DISTANT MEMORIES: HOW DO LLMS REMEMBER FINANCIAL DATA?

Anonymous authors
Paper under double-blind review

ABSTRACT

What do large language models remember about historical prices, and how can they access that knowledge? We ask models for the prices of financial instruments on particular dates, without tools or external data. A comparison of 60 language-model configurations on 8,096 questions shows large differences across models, instruments, and periods. Queries for consecutive days reveal detail that aggregate error can hide: some models reproduce daily variation in well-known series, while their answers about other instruments are little better than a smooth approximation. We also test repeated sampling, alternative question wordings, and the reverse question of identifying an instrument from its prices. Dates help substantially with reverse lookup, including when absolute price levels are removed. Finally, we examine financial text in public corpora. More frequent price-context mentions are associated with better recall.

1 INTRODUCTION

Large language models know a lot, including a lot of numerical facts such as historical prices of equities, indexes, currencies, and other instruments (Lopez-Lira et al., 2025). Do they seem to *understand* the data or have they just memorized the numbers? How can they access and reason with this data? How do they store the data: individual numbers, a compressed series, a fitted curve?

We set out to test this by asking language models questions. We are trying to understand what knowledge is in the weights, not what the model can do when given access to the web or a database. The recall experiments therefore query models without tools; a separate search-enabled control is reported in the appendix.

Our key motivation is to use knowledge of financial data as a lens to understand how LLMs store and retrieve information, in particular numerical information. But we are also interested in this from the perspective of LLMs reasoning about financial matters, such as trading.

One might reasonably wonder why it would matter that a model knows stock prices, when it could just access them with a tool call. Would it not be enough with a small cognitive core that gets its information from the outside world whenever it needs it? But using tools takes time. To reason about a question, it helps to have facts readily available, even if they are approximate and need to be checked before acting on the result. For example, to evaluate today's market action, it helps to have a baseline expectation of a reasonable level for Tesla stock, USD/EUR, or 10-year Treasury yields, and which of these is usually most volatile. More fundamentally, you need to have an idea about which instruments there are, what they are typically worth, and perhaps how they relate to each other to even begin to form hypotheses that can be evaluated by querying fact sources or running simulations. You need to know things to be able to ideate. For these reasons, we think it matters to what extent the model knows financial facts even before using tools.

We first ask how accurately models can recall a price given an instrument and a date. We compare accuracy with the models' stated confidence, look more closely at their answers for consecutive days, and ask how much the answers change across repeated queries and different wordings. Next, we reverse the question: given prices and dates, can a model identify the instrument? Finally, we examine public corpora to see what kinds of financial information they contain and how their coverage relates to recall.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

```
You are a participant in a research study measuring how accurately
language models recall historical financial data seen during
training. Answer purely from memory - do not use tools, search, or
calculations from other data.

Rules:
1. You MUST always provide a single numeric point estimate, even if you
    are very unsure. This is an estimation study: a rough guess carries
    information; a refusal does not. "I don't know" or a range is not an
    accepted answer.
2. Express uncertainty through the confidence field, never by refusing.
3. Reply with ONLY a single-line JSON object, no other text: {"price": <
    number>, "confidence": "low" | "medium" | "high"}
```

```
What was the closing price of Apple stock (ticker: AAPL) on 1997-04-03?
    Give the price in USD, as it was actually quoted on that day (do NOT
    adjust for stock splits that happened after that date).

Reply with only: {"price": <number>, "confidence": "low" | "medium" | "
    high"}
```

Figure 1: The system prompt for price recall and an example equity question. The model must give a point estimate and report uncertainty separately. Other instrument kinds use the same system prompt, with user questions specifying the relevant units and quoting conventions.

2 METHODOLOGY

We selected models spanning several providers and sizes, including both closed and open weights. The main comparison contains 60 language-model configurations and two decision-model baselines; thinking settings and local quantized versions count separately. Appendix A lists their coverage across experiments. Most cloud models were accessed through OpenRouter, Claude models through Anthropic's API, and local models through Ollama. The collection code and question generators were developed with assistance from LLMs.

1. We collect and cache historical values from Yahoo Finance and FRED. For equities, we reconstruct prices as quoted at the time by reversing subsequent split adjustments.
2. We generate questions deterministically from these records. The comparison set samples dates within five-year eras for each instrument; the dense set asks about every observation day in a selected window.
3. We query each model and save its answer, raw response, and any returned reasoning trace. The ground truth is used for scoring and is never shown in price-recall prompts.
4. We compare the answers with the corrected reference data. New questions motivated by the results are collected as separate experiments.

The prompt requires a numeric point estimate even when the model is uncertain, with uncertainty reported as `low`, `medium`, or `high` confidence (Figure 1). Question wording varies by instrument kind to specify units and, for equities, the as-traded price convention. If an answer cannot be parsed, the runner makes one retry with a stricter format instruction. A number extracted from an unfinished reasoning trace is not treated as a valid answer.

We found several problems in the source data during early testing. In some cases, frontier models agreed with each other and disagreed with our reference values by a simple factor. For example, Toyota had a 5-for-1 stock split in 2021, but Yahoo's split feed listed it twice, causing our initial reconstruction to apply the factor twice. Other issues included currency-unit mismatches, duplicate or spurious splits, a corrupted Time Warner history, and a Bed Bath & Beyond ticker reassigned to a different company. We corrected the records or excluded the affected questions. Model agreement helped us find these cases; we then checked the underlying records. All price results reported here rejoin the saved answers to the corrected question files.

Table 1: Question sets used in this paper. Counts refer to distinct questions, before repetitions or model configurations. Supplementary experiments use the separate panels described in the appendix.

Question set	Questions	Instruments	Dates
Price comparison, corrected	8,096	145	1928–2024
Dense daily panels	12,787	16	1983–2024
Reverse lookup, each of four conditions	480	16	1928–2024

Table 2: Price recall on the corrected comparison set ($N = 8,096$). We show the ten highest-scoring configurations and five smaller or lower-scoring models. Median absolute log error is in log units; accuracies and their confidence intervals are percentages. Intervals resample whole instruments. Accuracy is conditional on a valid answer; all 62 configurations and their coverage appear in Appendix A.

Model	n valid	Med. $\lvert \ln \rvert$	Within 5% [95% CI]	Within 2×
Claude Fable 5	8,090	0.027	62.6 [56.8, 68.3]	96.9%
GPT-5.6 Sol	8,096	0.031	59.8 [54.5, 65.2]	94.8%
Gemini 3.6 Flash	8,095	0.033	58.4 [52.8, 64.0]	96.5%
Gemini 3.1 Pro	8,096	0.033	57.5 [51.7, 63.3]	95.3%
Gemini 3 Flash	8,093	0.037	56.0 [50.3, 61.7]	93.7%
Claude Opus 5 (think)	8,094	0.038	55.8 [49.9, 61.7]	94.5%
Claude Opus 5 (no-think)	8,096	0.038	55.6 [49.9, 61.2]	94.4%
Grok 4.7	8,096	0.039	54.9 [49.0, 60.7]	94.8%
Kimi K3 (think)	8,095	0.059	46.5 [41.3, 52.0]	93.0%
Kimi K3 (no-think)	8,089	0.070	44.4 [39.3, 49.4]	89.9%
DeepSeek v3	8,095	0.116	36.0 [31.3, 40.8]	82.4%
Llama 3.3 70B	8,096	0.155	28.8 [24.8, 33.0]	80.1%
Qwen3.5 122B-A10B	8,092	0.243	21.3 [17.6, 25.4]	78.1%
Qwen3.5 9B	7,993	0.478	10.3 [8.4, 12.4]	60.1%
Gemma 3 270M (q4)	7,732	1.457	1.8 [1.2, 2.3]	25.1%

Our primary error measure is $|\ln(\hat{p}/p)|$, where $\hat{p}$ is the answer and p the true value. It treats an overestimate by a factor of two and an underestimate by a factor of two equally. We report this metric in log units. The separate "within 5%" measure uses $|\hat{p} - p|/p \leq 0.05$; a log error of 0.05 is only approximately a 5% percentage error. Price metrics use valid positive numeric answers, and we report their number so that failures are visible. For accuracy in the main comparison, we give 95% intervals by sampling the instruments from the questions that will be used. We repeat this process 5,000 times and use the accuracy output to compute the confidence interval. These rules don't apply to the supplementary experiments provided in the appendices; these experiments' instruments, models, and scoring rules are described in the appendix.

3 Do the models know historical prices?

The first and most fundamental question is to what extent models know prices of various instruments. The canonical comparison has 8,096 corrected questions across 145 instruments: 95 equities, 11 indexes, 6 rates, 11 foreign exchange pairs, 8 commodities, 6 exchange-traded funds, and 8 cryptocurrencies. Every configuration has a recorded attempt on every question; valid-answer counts are reported separately. Jev 1.13 and OpenJev 0.1 are decision-model baselines: their numerical scores are mapped onto an instrument's historical price range supplied by the experiment. These baselines receive price-range information unavailable to the language models. The full roster and instrument universe appear in Appendix C and Appendix D.

The main results are summarized in Table 2. Models do remember historical prices, but the accuracy varies considerably. Claude Fable 5 answers 62.6% of valid questions within 5% of the true value, compared with 59.8% for GPT-5.6 Sol and 58.4% for Gemini 3.6 Flash. These percentages are conditional on a valid answer; the respective counts are 8,090, 8,096, and 8,095. At the other end of the size range, the smallest models often repeat a small set of plausible-looking numbers across very different instruments.

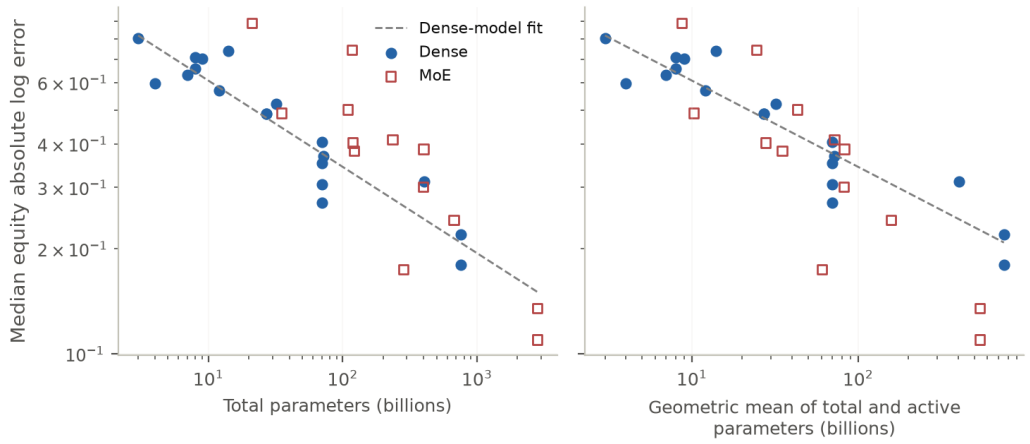

Figure 2: Equity recall error on the corrected comparison set, against total parameter count (left) and the geometric mean of total and active counts (right). Circles are dense models and squares are mixture-of-experts models. The dashed line is a log–log fit to the dense models. Only configurations with parameter counts in the collection rosters are shown.

Among open models with recorded parameter counts, larger models generally have lower equity recall error (Figure 2). Training data and procedures also vary across these models. Active parameter count alone does not explain the mixture-of-experts results: Llama 4 Scout and Maverick have the same recorded active count but different recall accuracy. Plotting the geometric mean of total and active parameters brings many sparse models closer to the dense-model trend. Newer versions are not invariably better than their predecessors.

The differences between instrument kinds are also large (Figure 3). For Claude Opus 5 with thinking, median absolute log error is 0.005 for indexes and 0.004 for foreign exchange, compared with 0.070 for equities. Equities include a much wider range of fame and difficulty. Even within that category, a model can know an instrument approximately without knowing its exact daily closes: Opus has a median absolute percentage error of 13.7% on Lakeland Industries, an obscure small-cap stock.

Recall is generally better for recent dates. GPT-5.6 Sol's median absolute percentage error is 1.2% in 2020–2024 and 4.9% in 1960–1964; DeepSeek v3's corresponding errors are 5.2% and 41.6%. Older eras have fewer instruments, so the pooled curves combine recency with changes in instrument composition. Appendix Figure 8 holds instrument identity fixed.

Do the models know when they know an answer? Several of the most accurate models mostly report low confidence, while some much weaker models frequently report medium or high confidence (Figure 4). Within a model, however, higher-confidence answers are often substantially more accurate: Claude Fable 5 is within 5% on 97% of the answers it labels medium and on 57% of those it labels low, and the corresponding figures for Llama 3.3 70B are 48% and 15%. Confidence thus helps distinguish a model's better answers, even though the labels mean different things across models. These are ordinal self-reports, so we assess their relationship with accuracy rather than treating them as probabilities. The breakdown by decade shows more medium- and high-confidence answers on recent questions (Appendix F.5).

Do models know individual daily values or only the rough shape of a series? Our *dense* set contains 12,787 observations from 16 instruments, each over a three-year window: for example, Apple in 1983–1986, the S&P 500 in 2007–2010, and Bitcoin in 2017–2020.

We interpolate linearly in log price between every 21st valid answer and, separately, between the true prices at those anchors. We compare these curves with daily answers on the same non-anchor days (Appendix B). The residual correlation, r, measures agreement between the model's and the true series' deviations from their respective interpolated curves.

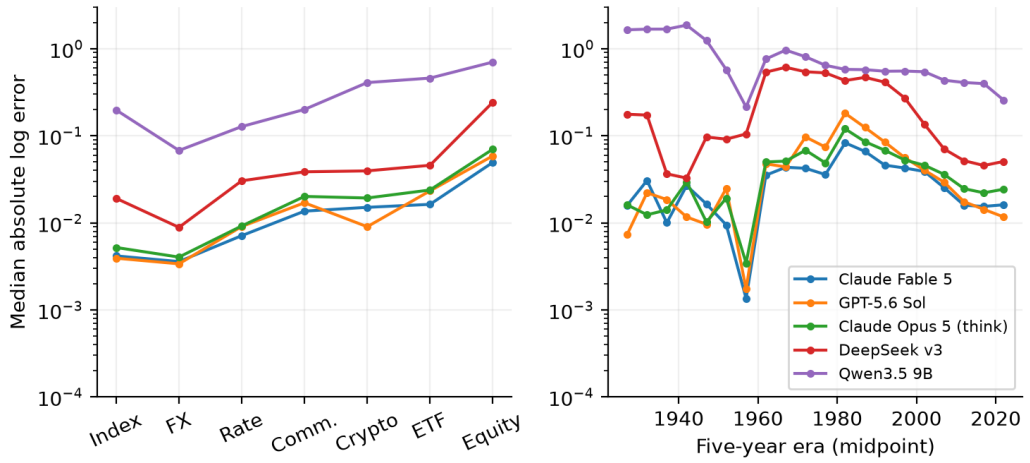

Figure 3: Median absolute log error by instrument kind and five-year era on the corrected comparison set. The curves describe the available questions in each group, rather than a balanced panel of instruments across eras. The dip of the strongest models around 1955 is a composition effect, since only rates and indices have questions there; Appendix Figure 8 holds the instrument set fixed.

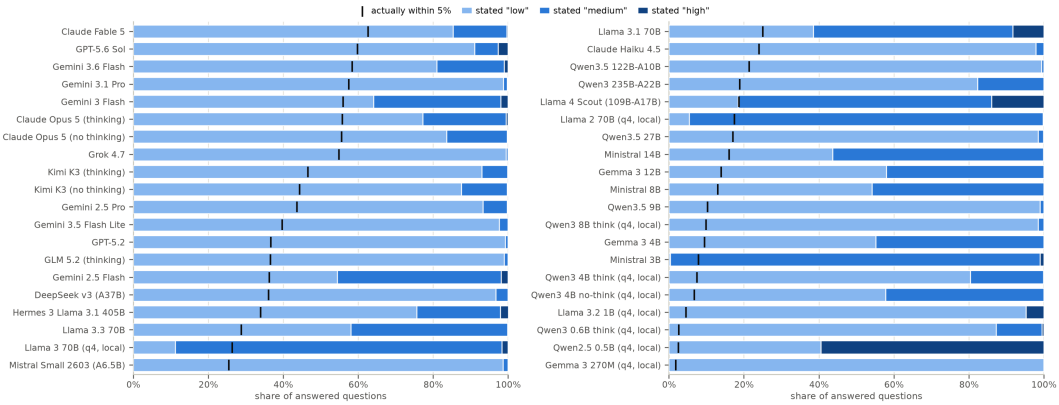

Figure 4: Stated confidence versus delivered accuracy for forty configurations, ordered by accuracy. Bars show the mix among answers with a parsed confidence label; black ticks show the share of numeric answers within 5% of truth. The confidence analysis uses the canonical questions and the scoring conventions in Appendix F.5.

Figure 5 show three regimes of recall, named for the patterns in the answers (check appendix figure 22 for additional graphs that shows these patterns):

- **Lookup Table:** Daily answers closely reproduce the true price path, including detail that log-linear interpolation from the true monthly anchors misses.
- **Anchor Points:** Interpolating the model's own answers at roughly monthly spacing changes the error very little; the daily queries add little detail between those anchors.
- **Coarse Level Prior:** Answers fluctuate around a rough price level, and smoothing improves them because much of their daily variation is error.

All three regimes can occur within one model, depending on the instrument and period. For example, Fable's residual correlation is approximately 1.00 on the S&P 500 and 0.87 on Tesla, but near zero on Air T.

Accurate recall can also coexist with systematic rounding. On US-dollar equity questions from 2002 onward, the five Anthropic configurations put 53–77% of their answers at prices ending in 50 cents,

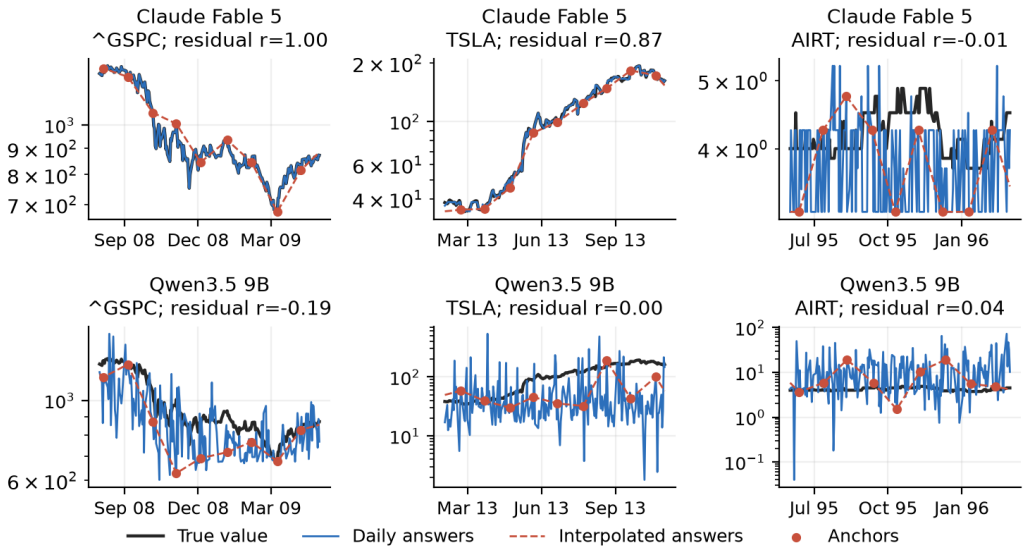

Figure 5: Lookup Table, Anchor Points, and Coarse Level Prior regimes. Daily answers and interpolation between the model's own answers at 21-observation intervals. Plots show selected portions of the three-year windows; statistics use all eligible non-anchor days in each window. The top row shows Fable recalling daily detail for the S&P 500 and Tesla, but little such detail for Air T, where its daily answers alternate between a few round price levels. In the bottom row, smoothing Qwen3.5 9B's answers reduces error on all three series. The reported r is the correlation of model and true residuals around their respective interpolated curves.

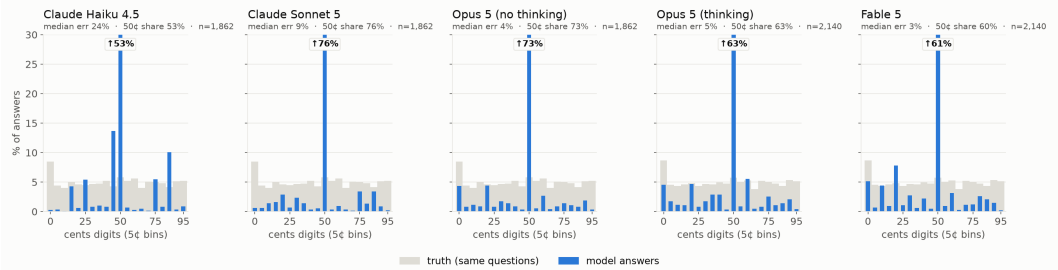

Figure 6: Cents digits for five Anthropic configurations on the same 2,140 US-dollar equity questions from 2002 onward. Blue shows model answers; grey shows the true values. Each model concentrates many answers at 50 cents. These are memory-only runs.

whereas the true prices end that way only 1.8% of the time (Figure 6). The preference for round numbers persists even in models that recall the price path accurately.

4 Reverse Lookup

If we give a model prices and dates, will it know which instrument they belong to? This asks whether information that can be retrieved in one direction can also be accessed in the other. We use 16 candidate instruments and show $k \in \{1, 3, 5, 10, 50\}$ prices sampled from a one-year window. There are six windows per instrument at each k, giving 480 questions per condition. The model chooses from all 16 instruments, so uniform random choice would achieve 6.25% accuracy. Units are omitted because they would often reveal the instrument.

We test four conditions on matched traces:

- Raw prices with a date for each observation.

Table 3: Instrument identification among 16 candidates (chance: 6.25%). There are 96 planned questions at each trace length. Raw prices retain their level; normalized prices are divided by the first price. Dates are either included for every point or omitted. Missing collection coverage is reported in Appendix A; invalid answers count as incorrect among recorded questions.

Model	Raw, dates $k = 1$	Raw, dates $k = 10$	Raw, no dates $k = 10$	Norm., dates $k = 10$	Norm., no dates $k = 10$
Claude Opus 5 (think)	88.5%	94.8%	53.1%	68.8%	16.7%
Claude Fable 5	86.5%	93.8%	49.0%	75.0%	8.3%
GPT-5.6 Sol	85.4%	92.7%	66.7%	78.1%	18.8%
Grok 4.5	78.1%	90.6%	46.9%	66.7%	9.4%
Gemini 3.6 Flash	77.1%	84.4%	39.6%	65.6%	11.5%
Grok 4.6	77.1%	87.5%	47.9%	76.0%	6.2%
Grok 4.7	76.0%	87.5%	55.2%	69.8%	9.4%
Qwen3.5 122B-A10B (fp8)	67.7%	77.1%	38.5%	32.3%	10.4%
DeepSeek v4 Flash	65.6%	74.0%	20.8%	29.2%	4.2%
Qwen3.5 122B-A10B (bf16)	65.6%	79.2%	37.5%	31.2%	6.2%
Qwen3.5 122B-A10B	62.5%	79.2%	39.6%	31.2%	10.4%
Qwen3.5 122B-A10B (fp4)	56.2%	72.9%	37.5%	24.0%	9.4%
Qwen3.5 35B-A3B	55.2%	62.5%	26.0%	14.6%	9.4%
DeepSeek v3	16.7%	31.2%	21.9%	11.5%	8.3%
Qwen3.5 9B	14.6%	15.6%	9.4%	13.5%	8.3%
Llama 3.1 70B	6.2%	14.6%	13.5%	13.5%	5.2%

- The same raw prices in chronological order, with dates omitted.
- Dated prices divided by the first price in the trace, so the first value is 1 and absolute price level is removed.
- The same normalized values with dates omitted.

The undated observations are not necessarily from consecutive days. The four conditions are paired at a given k; different values of k use separately sampled windows, rather than progressively extending the same trace. Normalization preserves relative changes, including any stock-split jumps. At $k = 1$, a normalized trace contains only the value 1; with dates shown, the date and the candidate list can still provide information.

Frontier models are good at identifying instruments from a single dated price. Opus with thinking reaches 88.5%, Fable 86.5%, and Sol 85.4% across all 96 questions at $k = 1$, counting invalid answers as incorrect. At $k = 10$, their accuracies are 94.8%, 93.8%, and 92.7% (Table 3). The candidate set supplies useful clues: a price in the thousands already excludes many instruments.

Removing the price level makes the dates especially useful. With ten normalized points, Sol identifies 78.1% of instruments when dates are included and 18.8% when they are omitted; the corresponding Opus figures are 68.8% and 16.7%. Removing dates sharply reduces accuracy, even though some models remain above chance. With only three normalized dated points, Sol and Opus already reach 67.7% and 64.6%. Dates locate a pattern in history and supply clues about instrument lifetimes and stock splits.

5 Data provenance

More famous instruments tend to be remembered better. A natural explanation is that models have seen more information about them during training, perhaps including more complete price histories. But what kind of information is it? Do the prices come from tables, newspaper articles, reports, or something else?

Most model providers do not release their full training data, so we cannot answer this directly for the strongest models. Public corpora nevertheless let us test whether financial coverage is associated with recall. We use infini-gram (Liu et al., 2024) to count price-context phrases in Dolma v1.7, such as "shares of {name}" and "{name} stock." For each equity, we take the larger of these two counts. This avoids some of the ambiguity of bare tickers, although the phrases need not state a price or refer

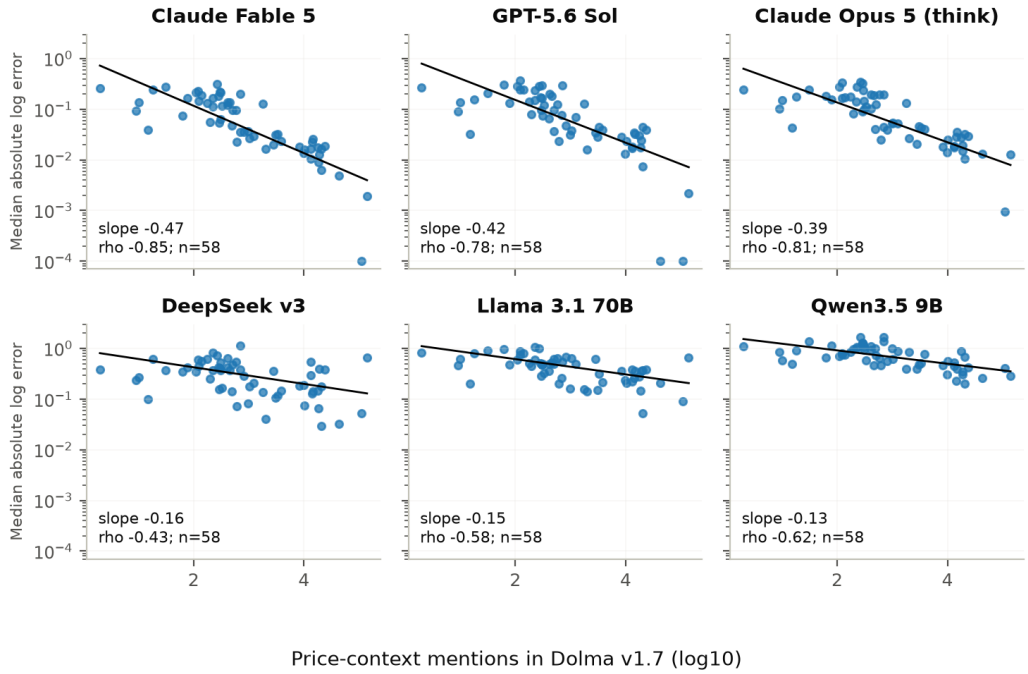

Figure 7: Price-context mentions in Dolma v1.7 versus recall error on 58 equities, using the corrected comparison set. Each dot is one equity. Lines use the median pairwise slope in log–log space; zero errors are clipped at 10^{-4} for plotting and fitting. The count is a proxy for exposure, not a measurement of any of these models' actual training examples.

to a date in our benchmark. The exposure analysis covers 58 of the 95 equities, for which positive price-context counts are available.

Figure 7 relates those counts to median absolute log error for each equity. For Fable, Sol, and Opus, Spearman correlations range from -0.85 to -0.78. The fitted log–log slopes range from -0.47 to -0.39: across equities, ten times as many mentions is associated with about 2.5–3 times lower error. The relationship is weaker for some other models; DeepSeek v3 has a rank correlation of -0.43 and a slope of -0.16.

Language coverage illustrates a further limitation of this proxy. Our Japanese phrase searches in Dolma v1.7 find little Toyota price text. A search of FineWeb-2's Japanese split, in contrast, finds 480,792 documents containing the Japanese term for closing price and 92,644 documents matching our Toyota-near-yen-price probe. The latter is a broad text match and can include yen amounts unrelated to a share price. The Dolma query returns 1,903 occurrences of the closing-price term. The searches count documents in FineWeb-2 and phrase occurrences in Dolma, so the totals are not directly comparable. They do show substantial Japanese financial coverage beyond our Dolma searches.

OLMo 3 was trained on Dolma 3, not the Dolma v1.7 index used above (Team Olmo et al., 2025). Apertus uses multilingual data derived from FineWeb-2 with additional processing (Project Apertus et al., 2025; Penedo et al., 2025). On 48 Toyota questions, the local OLMo 3.1 32B model answers two within 5%; Apertus 8B answers seven and Apertus 70B six. Yet Apertus 8B has a larger median log error than OLMo (0.89 versus 0.62), while Apertus 70B is better at 0.15. Explaining those differences will require tracing answers to documents in their actual training mixtures.

6 Related Work

Petroni et al. (2019) study factual knowledge in model parameters through factual queries, and Roberts et al. (2020) through question answering without external documents. Lopez-Lira et al.

(2025) ask the closely related question of whether models recall historical economic and financial values, and whether masking identifiers prevents that knowledge from being used. Such recall also matters for interpreting historical financial evaluations: Glasserman & Lin (2023) examine look-ahead bias in stock-return predictions based on GPT sentiment analysis. We use financial histories to compare approximate price knowledge with recovery of daily detail across many instruments and models.

Work on memorization shows that models can reproduce training examples (Carlini et al., 2021), with reproduction affected by model size, repetition, and prompting context (Carlini et al., 2023). For factual knowledge, Kandpal et al. (2023) relate recall to the frequency of relevant information in the training corpus. Allen-Zhu & Li (2025) study knowledge capacity under controlled training conditions. Our exposure analysis asks a related question using naturally occurring financial text. Work such as Geva et al. (2021) examines internal mechanisms of factual recall.

The form of the query can change what a model retrieves. Jiang et al. (2020) examine the effect of prompts on factual retrieval, while Elazar et al. (2021) study consistency across phrasings. Berglund et al. (2024) show that an association learned in one direction need not be available in the reverse direction. Our reverse-lookup task differs because it provides a fixed candidate set and numerical clues, but asks a similar question about access to stored information. Historical prices also require both numerical precision and temporal specificity, connecting this study to work on numeracy (Wallace et al., 2019) and time-dependent knowledge (Dhingra et al., 2022).

Dolma (Soldaini et al., 2024), FineWeb-2 (Penedo et al., 2025), and model releases with documented training data such as Apertus and OLMo (Project Apertus et al., 2025; Team Olmo et al., 2025) make it possible to examine the text available for learning these facts. We use these resources to investigate sources of financial knowledge.

7 Discussion

Models know a surprising amount about historical prices, down to daily movements in some series. The three regimes—Lookup Table, Anchor Points, and Coarse Level Prior—describe different levels of detail in that knowledge, and can coexist within a single model. Confidence adds another dimension: a model can answer accurately while reporting low confidence, yet its higher-confidence answers are often better still. Reverse lookup shows how much access depends on the question: dates make a price history much easier to identify, even after absolute price level is removed. More frequent price-context mentions are also associated with better recall. Together these results connect the detail of financial memory with the information available about an instrument and the cues used to retrieve it. Further experiments on repeated sampling and wording, post-training, confidence, synthetic-data models, and tasks that use recalled values are reported in Appendix F.

The experiments measure answers, so the internal representation remains an open question. Beating monthly log-linear interpolation establishes daily detail beyond that baseline, but a richer reconstruction could also produce it. Corpus counts measure public coverage rather than each model's training exposure, and the comparisons between released models also vary in training and serving. Our instruments are constrained by available historical feeds, and the dense panel has one window per instrument. The benchmark dates are not restricted to a verified cutoff for each model; historical accuracy should therefore be distinguished from forecasting ability. Rolling model aliases also limit exact reproducibility.

AI use statement

We used generative AI tools to help develop experiment code, inspect results, and draft and revise the manuscript. The evaluated models were also the subjects of the experiments described above. AI assistance included checking numerical claims against saved records and editing prose; it did not replace the recorded model responses or the reference financial data. The authors are responsible for the final analysis and manuscript.

Ethics statement

The study measures model responses to questions about historical financial data and examines public text corpora. Accurate historical recall should not be interpreted as evidence of an ability to forecast future prices. In particular, remembered outcomes can confound evaluations of financial decisions on historical data.

Reproducibility statement

The accompanying experiment code contains the question generators, provider configurations, cached reference data, and saved responses used in the analyses. Appendix A reports coverage, and Appendix B gives the dense-panel windows. The paper's evidence files record the comparison question IDs, model labels, source-file hashes, and derived tables. The main analyses and selected appendix results can be rebuilt without further model calls using `scripts/paper_audit.py` in the experiment repository; other supplementary results are preserved as analysis snapshots with their own scoring conventions.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/26d3c9a66836ded8f34a944f2bfe868e-Abstract-Conference.html.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.12288.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2202.07646.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273, 2022. doi: 10.1162/tacl_a_00459. URL https://aclanthology.org/2022.tacl-1.15/.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410. URL https://aclanthology.org/2021.tacl-1.60/.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Paul Glasserman and Caden Lin. Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. *arXiv preprint arXiv:2309.17322*, 2023. doi: 10.48550/arXiv.2309.17322. URL https://arxiv.org/abs/2309.17322.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. doi: 10.1162/tacl_a_00324. URL https://aclanthology.org/2020.tacl-1.28/.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15696–15707. PMLR, 2023. URL https://proceedings.mlr.press/v202/kandpal23a.html.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2401.17377.

Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. The memorization problem: Can we trust LLMs' economic forecasts? *arXiv preprint arXiv:2504.14765*, 2025. doi: 10.48550/arXiv.2504.14765. URL https://arxiv.org/abs/2504.14765.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro von Werra, and Thomas Wolf. FineWeb2: One pipeline to scale them all—adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*, 2025. doi: 10.48550/arXiv.2506.20920. URL https://arxiv.org/abs/2506.20920.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

Project Apertus et al. Apertus: Democratizing open and compliant LLMs for global language environments. *arXiv preprint arXiv:2509.14233*, 2025. doi: 10.48550/arXiv.2509.14233. URL https://arxiv.org/abs/2509.14233.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5418–5426. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.437. URL https://aclanthology.org/2020.emnlp-main.437/.

Luca Soldaini et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2024. URL https://arxiv.org/abs/2402.00159.

Team Olmo et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025. doi: 10.48550/arXiv.2512.13961. URL https://arxiv.org/abs/2512.13961.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do NLP models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5307–5315. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1534. URL https://aclanthology.org/D19-1534/.