a ten-LLM de-anonymization panel whose strongest attacker reaches 10.2% ticker Top-5 and at most 1.5% joint recovery, and treats the blinded condition as nearly clean [14]. BlindTrade states that its anonymization effect is not directly validated [8].

We perform a representation audit rather than propose a new generic anonymizer. We instantiate three public numeric schemas: KTD-Fin's paper-specified masked OHLCV representation, BlindTrade's named indicator inputs, and a code-derived port of the S&P 500 fundamentals arm of Epistemic Blinding [1]. For each, we ask a simple question: can a database adversary recover the security and, when applicable, the historical period from only the fields the model sees? Our primary attack is simple by design: nearest-neighbor linkage against a different public data provider. If even this baseline recovers the identity, the blinding claim fails without any appeal to a stronger adversary.

Our contributions are:

(1) a threat model separating *parametric recognition*, *database linkage*, *agentic construction*, and *downstream exploitation*, with a reproducible audit harness for finance-facing blinding protocols.
(2) a preregistered, sealed confirmatory study: the attacker recovers 99.0%, 94.9%, and 21.1% of packets across the three schemes, all five family tests survive correction, and recovery holds through price rebasing and a search over millions of candidates.
(3) a sealed mitigation study on 600 firms: continuous summaries stay linkable, and the binned release that stops the attack also fails our predeclared signal checks.
(4) frozen agent studies: coding agents from two model families, given only generic Python and raw market history, build working open-set attacks without our matcher. Control conditions test spontaneous discovery and downstream benefit.

What follows from recoverability is precise: the inference "the system could not know the identity from this packet" no longer holds, and our agent studies show that a coding agent asked to audit anonymity can build the attack itself, without being given our matcher. What we do not claim is that agents invent the attack unprompted, that recovery reflects parametric memorization, or that recovered identity improves financial decisions. The first and third we measure directly in Section 6: neutral-prompt agents discover identities only occasionally, and a randomized assay finds no forecasting benefit. The second we leave open.

2 Blinding Claims and Threat Models

2.1 Three audited protocols

Each audited protocol removes labels but keeps a rich numeric payload. Table 1 records what each representation hides, what remains, and where our instantiation departs from the original evaluation.

KTD-Fin. KTD-Fin masks stock and date handles end-to-end while preserving numeric signals computed on the original series. Its tools return OHLCV bars, factors, and prior close. Aliases are stable within an episode, and dates become relative indices. No official implementation was located at audit time, so we test a reduced, protocol-favorable numeric instantiation rather than its full probe payload. Licensed dual-provider CSI300 data for its Chinese A-share universe were also unavailable, so the stated estimand is resistance of the paper-specified representation on a frozen U.S. universe.

BlindTrade. BlindTrade replaces tickers with synthetic IDs and feeds specialist agents technical features, but its published prompts retain absolute execute and cutoff dates. We reconstruct the nine named numeric inputs of its momentum and mean-reversion agents, including price-to-moving-average ratios, ADX, MACD histogram, RSI, relative volume, Bollinger position, 20-day deviation, and stochastic K. Code, indicator equations, precision, and the numeric user-side serializer are unavailable, so we report a single reconstructed stock-day snapshot and a multi-day accumulated-exposure trace separately. The audit deliberately suppresses dates, randomizes IDs per packet, uses full pre-window warm-up, and rounds to four significant digits, all protocol-favorable choices. Because the source serializer is unspecified, a descriptive development-split grid varies it. Recovery is unchanged at 94.9% from two to five significant digits and under fixed-decimal rounding or truncation. Per-feature z-scoring defeats only a matcher that ignores the published schema. We do not simulate news, risk-regime inputs, or the downstream graph and optimizer.

Epistemic Blinding. The released tool replaces the ticker with a generated company code while passing sector and fundamentals to the LLM. Our port follows the released code rather than the paper's description: the code selects already-ratio-valued columns, retains sector, and applies no within-sector normalization. We regression-test the mapping and 17-feature schema against the released commit, then use a new data vintage and a rowwise attack. Because the released company codes can be inverted from public repository artifacts, the database attack ignores the code entirely and identifies firms from sector and fundamentals alone.

2.2 Four stages of evidence

Let X denote the rendered packet, I its hidden identity and period, D a public auxiliary database, and A an LLM agent.

Parametric recognition measures $P(\hat{I} = I \mid X, A)$ without tools. It depends on what a particular model memorized and whether prompting elicits that memory. **Database linkage** measures $P(\hat{I} = I \mid X, D)$ under an explicit search universe and matching rule. It is a property of the representation relative to public auxiliary information. **Agentic construction** measures whether a tool-using model can create and execute the linking procedure from an objective and generic resources, rather than merely call a supplied reverse linker. **Downstream exploitation** asks whether supplying or recovering I changes decisions or utility. These stages are not interchangeable. Database linkage shows that identity is present relative to D, not that an agent will discover or exploit it; agentic construction makes the vulnerability operational under a specified workflow without establishing spontaneous use or financial benefit; and a failed LLM guess does not certify that the representation is unlinkable.

This distinction parallels re-identification work, where linkage against auxiliary data is the canonical test of a released representation [10, 12]. Here no personal privacy harm is claimed: securities and prices are public. The protected object is the validity of a contamination-controlled evaluation.

3 Audit Design

3.1 Data and sealed splits

Attacks link packets built from one public data provider against an index built from a different one, so recovery cannot be an artifact of exact file identity. Packets (Source A) use adjusted daily Yahoo Finance OHLCV. The primary attack index (Source B) uses Alpaca's IEX feed, which observes the same public market but differs in venue coverage, OHLC extremes, volume, and data processing. The fundamentals attack links Yahoo packet fields against SEC EDGAR company facts. Same-source Yahoo indices are positive controls, not primary evidence.

The universe is a frozen S&P-500-style snapshot of 503 symbols, written U500. Development uses 100 sector-stratified firms (U100) and dates in 2022–2023. Confirmatory packets use the remaining 403 tickers and dates from July 2024 through May 2025, so the confirmatory firms and periods are disjoint from everything the attacks were tuned on. After coverage rules, the confirmatory price library contains 392 packet-source and 486 attack-index symbols, and the EDGAR index contains 498 firms. Each primary trajectory sample contains one non-overlapping packet from each of 100 distinct firms. One true trajectory is absent from Source B, leaving $n = 99$ under the predeclared coverage rule. The fundamentals sample contains 375 distinct firms.

The representation-audit design, adapters, outcomes, and split were frozen before attacks touched development data. After development, the composite weights and five-test family were hash-frozen before a single confirmatory packet was built. The sealed confirmatory split (A2) ran exactly once, and every figure reads the committed JSON artifacts rather than recomputing attacks. The

Table 1: Audited representations and fidelity tiers. Results test the declared U.S.-equity instantiations, not exact replications of the source trading studies.

Protocol	Stated purpose	Packet available to the LLM	Fidelity	Audit adaptation
KTD-Fin [14]	Separate historical memory from investment reasoning	Alias and relative day labels; original-series OHLCV and derived features	Paper-specified representation stress test	Reduced numeric payload; CSI300 → frozen U.S. universe; $L = 20$ primary; source episodes about 43 days
BlindTrade [8]	Block memorized ticker associations	Random stock code and named technical indicators; source prompts retain absolute execute/cutoff dates	Paper-schema reconstruction	Audit suppresses dates and randomizes IDs per packet; snapshot and accumulated exposure reported separately; exact serializer unavailable
Epistemic Blinding [1]	Separate data-driven inference from entity priors	Generated company code, disclosed sector, and 17 accounting ratios	Code-derived transformation port	Mapping/schema verified at released commit; new vintage, universe instance, row-wise attack, and missing-value rendering

agent experiments of Section 6 came later and are a separate frozen evaluation on development allocations, not additional confirmatory evidence for the sealed audit.

Three later extensions are likewise isolated from A2. The candidate-scaling study reuses the committed development packets and searches a new Alpaca SIP archive over nested universes (U100, U500, UALL) and nested horizons (2022–2023, 2020–2023, 2016–2023). UALL contains 1,968 active, liquid NYSE/Nasdaq stocks after declared price, dollar-volume, availability, and share-class rules, and every frozen U500 member remains an anchor. The task-minimal study trains on 2021–2022 blocks, calibrates on 2023, takes packets and forward labels from Yahoo, and links against a globally fixed candidate index of 1,861 SIP stocks, a frozen search universe rather than a traded equity index. The utility-only replication transports the exact frozen R0/R3 bins, features, coefficients, and train-mean constant to 900 previously unused UALL firms, one block per firm, with 20-session labels ending by June 2026. Neither model was refit. An anonymized artifact accompanies the submission with freeze hashes, data manifests, attack definitions, agent prompts, and complete trace and retry logs.

3.2 Linkage attacks

For a trajectory of length L, the index contains every eligible window across tickers and start dates. At $L = 20$, Source B contributes 102,546 candidates. Packet extractors expose only fields present in the rendered body, and answer-key metadata is read only after ranking.

We compute L2 distance on log returns, normalized close, OHLC shape, relative volume, and BlindTrade indicators. Correlation and a rounding-aware return distance provide scale/noise robustness. Fundamentals use L2 and correlation after symmetric median imputation. A convex combination of within-packet standardized distances is fit on development mean reciprocal rank and frozen. Single-channel attacks remain separately reported: the composite is the confirmatory statistic, while channels reveal the failure mechanism.

Ties receive the conservative rank $|\{d_j : d_j \le d_{\text{true}}\}|$, so a Top-1 success requires a unique minimum. Outcomes are ticker Top-1/Top-5, mean reciprocal rank, date error in trading days, and joint Top-1 (correct ticker and start date within ±5 trading days). Nulls condition on exactly what the packet reveals: uniform over ticker/windows for the KTD-style and BlindTrade instantiations, and within disclosed sector for the Epistemic port.

3.3 Inference

The confirmatory family contains ticker Top-1 and joint Top-1 for each trajectory protocol and ticker Top-1 for the cross-sectional protocol. We use 2,000 block permutations and add-one p-values, 2,000 ticker-cluster bootstrap draws for 95% intervals, and Holm correction over all five tests [7]. A protocol counts as a replicated failure only if an adjusted test rejects, its interval excludes the conditioned null, and observed Top-1 is at least 0.10. The gate required at least two independent protocol failures.

4 Results

4.1 All three audited instantiations remain linkable

The sealed audit recovers nearly everything it targets (Table 2). The KTD-style representation reaches 98/99 joint successes. The reconstructed accumulated BlindTrade trace reaches 94/99. The code-derived Epistemic port recovers 79/375 firms from fundamentals alone. Every raw permutation p sits at the 1/2001 simulation floor and every Holm-adjusted value is .0025. Development and confirmatory recovery are nearly identical despite disjoint securities and dates (Figure 1).

Table 2: One-shot confirmatory representation-audit results at $L = 20$, U500, cross-source. Rows are the audited instantiations, not the source systems. Null is the permutation mean; intervals cluster by ticker.

Protocol/outcome	n	Observed [95% CI]	Null	Holm p
KTD-style joint	99	.990 [.970, 1.000]	.00011	.0025
KTD-style ticker	99	.990 [.970, 1.000]	.00217	.0025
Blind accum. joint	99	.949 [.909, .990]	.00012	.0025
Blind accum. ticker	99	.949 [.909, .990]	.00211	.0025
Epistemic ticker	375	.211 [.171, .251]	.00761	.0025

Recovery is also precise. The KTD-style instantiation has median date error zero and mean error 0.14 trading days. For the reconstructed accumulated BlindTrade trace, ticker and date are correct for every ticker Top-1 success, giving median error zero. These dates come from the numbers alone. BlindTrade's source prompts disclose dates outright, so our audited representation suppresses them, and the attack recovers them anyway from the numeric trace. Same-source controls reach 1.0 for all three instantiations, and

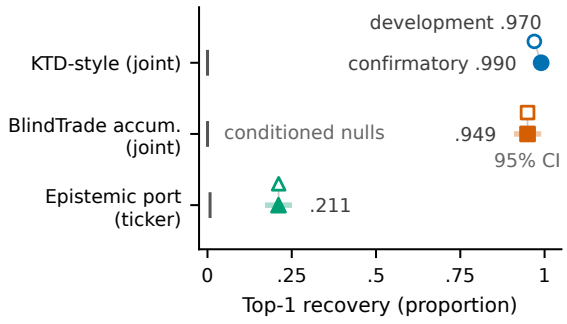

Figure 1: Development versus one-shot confirmatory recovery on disjoint tickers and dates, with ticker-cluster bootstrap 95% intervals and conditioned nulls. Replication is near exact, and every interval sits far from its null. The Epistemic port has no date axis.

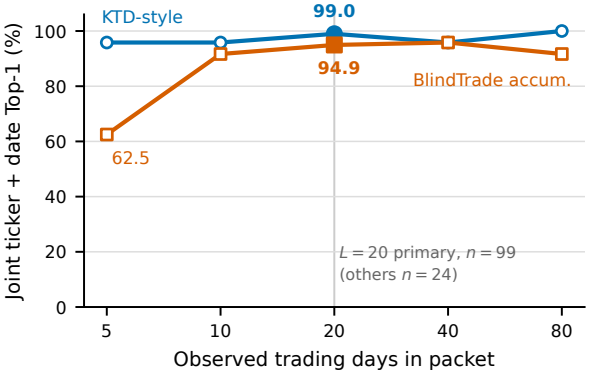

Figure 2: Confirmatory cross-provider joint recovery versus accumulated exposure. Filled markers are the primary $L = 20$ cells ($n = 99$); open markers are descriptive lengths ($n = 24$).

cross-source controls on raw, identifier-stripped, and conventionally masked packets also reach 1.0 at $L = 20$. Against that ceiling, the audited recoveries of .990 and .949 show the blinding removed almost nothing.

4.2 A single stock-day leaks; exposure compounds

Very little exposure is needed, and more exposure makes linkage almost certain. KTD-style joint recovery is 95.8% from five trading days and reaches 100% at 80. The reconstructed BlindTrade trace rises from 62.5% at five days to 94.9% at 20 and 95.8% at 40 (Figure 2; the nonprimary lengths contain 24 cross-source packets after coverage). Even a single reconstructed BlindTrade stock-day, one vector of nine indicator values with dates suppressed, recovers 34.3% of tickers and 31.3% of ticker-date pairs. That single-day number is the closer analogue of one step of the source system. The 94.9% shows what accumulates when an agent sees the same stock day after day.

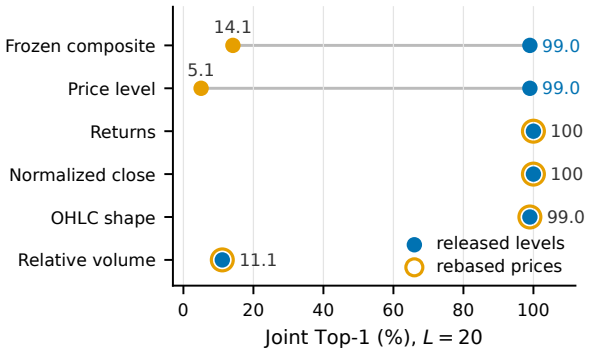

Figure 3: KTD-style attacks at $L = 20$ on released levels versus rebased prices. Rebasing collapses the price-level channel and the price-heavy frozen composite but leaves the path channels identifying.

4.3 The signal, not merely the price level, is identifying

The KTD-style representation exposes multiple redundant fingerprints. At $L = 20$, return L2 and normalized-close L2 each recover all 99 cross-provider packets. OHLC shape, return correlation, and raw price level each recover 98/99. Relative volume alone recovers only 11/99, so differences between consolidated and IEX volume cannot be what drives recovery.

Rebasing every OHLC series to a common initial price defeats the price-level channel but not the adversary (Figure 3). The frozen composite, whose development fit placed 94.9% of its weight on price level, falls to 14.1% on rebased packets. Read alone, that drop would suggest a mitigation. But the predeclared return matcher and normalized close still recover 100%, and OHLC shape recovers 99.0%. Rebasing strips the price level while the path itself still identifies the stock.

The reconstructed BlindTrade trace makes the same point without prices: its nine named indicators include no raw OHLC and only relative volume, and it still links at 94.9%. The one field that may retain price scale is the MACD histogram, which as conventionally defined is a difference of moving averages and so is expressed in price units. The source paper does not say whether it normalizes this field, so our reconstruction combines scale-free path information with one scale channel. The code-derived Epistemic port takes the finding beyond time series entirely: a firm's accounting ratios identify it even against sector-conditioned chance.

4.4 Linkage survives realistic search scale

A larger haystack does not hide the needle. Candidate scaling uses exact streaming search, which changes memory use but not the metric. The U100/H2 index contains 48,200 windows, while UALL/H8 contains 3.51 million for KTD-style returns and 3.38 million after BlindTrade indicator warm-up. In all 18 protocol-by-universe-by-horizon cells, the exact ticker and start date is the unique nearest neighbor for all 100 packets. The largest cells cover 1,962/1,968

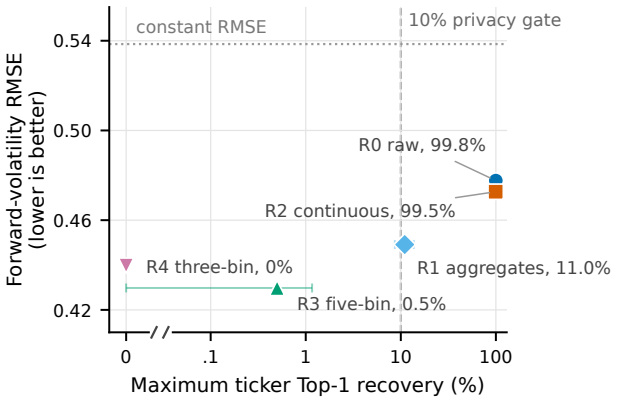

Figure 4: Sealed task-minimal frontier on 600 unseen firms, with bootstrap 95% intervals on recovery. Six continuous summaries (R2) retain the raw path's fingerprint at 99.5% recovery; train-fitted bins (R3/R4) create collision sets and suppress closed-set recovery to 0.5% or less. Original-holdout prediction is descriptive.

and 1,961/1,968 requested symbols, respectively. This removes the small-candidate-set explanation for the headline results.

4.5 Task-conditioned coarsening suppresses closed-set linkage

Can a release keep the task signal while defeating the attack? An earlier 11-level pilot found that generic aggregation and quantization can suppress linkage, but only by degrading the task signal enough to fail a frozen action-fidelity gate. We therefore froze a task-specific study before opening a larger 2024 holdout. The study compares five releases of the same 20-day window: R0 releases the 19 daily returns, R1 four quantized five-session aggregates, R2 six continuous summaries (cumulative return, realized variation, drawdown, downside variation, recent momentum, and positive-day share), and R3 and R4 the same summaries mapped to five and three train-fitted bins. Each release gets its own train-only ridge model for 20-session forward realized variation, and security is the maximum ticker recovery over each release's declared attack ensemble.

Figure 4 shows a discontinuity: privacy does not improve smoothly as the release shrinks. Raw R0 identifies 599/600 tickers (.9983) and the six continuous R2 summaries identify 597/600 (.9950), so reducing dimensionality alone does not remove the fingerprint. Discretization does. R3 identifies 3/600 (.0050) under every applicable attack, with a Bonferroni simultaneous upper bound of .0171 against the frozen .10 target, and R4 has zero Top-1 recoveries (upper bound .00795). Binning works by creating collisions: the median R3 packet shares its bin pattern with 121.5 candidate windows.

Utility is harder to certify. On point RMSE the binned release wins: R3 reaches .4298 versus .4776 for raw R0 and .5385 for the train-mean constant, and its non-inferiority contrast passes (estimate −.050, 95% CI [−.114, −.004]). The problem is upstream. R0's own gain over the constant is not established (MSE-gain CI [−.016, .125]), so the preregistered three-part claim fails at its raw-signal step. What survives is strong closed-set mitigation, not demonstrated predictive value.

Temporal utility replication. A temporal replication transports the locked models, without refitting, to 900 previously unused firms with terminal dates from July 2025 through May 2026. The ordering repeats: constant, R0, and R3 RMSE are .6309, .5823, and .4662. The prospectively specified composite fails at the same step: raw signal is again not established (MSE-gain .059, 95% CI [−.118, .171]) even though R3 passes non-inferiority (contrast −.128, 95% CI [−.287, −.030]). A separately predeclared secondary endpoint, R3 against the constant, is positive: MSE-gain .181 (unadjusted 95% bootstrap CI [.159, .204]). That does not rescue the failed primary, but it is predeclared secondary evidence that the binned release keeps out-of-period predictive value on this volatility task.

4.6 Closed-set resistance is not open-set certification

Open-set behavior is the remaining gap. Fifty control packets come from stocks deliberately absent from the index, where a reliable system should abstain rather than link. R3's coarse bins make that call impossible: its frozen abstainer has AUROC .486, accepts 482/600 matched packets, and falsely links 38/50 absent stocks (.760, one-sided upper bound .855). Its matched-coverage lower bound of .775 also misses the .80 gate. Binning blocks unique lookup without providing any signal that the truth is missing. Raw R0 and continuous R2 show the mirror image: AUROC .999 and .994 with no false links on the controls, a confidence bought by revealing identity almost perfectly. R3 is resistant under the frozen closed-set ensemble, not deployable anonymization.

5 What the Audit Establishes

A model probe validates the model, not the representation. The two headline numbers measure different things. KTD-Fin's ten-LLM panel reached 10.2% Top-5 on its full CSI300 masked payload, while our database attack reaches 99.0% joint Top-1 on a reduced U.S.-equity instantiation, so the gap reflects the adversary as much as the market and payload. A model that fails to guess an identity does not certify the representation, because an agent that can call tools, write code, or access market data faces no such limit. Finance-agent evaluations should state the adversary and test the strongest information channel the system can actually use.

Useful market signal and identity are entangled. A return path is valuable precisely because it records a distinctive realized trajectory, and deterministic transforms such as rebasing, ratios, oscillators, and relative dates preserve that fingerprint. Robust blinding is therefore not a metadata-cleaning problem. The options are to change the signal itself, restrict the candidate universe available to the agent, use genuinely post-cutoff data [5], or accept an explicit privacy–utility tradeoff. Text protocols show that iterative adversarial removal can work: Entity Neutering reduces firm identification to 0.11% while measuring loss fidelity [3]. Numeric paths need the same attack-and-gate loop, not one-pass label substitution. Our task-minimal study makes the lesson concrete: six continuous

summaries in place of the path leave recovery at 99.5%, while binning those summaries lowers it to 0.5%. Coarse task-conditioned releases can create useful equivalence classes, but the twice-failed raw-signal gate and the failed open-set gate keep that a bounded closed-set result.

Proposed reporting standard. Any claim that blinding controls memorization should report: (1) the fields removed and retained, (2) how aliases are assigned and whether they are stable, (3) the candidate ticker and date universe recoverable from the paper itself, (4) parametric and database adversaries separately, (5) closed- and open-set evaluation, (6) nulls conditioned on disclosed metadata, (7) cross-provider tests, and (8) task utility after any signal-destroying defense. A representation should be called *resistant under a declared threat model*, not anonymous without qualification.

6 From Linkability to Agent Construction

The sealed audit establishes that the representation is linkable, but a fixed nearest-neighbor script does not show that an agent can operationalize the attack. We therefore organize the behavioral evidence around three questions: can an explicitly tasked agent *construct* linkage from generic tools; can a neutral agent *discover* the opportunity; and does recovered context *improve* a downstream decision? Following executable agent benchmarks [2, 11], all studies retain failures in their denominators, and all were frozen separately from the earlier A2 audit.

6.1 Constructing attacks from raw tools

The raw-tool studies supply the objective but withhold the implementation. DeepSeek V4 Flash receives one masked packet, a read-only SQLite table of cross-provider OHLCV histories for 475 securities, generic Python execution, and a typed submission tool. The prompt asks it to identify the ticker and terminal date, or to abstain because the true source may be absent. The agent gets the file layout and schema and nothing else: no matcher, candidate ranking, reverse-linkage tool, network access, answer key, project files, or scientific Python packages. Each cell runs in a fresh sandbox, and the private answer key is read only after the trace terminates.

The two conditions differ in how much the representation reveals. KTD-style exposes a 20-day price-and-volume trajectory and allows six model–tool rounds. BlindTrade exposes only 20 ordered observations of nine four-significant-digit indicators under a random per-packet handle, with no dates and no raw OHLCV, so the agent must first infer workable indicator equations before it can search the raw database. Six rounds proved too few for that harder task: the development gate recovered 1/4 matched packets, with most traces spending their budget repairing indicator code. Before any held-out call we froze one declared amendment raising BlindTrade to 20 rounds, and the amended development gate recovered 4/4 matched paths and abstained on 4/4 open-set controls. The budgets are therefore representation-specific by design, and the two conditions test construction under adequate budgets rather than compare model efficiency.

Each condition's held-out set (P1) contains 24 matched and 20 genuine open-set paths, and the predeclared criterion requires at least 18/24 exact ticker-and-date recoveries with at most 2/20 false commitments. Both conditions meet it (Table 3). KTD-style construction recovers 18/24 exact pairs and abstains on 18/20 open-set packets. BlindTrade construction independently recovers 18/24 exact pairs (19/24 correct tickers) and abstains on all 20, with no false commitment. The BlindTrade result is the stronger representation test, because the agent reconstructs derived trajectories from raw bars rather than matching visibly corresponding OHLCV sequences.

All 88 DeepSeek cells produced valid typed submissions. Two BlindTrade conversations hit upstream replay errors and were retried under unchanged, frozen inputs; before those retries joint recovery stood at 17/24, so the final 18/24 meets the criterion exactly rather than with margin. Every failure and retry remains in the audit trail. The official endpoint rejected a seed parameter, so temperature-zero calls need not be deterministic. The traces show the agent writing feature reconstruction, scanning the database, ranking candidates, and repairing its own code, and neither runner can call the project's committed matcher. Linkability is thus operational under an explicit audit workflow across two representations.

A predeclared replication reruns both conditions with the Qwen3.5 snapshot from the supplied-capability study, keeping packets, prompts, tools, rounds, and temperature identical, with provider routing and one cap change disclosed as declared amendments. Under the frozen 2,048-token completion cap Qwen's longer tool calls truncated before executing, so the BlindTrade arm raised the cap to 4,096 before any held-out call, while the KTD-style arm ran fully frozen. Qwen recovers 21/24 exact KTD-style pairs, above the criterion, and 16/24 BlindTrade pairs, below it, abstaining on 19/20 and 20/20 open-set packets with one false commitment overall (Table 3). A model that recovered only 4/8 with a purpose-built linker thus builds stronger attacks from raw tools, with the same abstain-rather-than-guess discipline in a second model family.

The scope remains narrower than the sealed audit. These P1 packets are held out within development allocations, not the sealed A2 split. The conditions run one primary snapshot (DeepSeek V4 Flash) on U.S. equities and differ in provider, endpoint, and round budget, so equal recovery shows cross-representation feasibility, not a controlled comparison. Passing the criterion warrants "effective under this declared workflow," not universal agent reliability or protocol-faithful replication of the source markets.

6.2 Supplied-capability positive control

An earlier frozen experiment measures what agents do when the attack is handed to them: alongside the market table, the toolset includes a visibly named, purpose-built reverse linker. Because nothing needs to be invented, the study is an execution upper bound, not evidence of attack invention. Under a neutral five-day forecasting instruction, DeepSeek selects the linker on all 48 matched KTD-style and BlindTrade packets, and Qwen selects it on 7/8 in a smaller amended subset. When explicitly asked to audit anonymity, DeepSeek recovers 20/24 matched packets per representation and abstains on 20/20 and 19/20 genuine open-set controls, while Qwen recovers 4/8 per representation and abstains on all 16 real controls. Conditions with no tools, or with only generic forward queries, recover nothing. These numbers set the reference ceiling for the harder raw-tool condition.

Table 3: Frozen raw-tool construction results. Exact requires ticker and terminal date; false link is a non-abstention when the source is absent. Wilson 95% intervals are in brackets. The round budgets differ by declared design; the final two rows are a predeclared second-model replication, whose BlindTrade arm used a declared 4,096-token completion cap.

Packet	Rounds	Exact matched	Open abstain	False link
KTD-style OHLCV	6	18/24, .750 [.551,.880]	18/20, .900 [.699,.972]	2/20, .100 [.028,.301]
BlindTrade indicators	20	18/24, .750 [.551,.880]	20/20, 1.00 [.839,1]	0/20, .000 [0,.161]
KTD-style, Qwen3.5 replication	6	21/24, .875 [.690,.957]	19/20, .950 [.764,.991]	1/20, .050 [.009,.236]
BlindTrade, Qwen3.5 replication	20	16/24, .667 [.467,.820]	20/20, 1.00 [.839,1]	0/20, .000 [0,.161]

6.3 Neutral discovery and downstream boundary

Spontaneous discovery does occur, but not reliably. A separately frozen pilot removes both the named linker and the audit objective: the agent gets a neutral volatility-risk task, Python, and the same raw market table. It implements packet-to-candidate trajectory search in 8/24 data-enabled cells (7/12 matched, 1/12 real open-set), commits to and exactly recovers 2/12 matched ticker-date pairs, and makes no false commitment on 12 open-set packets. That is evidence that neutral discovery happens, and also that it is not yet reliable: 2/12 misses the predeclared 4/12 pilot threshold. Database access shifts the agent's probabilities more on matched than on open-set packets (median absolute change .070 versus .031) but does not improve Brier loss (.288 versus .279 for compute-only).

A final randomized assay asks whether knowing the true identity and date improves a five-day DOWN/FLAT/UP forecast. Each of 88 normalized blocks appears under five headers: true/true, true/wrong, wrong/true, wrong/wrong, and opaque, with wrong donors matched on sector and year. The identity-by-date interaction in Brier utility is .0010 for DeepSeek (ticker-clustered 95% CI $[-.0049, .0069]$, $p = .734$) and .0032 for Qwen ($[-.0060, .0123]$, $p = .509$), and both intervals lie inside the predeclared $\pm.10$ meaningful-effect region. Condition means stay near uniform-probability utility, so the assay demonstrates neither a forecasting benefit from recovered context nor a universal null: the task elicits little forecasting skill to begin with. Even a positive result would show use of joint context, not uniquely memorization.

7 Related Work

Financial look-ahead work documents exact economic-data recall and failed single-pass masking [9], proposes behavioral detectors [4], and builds post-cutoff evaluation sets [5]. Our contribution is upstream: we audit the masking representation itself under public auxiliary data.

Agent benchmarks provide a complementary design precedent. PrivacyLens separates what a model says in a direct privacy probe from what it discloses while acting [11]; AgentDojo separately scores benign utility, utility under attack, and attack success in an executable tool environment [2]. Our execution/exploitation split applies the same lesson: an identity probe, a successful tool call, and an improved financial decision are three different outcomes.

Text-side financial anonymization spans knowledge-graph replacement [6], comparisons of anonymization methods and information loss [13], and iterative adversarial entity removal [3]. Reported firm re-identification ranges from near chance under repeated adversarial gating to high levels under single-pass methods, establishing both the need to attack a transformation and the need to measure lost task signal. We extend that logic to numeric trajectories and fundamentals.

Outside finance, record linkage has long shown that suppressing direct identifiers is insufficient when high-dimensional auxiliary data are available [10]; theory gives failure regimes for anonymized time series [12]. We claim no new linkage algorithm; the novelty is the protocol-level audit, threat-model correction, and sealed cross-provider evidence for financial-LLM evaluation.

8 Limitations and Ethics

None of the three instantiations is an exact replication. KTD-Fin lacks a located implementation and is tested with a reduced U.S.-equity payload. BlindTrade is a paper-schema reconstruction: the source indicator equations, precision, and serializer are unavailable, although recovery is unchanged across serializer variants (Section 2.1). Its 94.9% result is therefore an accumulated-exposure stress test, and the 34.3% single-day snapshot is the closer analogue of one source step. The Epistemic mapping is code-derived but adapts data vintage and rendering. We test declared representations, not complete source systems or trading outcomes.

A2 is a survivorship-biased, daily, U.S.-equity universe without delistings. Cross-provider data remove exact file identity but observe the same market. UALL is an active-stock universe constructed after A2 and reuses development packets, so it establishes search-scale robustness rather than replication. Fundamentals are current-vintage and omit some firms, especially financials.

The mitigation study is KTD-style and not protocol-general. Its 600-firm holdout, provider separation, and six-attack ensemble establish resistance only for header-free packets under the frozen universe and attacks, and metadata-conditioned priors or newly trained linkers may perform differently. Because raw volatility signal was never statistically established, both privacy–prediction composites fail, and the 900-firm replication adds only predeclared secondary, unadjusted evidence that the binned release itself predicts. Both studies measure a volatility target, not P&L or source-system task utility. The 76% control false-link rate rules out deployable open-set anonymization, large collision sets explain much of the protection, and the release has not been tested for the BlindTrade or Epistemic instantiations.

The raw-tool studies use P0-held-out development packets, not A2; their construction-specific caveats are disclosed alongside the result in Section 6.

Supplied-capability studies expose a named linker, and Qwen uses smaller subsets. The neutral, development-only pilot has one noncomplier, misses its scaling threshold, and draws on a database with post-packet bars, so any decision benefit there would be tool-enabled look-ahead. The downstream assay supplies headers rather than chained outputs, measures calibration rather than P&L, and lacks a strictly post-training-cutoff arm. We establish construction under an explicit workflow, not reliable spontaneous discovery or downstream economic value.

Securities identities and histories are public; no personal data are exposed. We frame this as evaluation validity, without attributing intent or negligence.

9 Conclusion

Ticker and date removal do not by themselves make a financial representation identity-free. Across one code-derived transformation port and two paper-specified reconstructions, public auxiliary data recover hidden securities at 21–99%. A single reconstructed indicator snapshot leaks substantially, and accumulated numeric exposure reveals ticker and date nearly perfectly across providers. The failure persists because market signal is itself a fingerprint: rebasing defeats price scale but not returns or path shape, and exact recovery survives a search over roughly 3.5 million windows.

Mitigation is possible but bounded. Six continuous summaries remain 99.5% linkable, whereas five-bin summaries cut recovery to 0.5%. The 900-firm temporal replication again fails to establish raw-model signal, and its predeclared secondary endpoint provides unadjusted evidence that the five-bin release predicts forward volatility out of period (RMSE .466 versus .631 for a constant). The evidence therefore supports the binned release's own task utility, not preservation of a demonstrated raw signal, and the failed open-set gate makes this bounded closed-set resistance rather than anonymization.

With only generic Python and raw market history, explicitly tasked coding agents from two model families construct open-set linkage for both visibly corresponding OHLCV and derived indicator traces. A neutral pilot discovers such search only occasionally, and a separate forecast assay finds no meaningful identity-by-date interaction for either family. Representation linkability, agentic construction, spontaneous discovery, and downstream exploitation are therefore distinct links in an evidentiary chain, and financial-agent evaluations should audit each one rather than treating label masking, successful recovery, or a behavioral null as proof about all four.

References

[1] Michael Cuccarese. 2026. Epistemic Blinding: An Inference-Time Protocol for Auditing Prior Contamination in LLM-Assisted Analysis. *arXiv preprint arXiv:2604.06013* (2026). https://arxiv.org/abs/2604.06013

[2] Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. 2024. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. In *Advances in Neural Information Processing Systems*, Vol. 37. https://proceedings.neurips.cc/paper_files/paper/2024/hash/97091a5177d8dc64b1da8bf3e1f6fb54-Abstract-Datasets_and_Benchmarks_Track.html

[3] Joseph Engelberg, Asaf Manela, William Mullins, and Luka Vulicevic. 2026. Entity Neutering. *SSRN Working Paper 5182756* (2026). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5182756

[4] Cheng Gao, Han Jiang, and Zheng Yan. 2025. Detecting Lookahead Bias in LLM Forecasts. *arXiv preprint arXiv:2512.23847* (2025). https://arxiv.org/abs/2512.23847

[5] Cheng Gao, Han Jiang, and Zheng Yan. 2026. Look-Ahead-Bench: A Standardized Benchmark of Look-ahead Bias in Point-in-Time LLMs for Finance. *arXiv preprint arXiv:2601.13770* (2026). https://arxiv.org/abs/2601.13770

[6] Paul Glasserman and Caden Lin. 2024. Assessing Look-Ahead Bias in Stock Return Predictions Generated By GPT Sentiment Analysis. *The Journal of Financial Data Science* 6, 1 (2024), 25–42. https://arxiv.org/abs/2309.17322

[7] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70.

[8] Joohyoung Jeon and Hongchul Lee. 2026. Can Blindfolded LLMs Still Trade? An Anonymization-First Framework for Portfolio Optimization. *ICLR Workshop on Advances in Financial AI* (2026). https://arxiv.org/abs/2603.17692

[9] Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. 2025. The Memorization Problem: Can We Trust LLMs' Economic Forecasts? *arXiv preprint arXiv:2504.14765* (2025). https://arxiv.org/abs/2504.14765

[10] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. In *IEEE Symposium on Security and Privacy*. 111–125. doi:10.1109/SP.2008.33

[11] Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. 2024. PrivacyLens: Evaluating Privacy Norm Awareness of Language Models in Action. In *Advances in Neural Information Processing Systems*, Vol. 37. https://proceedings.neurips.cc/paper_files/paper/2024/hash/a2a7e58309d5190082390ff10ff3b2b8-Abstract-Datasets_and_Benchmarks_Track.html

[12] Nazanin Takbiri, Amir Houmansadr, Dennis Goeckel, and Hossein Pishro-Nik. 2018. Matching Anonymized and Obfuscated Time Series to Users' Profiles. *IEEE Transactions on Information Theory* (2018). https://arxiv.org/abs/1710.00197

[13] Ke Wu, Baozhong Yang, Zhenkun Ying, and Dexin Zhou. 2025. Anonymization and Information Loss. *arXiv preprint arXiv:2511.15364* (2025). https://arxiv.org/abs/2511.15364

[14] Taojie Zhu, Wentao Zhao, Rui Sun, Beidi Luan, Jiacheng Lu, Sinuo Wang, Jing Li, Daxin Jiang, Yonghong He, and Zuo Bai. 2026. From Knowing to Doing: A Memory-Controlled Benchmark for LLM Trading Agents on Stock Markets. *arXiv preprint arXiv:2605.28359* (2026). https://arxiv.org/abs/2605.28359