PREDICTING WHERE RECALL FAILS: THE FACT MATTERS MORE THAN THE ENTITY IN BLACK-BOX LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models recall some facts reliably and fabricate others, and the standard predictor of where recall fails is entity popularity, which assigns one score to every fact about an entity. We ask whether factual reliability is a property of the entity or of the fact. Holding the entity fixed, we probe multiple numeric facts about the same entity and period in four knowledge domains (Major League Baseball, the National Hockey League, national economic indicators, and corporate finance), covering thirty statistics and seven models from six laboratories. The fact matters more than the entity. In the sports domains, the statistic being asked explains about half of the variance in answer error and the player about 5%. Entity popularity has no held-out predictive value, whereas prose redundancy, the Wikipedia mention rate of a statistic computed before any query, reaches a held-out R^2 of 0.41–0.44 and predicts error on statistics withheld from fitting (R^2 of 0.33–0.40), where features of the numbers themselves, such as their dispersion and magnitude, do not. Ordering statistics by prose redundancy places the least mentioned tier lowest in all four domains (fact-level exact permutation test, combined $p < 10^{-5}$); deviation correlation falls from $+0.894$ to $+0.163$ in baseball and from $+0.990$ to $+0.853$ in national statistics. A second pre-query measure, reconstructibility from co-occurring facts, indicates which low-redundancy facts are consistent with inference from headline statistics. For the least mentioned baseball statistics, models fabricate values while refusing fewer than 1% of questions, whereas for a season beyond their training data the non-answer rate rises to 24.8%. For two frontier models, the spread of repeated samples locates this boundary without ground truth, with an AUC of 0.952. The assessment needs only public text and API access to the model.

1 INTRODUCTION

Language models answer questions about well-known entities more accurately than questions about obscure ones, and entity popularity is the standard predictor of where recall fails (Kandpal et al., 2023; Mallen et al., 2023; Sun et al., 2024). An entity-level score, however, is the same for every fact about the entity, so it cannot tell a user which facts about a well-known entity are safe to ask. Comparisons across entities also confound overall prominence with the exposure of specific facts. A famous athlete's home runs and sacrifice flies belong to the same entity and the same database row, yet one appears in almost every biography and the other in about one in seven (Zhang et al., 2023).

We ask whether factual reliability is a property of the entity or of the fact. To separate the two, we hold the entity fixed. For each entity and period, we probe multiple numeric facts, so that entity popularity is constant while the mention density of the facts varies by more than two orders of magnitude. In baseball, the nine facts come from one database row, which also holds structural depth, the number of conditions attached to a statistic, constant. Two measures, both computed from public sources before any query, describe each fact. *Prose redundancy* is the mention density of the fact in the entity's Wikipedia article, and *reconstructibility* is how well the fact is predicted by co-occurring headline facts about the same entity. We test the approach in four knowledge domains documented

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.

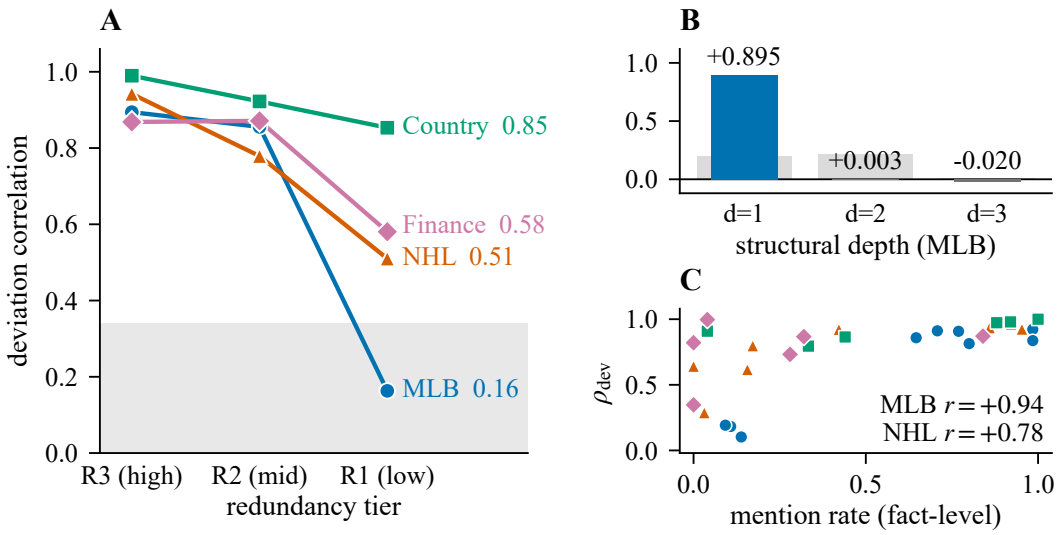

Figure 1: **Recall ordering among facts about the same entity in four domains. A.** Tier-level deviation correlation from R3 to R1, one line per domain; the gray band marks the largest 95th percentile of the block-permutation null over all domain–tier pairs. **B.** Depth reference from DEPTH-CONTROL (d=1 vs. B_0; d=2, d=3 vs. B_2). **C.** Fact-level Wikipedia mention rate vs. deviation correlation for all thirty statistics, colors and markers as in A; r is the within-domain Pearson correlation for the two nine-statistic domains. Table 2 reports the tier-level summary.

in different kinds of sources: two professional sports leagues, national economic indicators, and corporate finance.

The fact matters more than the entity. In the two sports domains, the statistic being asked explains 51% and 47% of the variance in answer error, and the player 5% (Table 1). Entity popularity adds no held-out predictive value, whereas a single pre-query number per statistic, its Wikipedia mention rate, reaches a held-out R^2 of 0.44 and 0.41, close to the 0.50 and 0.54 obtained when the identity of the statistic is known, and predicts the error level of statistics withheld from fitting ($R^2 = 0.40$ and 0.33). Features of the numbers themselves, such as their dispersion, magnitude, and share of zeros, do not predict held-out statistics.

The same ordering holds in all four domains. With tiers assigned before querying, the lowest-redundancy tier is the lowest-recall tier in each domain, and deviation correlation falls from $+0.894$ to $+0.163$ in Major League Baseball, from $+0.942$ to $+0.510$ in the National Hockey League, from $+0.868$ to $+0.581$ in corporate finance, and from $+0.990$ to $+0.853$ in national economic indicators (Figure 1, Table 2). An exact permutation test that treats each statistic as one observation gives a combined $p < 10^{-5}$ across the four domains. The baseball drop occurs at constant structural depth, so the depth effect measured in a separate control (Figure 1B) cannot produce it.

Reconstructibility separates two kinds of low-redundancy fact. For some, the recall signal largely disappears once co-occurring headline statistics are controlled for, a pattern consistent with inference from those statistics; for others, it remains intact.

At the low end of the ordering, in baseball, models answer less accurately than a league-median baseline while refusing fewer than 1% of questions; for a season beyond their training data, the non-answer rate rises to 24.8%. For the two frontier models, the spread of repeated samples, a standard black-box consistency signal, locates this boundary without ground truth with an AUC of 0.952. The $22{,}708$ response records from seven models and six laboratories are scored against official league records, World Bank data, and company financial statements. The assessment does not require knowledge of the training cutoff or access to logits, weights, or the training corpus.

2 RELATED WORK

Training cutoffs and temporal evaluation. Temporal holdouts are standard in language model evaluation, yet passive backtests split by a reported cutoff date inherently conflate temporal recency, data leakage, and model capability (Zhang & Stadie, 2026). Similar methodological challenges arise in studies of effective versus declared cutoffs (Cheng et al., 2024), temporal generalization gaps (Lazaridou et al., 2021), time-indexed probing (Dhingra et al., 2022), and continuously refreshed benchmarks (White et al., 2025). Our work predicts recall boundaries from fact-level corpus statistics and does not depend on a declared cutoff. Extraction attacks evaluate black-box models by recovering verbatim training sequences (Nasr et al., 2025; Shilov et al., 2026); our probe measures graded recall of individual numeric facts.

Contamination detection and knowledge boundaries. Existing contamination audits typically determine whether a specific datum was included in pretraining, using membership inference (Shokri et al., 2017; Shi et al., 2024), canonical-order likelihood tests (Oren et al., 2024), guided prompt completion (Golchin & Surdeanu, 2024), or dataset-level audits (Sainz et al., 2023). Our framework addresses boundary estimation instead of membership. It estimates where recall fails for an entire class of facts, producing a graded measure of recall quality and a classification of failure modes without requiring logit access or training-set inspection. Prior work formalizes knowledge boundaries along temporal, capability, and retrieval axes (Li et al., 2025), and recent evidence identifies recall as the primary bottleneck for parametric factuality (Calderon et al., 2026); fact-level prose redundancy provides a further predictive axis. Methods that estimate factual knowledge from hidden representations before generation (Gottesman & Geva, 2024) or detect guessing from internal attention mechanisms (Dentan et al., 2026) require white-box access; our approach operates entirely over black-box text outputs. Post-hoc hallucination detectors based on sampling consistency (Manakul et al., 2023) or semantic entropy (Kuhn et al., 2023) identify unreliable outputs after the model has responded; our framework operates before querying, predicting where recall will fail from public text alone, and Section 5.6 uses the same consistency signal to locate that boundary without labels.

Memorization, duplication, and frequency-recall relations. Empirical studies demonstrate that pretraining duplication strongly affects memorization and verbatim extraction (Carlini et al., 2021; 2023; Lee et al., 2022; Tirumala et al., 2022; Biderman et al., 2023; Allen-Zhu & Li, 2025). While intra-sequence redundancy describes token predictability within individual samples (Zhang et al., 2025a), we measure how often a fact is mentioned across documents. Studies using knowledge-probing benchmarks show that question-answering accuracy depends strongly on entity popularity and pretraining term frequencies (Petroni et al., 2019; Kandpal et al., 2023; Mallen et al., 2023; Sun et al., 2024; Razeghi et al., 2022; Smith et al., 2026). Recent work characterizes this relationship through training dynamics (Yu et al., 2023; Christoph et al., 2025), internal representations (Merullo et al., 2025; Wang et al., 2024), surface-form effects (Nishida et al., 2026), the interplay of entity frequency and context (Fouilhé et al., 2026), retrieval-mechanism diagnostics (Haviv et al., 2023), and decompositions of memorization into recitation, reconstruction, and recollection (Prashanth et al., 2025; Saynova et al., 2025). These studies vary frequency across entities or topics, so entity prominence and the exposure of any particular fact covary. Kandpal et al. (2023) count documents in which question and answer entities co-occur, and Smith et al. (2026) relate recall of scholarly references to topic-level representation across model sizes, but neither holds the entity fixed. Closest to our design, Niimi (2025) show in a single model that fields within one bibliographic record are memorized at different citation-count thresholds, and Zhang et al. (2025b) examine n-gram coverage as a pre-query hallucination signal. Frequency proxies built from Wikidata track pretraining counts only loosely (He et al., 2025), so we test the proxy directly against held-out recall error (Section 5.1). Our design evaluates multiple facts about the same entity and period across seven models, holding entity popularity constant and, in baseball, structural depth as well, measures coverage at the level of the individual fact, and shows that the resulting ordering holds in four knowledge domains.

3 METHOD: THE RECALL PROBE

The recall probe evaluates factual recall of numeric facts from black-box models in three stages. First, prose redundancy and reconstructibility are computed from public sources before any query. Second, standardized prompts are submitted to the target model under controlled decoding. Third,

responses are scored as deviations from a hierarchy of baselines, tested against unit-level permutation nulls, and classified by failure mode.

3.1 PROBE COMPONENTS

Deviation-space scoring. For a fact with ground truth y, model answer a, and reference baseline B, the unit of analysis is the pair $(a - B, y - B)$. The primary evaluation metric is the anomaly correlation of forecast verification (Murphy & Epstein, 1989), which we refer to as the deviation correlation:

$$\rho_{\text{dev}} \;=\; \text{corr}(a - B,\; y - B)\,. \tag{1}$$

When statistics with different scales are pooled into a tier T, each deviation is divided by the cross-entity standard deviation s_j of statistic j, and the correlation is taken over all (entity-period i, model m, statistic $j \in T$) records: $\rho_{\text{dev}}(T) = \text{corr}\left((a_{imj} - B_{ij})/s_j,\; (y_{ij} - B_{ij})/s_j\right)$. Absolute accuracy on numbers with strong prior distributions mainly measures adherence to the prior. A model that outputs typical values has $a \approx B$ and $\rho_{\text{dev}} \approx 0$, so a positive correlation requires predictions that track entity-specific deviations. The baselines form a hierarchy of increasing specificity: B_0 (population median across probed units, or a pre-test-period constant in DEPTH-CONTROL), B_1 (entity-period mean), B_2 (B_1 adjusted for subgroup effects), and B_{persist} (previous-period value, used in POST-CUTOFF). Scoring against the strongest applicable baseline separates recall from guessing typical values (Appendix A).

Unit-level block permutation null. The same (entity, period) units appear across all models and facts, so cell-level permutations underestimate variance. The null distribution instead permutes (entity, period) units as intact blocks. All statistics of a unit move together, as do its ground-truth values and baselines, while the outputs of every model stay fixed. Every permutation p-value and null threshold derives from this block-level distribution. We use $\text{NPERM} = 5000$ permutations for tier-level and partial-correlation tests and $\text{NPERM} = 2000$ for fact-level tests, with add-one p-values (Phipson & Smyth, 2010). Tests controlling for co-occurring facts residualize both predictions and true values before block permutation (Freedman & Lane, 1983; Winkler et al., 2014).

Failure-mode classification. The probe separates non-answers (explicit refusals and answers consisting only of a zero) from answers that are less accurate than the baseline. Answers that differ from the truth by a factor of 300 or more, mostly because the quantity is stated in the wrong unit, form a third category; they are counted and excluded from correlation and error statistics (Appendix B).

3.2 TWO PRE-QUERY MEASURES

Prose redundancy. Prose redundancy quantifies how frequently a fact appears within narrative text. It is the fact-level counterpart of the entity-level frequency proxies used in prior work (Kandpal et al., 2023; Mallen et al., 2023), and it counts mentions per statistic within the entity's own article instead of documents per entity. At the entity level, we record annual Wikipedia pageviews as a measure of entity prominence (Mestyán et al., 2013) (Appendix E). At the fact level, we retrieve the plain-text extract of every entity's Wikipedia article through the MediaWiki API (8 August 2026 for the sports domains, 18 September 2026 for the other two), which contains the article prose and section headings without infoboxes or tables, and tally mentions matching the noun phrases of each statistic. We assigned the statistics of each domain to three tiers by how prominently each statistic is usually reported, before querying and validated the assignment by text counts (Appendix C). Mean mentions per article decrease from R3 to R1 in every domain: $12.52 > 7.81 > 0.17$ in MLB and $19.56 > 0.50 > 0.01$ in NHL (65 and 64 players, nine statistics each), and $8.08 > 1.10 > 0.48$ in national statistics and $1.76 > 0.76 > 0.00$ in corporate finance (25 entities, six statistics each; Tables 11, 12, 13, and 14).

Reconstructibility. Reconstructibility measures how effectively neighboring headline facts predict a target statistic, capturing whether an answer can be derived without memorization. The notion corresponds to the reconstruction category in the memorization taxonomy of Prashanth et al. (2025); we operationalize it for numeric facts directly from the published data as the multiple correlation R between each statistic and the headline (R3) statistics of the same unit. Reconstructibility varies across tiers and within tiers (Section 5.4), and like prose redundancy it is available before any query.

4 EXPERIMENTAL DESIGN

The evaluation spans four knowledge domains and ten experiments totaling 22,708 response records across seven models (Table 3). The four domains are documented in different sources. MLB baseball (nine season-batting statistics from a single StatsAPI row) and NHL hockey (nine official regular-season statistics, with the six R3 and R2 statistics taken from the league's summary report and the three R1 statistics from its realtime report) are documented mainly in narrative prose, and their statistics span the widest range of mention density. National economic indicators (six country-level statistics: GDP, population, life expectancy, unemployment, inflation, Gini coefficient) are documented across encyclopedias, textbooks, government databases, and news reporting. Corporate finance (six metrics: revenue, net income, operating income, total assets, current ratio, debt-equity ratio) is documented mainly in regulatory filings, with sparse narrative coverage.

EQUAL-DEPTH isolates redundancy by probing nine statistics that come from the same MLB season-batting row, holding structural depth and entity popularity strictly constant. NHL-REPLICATION applies the same protocol to NHL regular-season statistics. COUNTRY and FINANCE probe six statistics each on 25 entities, using the same Wikipedia-based redundancy-tier methodology. DEPTH-CONTROL measures the effect of structural depth at three levels: a season batting average (d=1), the average against left-handed pitchers (d=2), and the average against left-handed pitchers in home games (d=3). POST-CUTOFF extends the framework to temporal boundaries by probing three high-redundancy statistics across 2023–2025 against the persistence baseline B_{persist}. Four within-domain controls (depth-scan, anchor-ablation, positive-control, repeat-sampling) are documented in the appendix.

The four cross-domain experiments have distinct roles. EQUAL-DEPTH separates redundancy from structural depth, NHL-REPLICATION tests the ordering in a second sport, and COUNTRY and FINANCE test it in domains where narrative prose is one source among several.

Preregistration. Preregistered criteria evaluate tier ordering in the equal-depth setting (Appendix F): R1 permutation $p \geq 0.05$ indicates a redundancy effect, $p < 0.05$ with $\rho_{\text{dev}} \geq 0.3$ indicates that a depth effect survives, and values in between yield an intermediate verdict.

Models and ground truth. Seven models from six independent laboratories: DeepSeek V4 Pro, Zhipu GLM-5.2, Alibaba Qwen3.6-27B, Nex-N2.5 Pro and Mini (Jin et al., 2024), GPT-5.5, and Claude Opus 5. All receive identical prompt templates under standardized decoding (temperature 1.0, single sample per cell, chain-of-thought disabled). Ground truth derives from official league records (MLB StatsAPI and Statcast, NHL API), World Bank Open Data (2022), and company income statements and balance sheets for fiscal year 2023 obtained through Yahoo Finance (Appendix B).

5 RESULTS

5.1 THE FACT, NOT THE ENTITY, EXPLAINS RECALL ERROR

We first ask how much of the variation in answer error is carried by the entity and how much by the fact. The unit is one (player-season, statistic) instance, and the outcome is its absolute normalized error $|a - y|/s_j$ averaged over the seven models. All models include the distance of the true value from the cross-entity median, which mechanically affects error. Table 1 reports the variance shares and held-out R^2 under two cross-validation schemes: folds grouped by player, and leave-one-statistic-out.

The statistic being asked explains about half of the variance in error, and the player about a twentieth. Entity popularity, the standard predictor, adds nothing to held-out prediction. The mention rate of the statistic, a single number computed before any query, recovers most of what knowing the identity of the statistic provides and predicts the error level of statistics that were not used for fitting. Other statistic-level features predict held-out statistics less well. The dispersion, magnitude, and share of zeros of the true values, whether the statistic is a rate, and the general-English frequency of its name reach at best -0.09 in MLB and 0.20 in NHL one at a time (the latter for name frequency) and -0.31 and 0.01 together; pooled over the eighteen sports statistics, the mention rate alone reaches 0.38 against -0.43 for the competing features combined. Within a statistic, players whose

Table 1: Entity versus fact in the sports domains. Instance-level outcome: absolute normalized error averaged over seven models (MLB: 657 instances, 64 players; NHL: 630 instances, 57 players; nine statistics each). Every model includes $|y - B_0|/s_j$. Popularity: $\log_{10}$ annual pageviews of the player-season. Mention rate: one pre-query number per statistic. Held-out R^2 uses five folds grouped by player, or leaves out one statistic at a time. †Competing statistic-level features: log coefficient of variation of the true values, log median true value, share of zero true values, general-English frequency of the statistic's name, and (MLB) rate versus count.

	MLB	NHL
Share of error variance explained by the statistic	0.51	0.47
Share of error variance explained by the player	0.05	0.05
Held-out R^2, folds grouped by player		
with entity popularity	−0.01	0.03
with statistic identity	0.50	0.54
with mention rate of the statistic	0.44	0.41
Held-out R^2, leave one statistic out		
with entity popularity	−0.14	−0.14
with mention rate of the statistic	0.40	0.33
with the best single competing feature†	−0.09	0.20
with all competing features†	−0.31	0.01

articles mention it more often are also answered more accurately after controlling for popularity (Appendix G).

5.2 Recall ordering across four knowledge domains

Table 2 reports tier-level deviation correlations and normalized errors for all four domains under one scoring protocol. Each statistic is normalized by its cross-entity standard deviation, B_0 is the cross-entity median, and significance comes from entity-level block permutation. Tiers were assigned by how prominently each statistic is usually reported, before querying, and validated by text counts in every domain (Section 3.2). An exact permutation test that treats each statistic as one observation and enumerates every assignment of statistics to tiers of the same sizes gives $p = 0.002$ (MLB), 0.004 (NHL), 0.044 (national statistics), and 0.144 (corporate finance) for the R3–R1 gap in fact-level deviation correlation, and $p < 10^{-5}$ when the four domains are combined.

Sports domains. In MLB and NHL, Wikipedia mention counts vary by orders of magnitude across statistics within the same schema row, and narrative prose is the primary channel through which these facts enter pretraining corpora. Both domains show the preregistered ordering. MLB shows the largest drop across tiers, with ρ_{dev} dropping from $+0.894$ (R3) through $+0.855$ (R2) to $+0.163$ (R1), an 82% decline. NHL replicates the ordering ($+0.942$ to $+0.510$) with a shallower slope; Section 5.4 attributes part of the NHL R1 signal to reconstruction from headline statistics.

National statistics. Country-level indicators are documented across encyclopedias, textbooks, government databases, and news reporting in addition to Wikipedia. Recall is high in every tier and still ordered by redundancy, with ρ_{dev} of $+0.990$ (R3), $+0.922$ (R2), and $+0.853$ (R1), and normalized error rises from 0.030 to 0.342 while remaining well below the B_0 baseline in every tier (0.43–0.56). In every model, R1 error exceeds R2 error and R3 error is lowest (Table 8). Near-exact recall of macroeconomic series within the training window has been reported independently (Lopez-Lira et al., 2025); holding the entity fixed further shows that, even in this densely documented domain, recall follows fact-level mention density.

Corporate finance. Financial metrics are reported in regulatory filings and receive sparse narrative coverage; the two R1 metrics have no Wikipedia mentions in any company article. R3 and R2 are tied ($\rho_{\text{dev}} = +0.868$ and $+0.871$), R1 is lowest ($+0.581$), and R1 normalized error (0.320) remains below the baseline (0.555). Section 5.4 shows that the recall signal for one R2 metric, operating income, is consistent with reconstruction from revenue and net income, while the signal for the R1 current ratio is not explained by the headline metrics.

Table 2: Deviation correlation by redundancy tier across four knowledge domains. ρ_{dev}: deviation correlation vs. B_0 (cross-entity median per fact); MAE(z): normalized absolute error (bold if worse than baseline); block permutation null: 5,000 unit-level shuffles. The lowest-redundancy tier is the lowest-recall tier in every domain. Depth reference from DEPTH-CONTROL: d=2 (+0.003), d=3 (−0.020); see Figure 1B.

Tier	n	ρ_{dev}	perm. p	null p_{95}	MAE(z)	B_0 MAE(z)
MLB (EQUAL-DEPTH)						
R3 high	1,566	+0.894	<0.0002	0.144	0.225	0.758
R2 mid	1,570	+0.855	<0.0002	0.141	0.305	0.761
R1 low	1,564	+0.163	<0.0002	0.063	**1.084**	0.780
NHL (NHL-REPLICATION)						
R3 high	1,642	+0.942	<0.0002	0.174	0.177	0.939
R2 mid	1,672	+0.779	<0.0002	0.141	0.404	0.818
R1 low	1,654	+0.510	<0.0002	0.103	**0.797**	0.780
National statistics (COUNTRY)						
R3 high	346	+0.990	<0.0002	0.314	0.030	0.433
R2 mid	350	+0.922	<0.0002	0.341	0.152	0.563
R1 low	343	+0.853	<0.0002	0.337	0.342	0.552
Corporate finance (FINANCE)						
R3 high	349	+0.868	<0.0002	0.249	0.159	0.690
R2 mid	346	+0.871	<0.0002	0.243	0.146	0.493
R1 low	343	+0.581	<0.0002	0.199	0.320	0.555

Across the four domains, the gap between the top and bottom tiers ranges from 14% (national statistics) to 82% (baseball). The gap varies with the domain; the position of the lowest tier does not. Figure 1C places all thirty statistics on common axes of Wikipedia mention rate and fact-level deviation correlation. In the sports domains, where nine statistics per domain support a continuous test, the two are correlated at $r = +0.943$ and $+0.783$ (Section 5.3). Eight of the 2,085 country and finance answers differ from the truth by a factor of 300 or more, seven of them because of a wrong unit, and are excluded as a failure mode (Appendix B).

5.3 RECALL DEGRADATION IN THE SPORTS DOMAINS

The sports domains span the widest range of mention density and show the largest drops across tiers. Across the nine statistics of each domain, mention rate correlates with deviation correlation at Pearson $r = +0.943$ in MLB ($p = 0.006$) and $+0.783$ in NHL ($p = 0.002$), and at $+0.896$ ($p = 0.004$) and $+0.770$ ($p = 0.006$) on log counts. Alternative keyword sets that keep the prose phrases give correlations of $[+0.681, +0.943]$ in MLB and $[+0.695, +0.783]$ in NHL (Table 5).

Preregistered tests. In MLB, each R1 statistic retains a significant positive correlation (hit-by-pitch $+0.183$, $p = 0.001$; grounded-into-double-play $+0.192$, $p = 0.001$; sacrifice flies $+0.102$, $p = 0.049$). MLB R1 ($+0.163$) falls in the preregistered intermediate band, which places an upper bound of about 0.16 on any independent depth effect; NHL R1 ($+0.510$) falls in the *depth effect survives* band, and residualization on headline statistics brings it to $+0.209$ (Section 5.4; Appendix F). At the same time, pooled MAE at R1 exceeds the B_0 baseline (1.084 vs. 0.780; Table 2). Models track the direction of each deviation but overshoot its size, so their answers are less accurate than the league median. Per-model results are in Tables 6 and 7.

Prominence does not produce the drop. Entity prominence is associated with slightly lower error at high redundancy but does not explain the tier ordering. The drop appears within both prominence strata, and the R1 reversal between strata (bench $\rho_{\text{dev}} = +0.310$ vs. star $+0.058$) leaves R1 the lowest tier in each (Table 4).

Failure mode: fabrication at low redundancy. At the low-redundancy boundary in baseball, models answer without abstaining and fabricate values. In MLB R1, normalized absolute error reaches 1.084, exceeding the league-median baseline of 0.780; all seven models individually exceed the baseline (Table 6). At R1, always predicting the league median would outperform every eval-

uated model, a form of confident confabulation (Farquhar et al., 2024). Error beyond the baseline is largest where mention density and reconstructibility are both lowest (MLB R1); NHL R1 error is at the baseline (0.797 vs. 0.780), and in national statistics and corporate finance R1 error stays below it. The non-answer rate never exceeds 2.3% in any redundancy tier, against 24.8% at temporal boundaries (Section 5.5). The seven models indicate when a question falls after their training period, but not when a fact is sparsely covered within it.

5.4 Reconstructibility and inference from headline statistics

Statistics in the same row are correlated, so a positive deviation correlation can come from direct memorization, from reconstruction out of neighboring facts, or from both. Reconstructibility, the multiple correlation R between a statistic and the headline (R3) statistics of the same unit, indicates in advance how much of a correct answer could be derived from the headline statistics. Residualizing predictions and true values on the co-occurring R3 statistics with the Freedman–Lane procedure ($\mathrm{NPERM} = 5000$) then tests whether a recall signal remains once linear reconstruction is removed.

Facts consistent with reconstruction. Where reconstructibility is high, the recall signal largely disappears under residualization. NHL takeaways and giveaways have reconstructibility 0.84 and 0.89 from goals, assists, and points; their raw correlations of $+0.645$ and $+0.642$ fall to $+0.157$ and $+0.036$ ($p = 0.21$ for giveaways). Corporate operating income has reconstructibility 0.86 from revenue and net income; its raw correlation of $+0.732$ falls to $+0.063$ ($p = 0.24$). Average R1 reconstructibility is 0.682 in NHL against 0.274 in MLB, and residualization brings the two sports into alignment, with NHL R1 dropping from 0.510 to 0.209 and MLB from 0.163 to 0.124 (Appendix D).

Facts with a direct recall signal. Where reconstructibility is low, the signal remains after residualization. National statistics at R1 have reconstructibility 0.19 from GDP and population, and the pooled R1 correlation moves from $+0.853$ to $+0.868$ under residualization; inflation and the Gini coefficient retain $+0.906$ and $+0.831$. The corporate current ratio, with no Wikipedia mentions and reconstructibility 0.38, retains $+0.795$, and total assets retain $+0.997$. For these values, the models' answers carry information that the headline statistics do not. A historical reconstruction baseline built from the previous season gives the same picture (Appendix L). For any fact in the lower tiers, reconstructibility computed from the published data therefore indicates whether a correct answer could have been derived from neighboring facts.

5.5 Post-Cutoff: a second kind of boundary

The probe receives no information about training cutoffs, yet a temporal boundary appears in the responses. Post-Cutoff queries three high-redundancy statistics for the 2023, 2024, and 2025 seasons and scores answers against the persistence baseline B_{persist}. Non-answer rates, which combine explicit refusals and bare-zero answers, are 0.0% across all 840 records for 2023 and 2024 and 24.8% for 2025 (Figure 2). On items answered by at least five models, pooled error is lower than the persistence error by a factor of 3.5 in 2023 (0.251 vs. 0.888) and 2024 (0.272 vs. 0.950), and approaches it in 2025 (0.837 vs. 0.736, bootstrap interval for the ratio $[0.925, 1.438]$). Three models abstain often on 2025 queries (GLM 88.3%, DeepSeek 41.7%, Qwen 38.3%); GPT-5.5 and Claude Opus 5 stay well below the persistence error (0.167 and 0.236), consistent with training data that extends into 2025; Nex-Pro and Nex-Mini exceed it (1.059 and 1.092; Table 10). At this boundary models abstain, continue to recall, or degrade (Yin et al., 2023; Wen et al., 2025), whereas at the redundancy boundary abstention is near zero. We define this boundary by behavior alone, because comparisons around a training cutoff cannot separate recency effects from leakage (Zhang & Stadie, 2026).

5.6 Detecting the knowledge boundary without ground truth

Pre-query tiers define which facts should lie inside and outside the knowledge boundary, so a model's answers can be scored on how well they separate the two. A receiver operating characteristic (ROC) analysis (Hanley & McNeil, 1982; Fawcett, 2006) of the absolute error $|a - y|$ in normalized units separates high- from low-redundancy facts in all four domains, with pooled AUCs of 0.808 (MLB), 0.834 (NHL), 0.795 (corporate finance), and 0.761 (national statistics). Per-entity AUCs average

0.806, 0.824, 0.798, and 0.756, and exceed 0.75 for 51 of 65 MLB players, 56 of 64 NHL players, 20 of 25 companies, and 13 of 25 countries. Separation is lowest in national statistics, where the tier gap is smallest (Section 5.2). With depth-based labels in MLB, the per-player mean AUC is 0.888 (Appendix Figure 3).

An auditor usually lacks ground truth. Signals computed from a single response, such as distance from the baseline or refusal, do not separate the depth classes of DEPTH-CONTROL (Table 15). Repeated-sample dispersion does (Manakul et al., 2023; Kuhn et al., 2023; Wang et al., 2023). Across the 159 R3 and R1 MLB cells answered by the two frontier models, the standard deviation across four repeated samples separates high- from low-redundancy facts with an AUC of $\mathbf{0.952}$, reaching 99.5% of the ground-truth AUC (0.957; Appendix J). Sampling variability at redundancy boundaries closely tracks prediction error, so the knowledge boundary can be located from response consistency without verified answers.

6 DISCUSSION

Factual reliability belongs to the fact more than to the entity. The statistic being asked explains about ten times as much of the error as the player, and among facts about the same entity recall follows the prose redundancy of the individual fact in all four domains. The ordering appears where narrative prose is the main source (sports), where it is one source among many (national statistics), and where regulatory filings are the main source (corporate finance). The agreement across seven models from six laboratories fits a token-frequency account of memorization (Carlini et al., 2023), under which facts that appear across many independently authored passages accumulate more gradient signal than facts confined to structured sources.

Holding the entity fixed isolates this fact-level effect from the entity-level popularity effects that prior frequency–recall studies measure (Kandpal et al., 2023; Mallen et al., 2023). In MLB, nine statistics sharing identical entity popularity and structural depth span a 72-fold range in mean mentions per article between the highest and lowest tiers, and the corresponding recall gap reaches 82%; the same ordering among facts about one entity appears in hockey, national statistics, and corporate finance. Entity-level evaluations assign one reliability level to all facts about an entity (Kandpal et al., 2023; Sun et al., 2024). These results show that the same entity yields accurate recall for one statistic and fabrication for another, so reliability assessment must operate at the level of the individual fact.

The two pre-query measurements play complementary roles. Prose redundancy orders the facts within a domain. Reconstructibility indicates which facts in the lower tiers could be recovered by inference from headline facts and which carry a recall signal beyond them, a distinction that a single accuracy number cannot make.

These findings make it possible to estimate factual reliability for each query before the model is queried, for models accessed through a text API without retrieval or tools. Fact-level redundancy, computed from public text, ranks the facts of a domain by expected recall. Reconstructibility, computed from the published data, indicates whether a low-ranked fact could still be recovered by inference. When repeated samples are available, their dispersion locates the boundary without ground-truth labels.

7 CONCLUSION

Whether a black-box language model recalls a fact depends more on the fact than on the entity. Holding the entity fixed and probing multiple facts about it, we find that entity popularity has no held-out predictive value, while prose redundancy, computed from public text before any query, predicts error on held-out statistics and orders recall among facts in all four knowledge domains tested, with a gap that ranges from 14% to 82%, and that reconstructibility identifies facts whose recall is consistent with inference from their neighbors. Where the drop from R3 to R1 is largest, models fabricate, and repeated-sample dispersion detects this boundary with AUC 0.952 without ground truth. Factual reliability is a fact-level property, and it is estimable from public corpus statistics before a model is queried.

AI USE STATEMENT

During manuscript preparation, a large language model was used to refine sentence fluency. All research contributions, including question formulation, experimental design, data analysis and interpretation, were carried out by the authors, who bear sole responsibility for the content.

Separately from manuscript preparation, language models appear in this work only as the seven probed models (Section 4).

REPRODUCIBILITY STATEMENT

All quantitative claims trace to released result files produced by the ten experiments across four domains (Section 4) plus two further analyses (the reconstruction baseline and the attenuation simulation, Sections 5.4 and Appendix L), and every figure and table in this paper is generated by deterministic scripts from those files with zero additional model calls. Appendix B documents prompts and decoding settings; Appendix F reproduces the preregistered judgment criteria verbatim; Appendix D provides complete per-model tables. Permutation tests use fixed seeds: $5{,}000$ permutations for tier-level, partial-correlation and continuous-coupling tests, $2{,}000$ for fact-level and two-factor tests, and $1{,}000$ for the block-correlation nulls and the attenuation simulation. Code and result files will be released.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Nitay Calderon, Eyal Ben-David, Zorik Gekhman, Eran Ofek, and Gal Yona. Empty shelves or lost keys? recall is the bottleneck for parametric factuality. *arXiv preprint arXiv:2602.14080*, 2026.

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, 2021.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. Dated data: Tracing knowledge cutoffs in large language models. In *First Conference on Language Modeling (COLM)*, 2024.

Daniel Christoph, Max Ploner, Patrick Haller, and Alan Akbik. From data to knowledge: Evaluating how efficiently language models learn facts. In *Proceedings of the ACL 2025 Workshop on Large Language Model Memorization (L2M2)*, 2025.

Jérémie Dentan, Davide Buscaldi, and Sonia Vanier. Guess or recall? training CNNs to classify and localize memorization in LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

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.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010.

Guilhem Fouilhé, Nicholas Asher, and Philippe Muller. Memory vs. context? influential factors of factual recall in language models. *arXiv preprint arXiv:2609.24238*, 2026.

David Freedman and David Lane. A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298, 1983. doi: 10.1080/07350015.1983.10509354.

Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Daniela Gottesman and Mor Geva. Estimating knowledge in large language models without generating a single token. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 248–264, 2023.

Yuan He, Bailan He, Zifeng Ding, Alisia Lupidi, Yuqicheng Zhu, Shuo Chen, Caiqi Zhang, Jiaoyan Chen, Yunpu Ma, Volker Tresp, and Ian Horrocks. Supposedly equivalent facts that aren't? entity frequency in pre-training induces asymmetry in LLMs. In *Proceedings of the Second Conference on Language Modeling (COLM)*, 2025.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. A comprehensive evaluation of quantization strategies for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12186–12215, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

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 (ICML)*, pp. 15696–15707, 2023.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pp. 29348–29363, 2021.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, 2022.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. Knowledge boundary of large language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. The memorization problem: Can we trust LLMs' economic forecasts? *arXiv preprint arXiv:2504.14765*, 2025.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9004–9017, 2023.

Jack Merullo, Noah A. Smith, Sarah Wiegreffe, and Yanai Elazar. On linear representations and pretraining data frequency in language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Márton Mestyán, Taha Yasseri, and János Kertész. Early prediction of movie box office success based on Wikipedia activity big data. *PLOS ONE*, 8(8):e71226, 2013. doi: 10.1371/journal.pone.0071226.

Allan H. Murphy and Edward S. Epstein. Skill scores and correlation coefficients in model verification. *Monthly Weather Review*, 117(3):572–581, 1989. doi: 10.1175/1520-0493(1989)117⟨0572:SSACCI⟩2.0.CO;2.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Junichiro Niimi. Hierarchical memorization in large language models: Evidence from citation generation. *arXiv preprint arXiv:2511.08877*, 2025.

Yuto Nishida, Naoki Shikoda, Yosuke Kishinami, Ryo Fujii, Makoto Morishita, Hidetaka Kamigaito, and Taro Watanabe. Revisiting non-verbatim memorization in large language models: The role of entity surface forms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. Proving test set contamination in black box language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

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, 2019.

Belinda Phipson and Gordon K. Smyth. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, 2010. doi: 10.2202/1544-6115.1585.

USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, Jyothir S V, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A. Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, Katherine Lee, and Naomi Saphra. Recite, reconstruct, recollect: Memorization in LMs as a multifaceted phenomenon. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Yasaman Razeghi, Robert L. Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 840–854, 2022.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10776–10787, 2023.

Denitsa Saynova, Lovisa Hagström, Moa Johansson, Richard Johansson, and Marco Kuhlmann. Fact recall, heuristics or pure guesswork? precise interpretations of language models for fact completion. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

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

Igor Shilov, Matthieu Meeus, and Yves-Alexandre de Montjoye. The mosaic memory of large language models. *Nature Communications*, 17:2142, 2026.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (S&P)*, pp. 3–18, 2017.

Matthew L. Smith, Jonathan P. Shock, Samuel T. Segun, Iyiola E. Olatunji, and Tegawendé F. Bissyandé. Predictable confabulations: Factual recall by LLMs scales with model size and topic frequency. *arXiv preprint arXiv:2605.18732*, 2026.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974. doi: 10.1111/j.2517-6161.1974.tb00994.x.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 311–325, 2024.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pp. 38274–38290, 2022.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Yifei Wang, Yuheng Chen, Wanting Wen, Yu Sheng, Linjing Li, and Daniel Dajun Zeng. Unveiling factual recall behaviors of large language models through knowledge neurons. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7388–7402, 2024.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Anderson M. Winkler, Gerard R. Ridgway, Matthew A. Webster, Stephen M. Smith, and Thomas E. Nichols. Permutation inference for the general linear model. *NeuroImage*, 92:381–397, 2014. doi: 10.1016/j.neuroimage.2014.01.060.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8653–8665, 2023.

Qinan Yu, Jack Merullo, and Ellie Pavlick. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9924–9959, 2023.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Jie Zhang, Qinghua Zhao, Chi ho Lin, Zhongfeng Kang, and Lei Li. Beyond frequency: The role of redundancy in large language model memorization. *arXiv preprint arXiv:2506.12321*, 2025a.

Shuo Zhang, Fabrizio Gotti, Fengran Mo, and Jian-Yun Nie. Measuring the impact of lexical training data coverage on hallucination detection in large language models. *arXiv preprint arXiv:2511.17946*, 2025b.

Zeyu Zhang and Bradly C. Stadie. Temporal leakage in LLM backtesting: Measurement, validation, and adjusted scores. *arXiv preprint arXiv:2608.02985*, 2026.

A WHY THE PROBE SCORES DEVIATIONS RATHER THAN PROBABILITIES

This appendix gives the empirical basis for two design choices in Section 3.1.

First-token logprob is not a calibrated probability. An alternative design asks a binary question for each plate appearance ("HIT or OUT?") and uses the first-token log-probability as a rate. When these cell-level predictions are marginalized back to one-dimensional rates, their correlation with published rates is $+0.054$ under a statistical framing and -0.086 under a persona framing. For home-run rates, two models imply a marginal rate of ~ 0.7 against a true rate of 0.031. A first-token logprob expresses a preference between two candidate tokens and does not measure a frequency, and recalling a memorized statistic resembles reporting a number, not sampling a random variable. The probe therefore scores reported numbers, and no probability-level quantity appears anywhere in this paper.

League-level quantities cannot separate recall from a constant. League splits of batting average sit within a few thousandths of the league mean, so an answerer that always returns the mean is nearly optimal. Model answers for league-level batting averages in their published form show deviations small enough to resemble digit-level recall, yet the models score worse than a constant baseline that does not draw on the test period, the league mean of 0.2446. A depth sweep on these quantities also degrades the naive baseline at the same rate as the model. Probed quantities therefore need conditional variation that exceeds both the naive baseline error and the answer granularity. For this reason the probe targets player-level facts, whose realized batting averages span 0.197–0.354 across the probed units.

B INSTRUMENT DETAILS

Prompts. Every prompt is one question followed by one format line. In the player-level experiments, the question stem is *"In the [season] Major League Baseball regular season, what was [player]'s [statistic]?"*, and in NHL-REPLICATION *"In the [2022–23|2023–24] National Hockey League (NHL) regular season, what was [player]'s [statistic]?"*. The format line differs by experiment, and it determines which numbers a prompt contains. In DEPTH-CONTROL it is *"Reply with a single number between 0 and 1, such as 0.250. No words."*, so the prompts of that experiment contain three numbers besides the season, the range bounds and the format example; Appendix G ablates the anchoring effect of the example. In EQUAL-DEPTH, NHL-REPLICATION and POST-CUTOFF it is instead *"Reply with only the number, in three decimal places. No words."* for the decimal-valued facts (batting average and OPS, two of the nine in EQUAL-DEPTH and two of the three in POST-CUTOFF) and *"Reply with only the integer. No words."* for every count-valued fact, which in NHL-REPLICATION is all nine. Neither line carries an example or a bound, so over those $11{,}025$ calls the season is the only number the prompt contains. The answer-range bounds used by the parser never appear in prompt text. **Answer parsing.** Temporal refusals ("the 2025 season has not happened") are treated as behavioral responses rather than instrument failures, since they carry information the probe is designed to capture. Answers outside the physical range of a statistic are retained. DeepSeek returned integers such as 73, 100, and 130 for facts whose probed values never exceed 23; such answers are the fabrications that characterize the low-redundancy tier, and filtering them would remove that behavior from the measurement. Discarding the ten answers in this class

Table 3: Ten experiments across four domains. Ground truth sources: MLB StatsAPI (with Statcast pitch-level data for DEPTH-CONTROL d=3 and DEPTH-SCAN), NHL API, World Bank Open Data, and company financial statements obtained through Yahoo Finance.

Domain	Experiment	Design	n	Question answered
Cross-domain design				
MLB	EQUAL-DEPTH	9 facts from one StatsAPI row, 3 tiers; 75 units × 7 models	4,725	Does recall follow redundancy at equal depth?
NHL	NHL-REPLICATION	9 season-level facts, 3 tiers; 80 units × 7 models	5,040	Does the ordering replicate in a second sport?
Country	COUNTRY	6 facts, 3 tiers; 149 entity–fact cells (25 entities) × 7 models	1,043	Does the ordering hold outside sports?
Finance	FINANCE	6 facts, 3 tiers; 149 entity–fact cells (25 entities) × 7 models	1,043	Does the ordering hold where prose coverage is sparse?
Within-domain controls				
MLB	DEPTH-CONTROL	65 players × $d1/d2/d3$; 75 units × 7 models	5,173	Isolate structural depth from redundancy
MLB	POST-CUTOFF	20 players × 3 facts × 3 seasons × 7 models	1,260	Temporal boundary detection
MLB	DEPTH-SCAN	133 league-level batting-average cells, $d1$–$d5$ × 7 models	931	Continuous depth profile
MLB	ANCHOR-ABLATION	80 d=1 cells (40 star, 40 bench), no-example prompt × 7 models	560	Format-example anchoring
MLB	POSITIVE-CONTROL	59 published league-level rates × 7 models	413	Verify probe detects genuine recall
MLB	REPEAT-SAMPLING	40 cells × 3 tiers × 7 models, 3 additional draws per cell	2,520	Label-free detector
Total response records			22,708	

Table 4: Two-factor breakdown: prominence × redundancy tier (EQUAL-DEPTH, MLB). r is the correlation between $\log_{10}$ pageviews and $|z$-error$|$; one-sided permutation p (for lower error with higher prominence).

			Star		Bench	
Tier	r(prom, $\|$err$\|$)	perm. p	ρ_{dev}	MAE(z)	ρ_{dev}	MAE(z)
R3	-0.093	0.003	$+0.902$	0.194	$+0.846$	0.253
R2	-0.021	0.118	$+0.822$	0.304	$+0.826$	0.305
R1	$+0.045$	0.94 (rev.)	$+0.058$	1.268	$+0.310$	0.922

would raise the R1 deviation correlation from $+0.163$ to $+0.238$, and the preregistered verdict is the same with or without them. Because a refusal can carry a number in the same sentence ("the 2025 season has not happened, so 0.000"), we scan every scored response for temporal and epistemic refusal language and reclassify any match.

Unit-scale answers. The country and finance prompts request a unit (trillions of dollars for GDP, billions for financial statements). Eight of the $2{,}085$ scored answers in these two domains differ from the truth by a factor of 300 or more. Seven of them differ by a factor between 700 and $2{,}400$, consistent with a value reported in billions where trillions were requested or in millions where billions were requested. The eighth is an operating-income answer for a company whose truth is near

zero, where the ratio rule applies although the normalized error is 0.30. We count these answers as a failure mode and exclude them from the correlation and error statistics of Table 2, because a correct value at the wrong scale and a fabricated value are different behaviors. The position of the lowest tier does not depend on this exclusion. On all scored records, Spearman correlations are $+0.896$, $+0.913$, and $+0.786$ for country R3, R2, and R1, and $+0.859$, $+0.851$, and $+0.747$ for finance. No answer in the sports domains falls in this class, since those prompts request counts or rates with no unit choice.

What counts as a non-answer. A non-answer is either an explicit refusal such as "the 2025 season has not yet happened" or a bare zero, which expresses the same abstention without the sentence. Both are behavioral signals, and a non-answer is scored as neither a correct nor an incorrect answer. Parse failures form a separate category and are never counted as non-answers.

Bare-zero answers. A response whose entire body is a single zero (`0`, `.000`, or `0.000`) contains no prose, so the refusal scanner cannot detect it, and naive scoring counts it as a numeric estimate. Season dependence shows that such responses are non-answers. With the same players, the same facts and byte-identical prompts, POST-CUTOFF produces none of them in 420 records on 2023 and none in 420 on 2024, but 54 of 370 on 2025. We therefore reclassify a response as a non-answer when, and only when, (i) its entire body is one zero token and (ii) zero lies outside that fact's ground-truth support, so that zero is not a possible value of the quantity. Condition (ii) keeps the rule from selecting on the outcome in the redundancy-tier analysis, where zero is a common truth. Stolen bases, sacrifice flies and power-play goals have a zero truth in 34 NHL and 6 MLB units, and for those facts a zero answer is scored as given. The condition does not bind in POST-CUTOFF, which probes only batting average, home runs and OPS; their minimum true values over the experiment's $1{,}260$ records are 0.197, 4 and 0.704, and no unit has a true value of zero. Every bare zero in that experiment is therefore removed, and each is a maximal-error record. In that experiment the rule affects 14.6% of scored 2025 answers, and the season dependence described above is its basis.

The rule is applied uniformly in all four domains. It changes none of the six tier-level correlations by more than 0.022: MLB $+0.872 \to +0.894$ (R3), $+0.854 \to +0.855$ (R2), $+0.163 \to +0.163$ (R1); NHL $+0.941 \to +0.942$, $+0.780 \to +0.779$, $+0.514 \to +0.510$. Neither preregistered verdict changes. Its effect is concentrated in POST-CUTOFF, where it lowers the pooled 2025 error over all scored items from 1.769 to 0.808, against a persistence baseline that moves from 0.716 to 0.725 (Section 5.5 reports the common subset of items answered by at least five models), and changes whether one model falls below the persistence baseline (Section 5.5).

Models and decoding. We evaluate seven models covering six families: DeepSeek V4 Pro, Zhipu GLM-5.2, Alibaba Qwen3.6-27B, two Nex-N2.5 variants (Pro and Mini), GPT-5.5, and Claude Opus 5. Every call uses `temperature`$= 1.0$, `max_tokens` of at least 2048, one sample per cell, and explicit chain-of-thought disabled, so answers are direct reports rather than derivations. Sampling at temperature one adds noise to every estimate and attenuates the reported correlations. Each sports-domain tier pools roughly $1{,}600$ records (about 350 in the other two domains), and a repeat-sampling check on a preregistered subset (Appendix J) shows that this noise is small relative to the tier differences. All models were queried through standard chat-completion API endpoints with no system prompt, no browsing or retrieval tools enabled, and no tool-use configuration; the prompt is the sole input to the model.

Roster. The EQUAL-DEPTH roster comprises 75 (player, season) units (38 in 2023 and 37 in 2024, with 10 players appearing in both), matching DEPTH-CONTROL, for 4,725 calls. One player (José Ramírez) lacks Wikipedia pageview data and is excluded from continuous prominence analyses; tier-level analyses retain him.

Instrument independence. Two conditions separate the measuring instruments from the models being measured. First, every model receives the identical prompt and decoding settings for a given cell, so no model is asked an easier question than another. Second, refusals are identified by a fixed regular-expression scanner for temporal and epistemic refusal language, so no model judges another model's answers. Every figure and table is regenerated by deterministic scripts from the released result files with zero additional model calls.

Baseline hierarchy. The four baselines are computed from ground-truth values only, without access to model outputs. B_0 is the median of ground-truth values across the probed entities for a given fact and period; in DEPTH-CONTROL, B_0 is a constant from the pre-test period. B_1 is the mean

ground truth for the same entity across the probed periods. $B_2 = B_1 + \hat{\delta}_g$, where $\hat{\delta}_g$ is the mean residual $(y - B_1)$ for the entity's subgroup (e.g., position in sports domains). B_{persist} is the entity's ground truth from the immediately preceding period, used in POST-CUTOFF. Each experiment is scored against the strongest baseline whose conditioning variables are available for all units in that experiment.

C REDUNDANCY: TIER ASSIGNMENT, TERM LISTS, MATCHING RULES, SENSITIVITY

Tier assignment precedes measurement. The three redundancy tiers are declared in each experiment's elicitation script, in the fact specification that also fixes the API field and the question wording, and are fixed before any call is issued or any article is retrieved. The assignment rule is editorial prominence in the sport's own reporting: R3 is the headline line of a box score or a player card (batting average, home runs, OPS; goals, assists, points), R2 is the second block of the same card (RBI, stolen bases, walks; power-play goals, shots, penalty minutes), and R1 is a quantity that appears only on a full statistics table and essentially never in a headline (hit-by-pitch, sacrifice flies, grounded-into-double-play; blocked shots, takeaways, giveaways). The Wikipedia mention counts were collected afterwards and serve as an independent check on this ordering. Every continuous analysis in this paper is computed from measured mention rates, not from tier index.

Term lists. MLB. AVG: "batting average"; AVG. HR: "home run(s)", "homer(s)"; HR, HRS. OPS: "on-base plus slugging" (hyphen optional); OPS. RBI: "run(s) batted in"; RBI, RBIS. SB: "stolen base(s)"; SB, SBS. BB: "walk(s)", "base(s) on balls"; BB, BBS. HBP: "hit by pitch(es)" (space or hyphen); HBP. SF: "sacrifice fly/flies", "sac fly/flies"; SF. GIDP: "double play(s)", "grounded into"; GIDP, GDP.
NHL. GOALS: "goal(s)". ASSISTS: "assist(s)". POINTS: "point(s)". PIM: "penalty minute(s)"; PIM, PIMS. PPG: "power-play goal(s)" (space or hyphen); PPG. SHOTS: "shot(s)"; SOG. BLK: "blocked shot(s)", "shot-block(s/ing/er)". TKA: "takeaway(s)". GVA: "giveaway(s)".

Matching rules. Phrases are matched case-insensitively, with the optional plural written into the pattern; word-boundary anchors are applied to any term that is a substring of ordinary English words (*homer*, *walk*, *sac fly*, *goal*, *point*, *shot*, *assist*). Abbreviations are matched as whole words and *case-sensitively*, so "SF" counts and "sf" does not. A fact's count in an article is the sum of matches over all of its patterns. The *mention rate* is the fraction of the domain's articles with at least one match, and the *total count* is the sum over articles, entering continuous analyses as $\log(1 + x)$. No lemmatizer, stemmer or parser is used, and no manual adjudication of individual matches is performed.

Documented sources of counting noise. The pattern "double play" also matches defensive-side narration (*turned a double play*), so the GIDP rate is an overestimate, and the bare word "walk(s)" also matches non-statistical uses. Bare capitalized abbreviations are rare in biographical prose and come mainly from table-header residue in the article text; case-sensitive matching prevents them from matching ordinary words, but they add a component that does not reflect mentions in prose. In the sensitivity analysis below, the row *Documented noise removed* restricts GIDP to "grounded into" and BB to "bases on balls", and the *Phrases only* rows remove the abbreviations.

Sensitivity to the term list. We recount the same cached article text under alternative term lists and recompute the fact-level correlation against the fixed per-fact deviation correlations. No model is queried, and no result in the main text is affected. Table 5 gives the result. The correlation persists under every list that keeps the prose phrases, and the per-article tier means remain monotone. This monotonicity is specific to the per-article transform. On the article-share transform, the MLB R3 and R2 tier means are adjacent and their order varies across lists, consistent with the R3/R2 boundary noted in Table 11; the R1 gap is unaffected on either transform. Removing the two documented noise sources lowers the correlation and leaves the ordering intact. The only list that inverts the R3/R2 tier means discards the prose phrases and counts only capitalized abbreviations; it measures table-header residue instead of prose restatement and is reported for completeness.

Table 5: Term-list sensitivity. Pearson correlation between the fact-level redundancy measure and the fact-level deviation correlation, recomputed under alternative term lists over the same 65 MLB and 64 NHL articles; the deviation correlations are held fixed. The R3/R2/R1 columns are tier-mean mentions per article, the quantity reported in Section 3.2. The first row of each panel is the list used throughout the paper.

Term list	n	r (rate)	r (log)	R3	R2	R1	mono.
MLB							
As reported (phrases + abbreviations)	9	+0.943	+0.896	12.52	7.81	0.17	yes
Phrases only	9	+0.815	+0.736	11.84	3.03	0.16	yes
Case-insensitive abbreviations	9	+0.943	+0.896	12.53	7.81	0.17	yes
Documented noise removed	9	+0.775	+0.738	12.52	6.53	0.12	yes
Noise removed, phrases only	9	+0.681	+0.614	11.84	1.76	0.11	yes
Abbreviations only (no prose)	9	+0.362	+0.498	0.68	4.77	0.01	no
NHL							
As reported (phrases + abbreviations)	9	+0.783	+0.770	19.56	0.50	0.01	yes
Phrases only	9	+0.783	+0.772	19.56	0.49	0.01	yes
Case-insensitive abbreviations	9	+0.783	+0.770	19.56	0.50	0.01	yes
Blocked shots narrowed	9	+0.783	+0.770	19.56	0.50	0.01	yes
Everyday-word facts dropped	6	+0.695	+0.647	13.02	0.27	0.01	yes

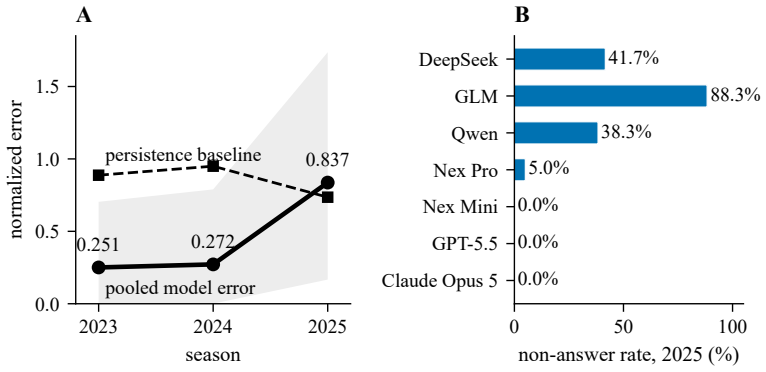

Figure 2: **Temporal boundaries in model responses. A.** Normalized error by season (common subset, ≥ 5 models answering); bold line is pooled mean, dashed line is B_{persist}, shaded band spans the model range. **B.** Non-answer rate by model for the 2025 season (2023 and 2024 are 0.0% on all models).

D Complete tables

Entity prominence is associated with lower error at high redundancy but cannot produce the R1 drop, because the Equal-Depth construction holds it fixed. By prominence stratum, the MAE comparisons are star d=1 0.0048 vs. B_0 0.0345; bench d=1 0.0084 vs. 0.0258; star d=2 model 0.0271 vs. B_2 0.0181 (bench 0.0316 vs. 0.0208); and star d=3 model 0.0475 vs. 0.0355 (bench 0.0528 vs. 0.0419). Across prominence quartiles (median pageviews 44,643 / 104,334 / 226,712 / 599,953), d=1 MAE is 0.0102 / 0.0085 / 0.0060 / 0.0030, and the relative improvement over B_0 is 47.3% / 66.3% / 81.0% / 93.5%.

E Player selection

MLB. For each season we rank qualified batters by plate appearances and take the top 20 as the high-prominence stratum; the low-prominence stratum is 20 players drawn uniformly at random (fixed seed) from those with 400–800 plate appearances. Players with fewer than 200 at-bats are excluded, and a split cell is built only if it rests on at least 40 at-bats. The two pools are defined on overlapping plate-appearance ranges, so five player-seasons qualify for both (two in 2023, three in

Table 6: Per-model deviation correlation and MAE(z) by redundancy tier (EQUAL-DEPTH, MLB). For every model, the R1 correlation lies far below the R2 and R3 correlations, and the size of the R1 drop varies across models.

	R1 low		R2 mid		R3 high	
Model	ρ_{dev}	MAE(z)	ρ_{dev}	MAE(z)	ρ_{dev}	MAE(z)
DeepSeek	$+0.178$	1.285	$+0.966$	0.062	$+0.929$	0.078
GLM	$+0.065$	1.345	$+0.827$	0.439	$+0.939$	0.161
Qwen	$+0.080$	1.183	$+0.639$	0.645	$+0.758$	0.556
Nex-Pro	$+0.179$	1.152	$+0.934$	0.229	$+0.942$	0.182
Nex-Mini	$+0.230$	0.979	$+0.694$	0.587	$+0.761$	0.514
GPT-5.5	$+0.403$	0.814	$+0.995$	0.031	$+0.999$	0.007
Claude Opus 5	$+0.342$	0.834	$+0.971$	0.143	$+0.982$	0.079

Table 7: Per-model deviation correlation and MAE(z) by redundancy tier, NHL. Same seven models and scoring as Table 6.

	R1 low		R2 mid		R3 high	
Model	ρ_{dev}	MAE(z)	ρ_{dev}	MAE(z)	ρ_{dev}	MAE(z)
DeepSeek	$+0.583$	0.788	$+0.882$	0.204	$+0.914$	0.169
GLM	$+0.560$	0.695	$+0.843$	0.387	$+0.981$	0.097
Qwen	$+0.326$	1.158	$+0.612$	0.707	$+0.893$	0.339
Nex-Pro	$+0.572$	0.728	$+0.777$	0.457	$+0.976$	0.140
Nex-Mini	$+0.346$	1.234	$+0.677$	0.582	$+0.874$	0.378
GPT-5.5	$+0.820$	0.441	$+0.949$	0.147	$+0.999$	0.008
Claude Opus 5	$+0.727$	0.525	$+0.855$	0.347	$+0.979$	0.117

2024) and are assigned to the low-prominence stratum; the realized split is therefore 35 high and 40 low rather than 40 and 40. The five overlapping player-seasons average 5.59 in $\log_{10}$ annual pageviews, above the high stratum's mean of 5.43 and the low stratum's mean of 4.99, so assigning them to the low stratum lowers the high mean, raises the low mean, and compresses the contrast in Table 4. Excluding them makes the correlations slightly stronger at R3 and leaves them unchanged at R2 and R1 (-0.105, -0.021, $+0.042$ with the five excluded vs. -0.096, -0.023, $+0.043$ with them retained, on the same scoring basis), so the reported values are conservative. Deduplication leaves 38 units in 2023 and 37 in 2024, 75 in total over 65 distinct players.

NHL. For 2022–23 and 2023–24 we take the top 20 scorers by points as the high-prominence stratum, and, among skaters with at least 60 games played, the 20 with the lowest ice time per game as the low-prominence stratum; ties break deterministically. The two criteria cannot intersect, so each season is 20 and 20, giving 80 units over 64 distinct players, 16 of whom appear in both seasons.

The two domains therefore use different low-prominence constructions (a random draw from a volume band in MLB, an extreme-value cut in NHL), so star–bench effect sizes are not directly comparable across domains.

F PREREGISTERED JUDGMENT CRITERIA AND VERDICTS

The criteria below are quoted verbatim from the registration made before the EQUAL-DEPTH run.

> R1 permutation $p \geq 0.05 \Rightarrow$ redundancy effect. $p < 0.05$ and $\rho_{\text{dev}} \geq 0.3 \Rightarrow$ depth effect survives. Anything in between $\Rightarrow$ intermediate verdict.

The observed MLB value is R1 $\rho_{\text{dev}} = +0.163$ with permutation $p < 0.0002$, which falls in the intermediate band ($p < 0.05$ but $\rho_{\text{dev}} < 0.3$). The verdict reported throughout this paper, intermediate and favoring the redundancy account, is the output of this rule. The tier correlation declines by 82% from R3 to R1, and any independent depth effect is bounded above by a correlation of about 0.16.

Table 8: Per-model deviation correlation and normalized error by redundancy tier, country statistics (unit-scale errors excluded). In every model, R1 error exceeds R2 error and R3 error is lowest.

Model	R1 ρ_{dev}	MAE(z)	R2 ρ_{dev}	MAE(z)	R3 ρ_{dev}	MAE(z)
DeepSeek	$+0.889$	0.301	$+0.988$	0.091	$+1.000$	0.009
GLM	$+0.856$	0.313	$+0.982$	0.127	$+1.000$	0.012
Qwen	$+0.788$	0.421	$+0.817$	0.221	$+0.943$	0.134
Nex-Pro	$+0.765$	0.415	$+0.974$	0.129	$+1.000$	0.010
Nex-Mini	$+0.883$	0.390	$+0.782$	0.268	$+0.991$	0.033
GPT-5.5	$+0.915$	0.246	$+0.986$	0.104	$+1.000$	0.010
Claude Opus 5	$+0.903$	0.306	$+0.980$	0.125	$+1.000$	0.010

Table 9: Per-model deviation correlation and normalized error by redundancy tier, corporate finance (unit-scale errors excluded). In every model, R1 error exceeds R2 error; the R1 correlation of GLM reflects the debt-to-equity ratio of a company with negative equity (Table 14).

Model	R1 ρ_{dev}	MAE(z)	R2 ρ_{dev}	MAE(z)	R3 ρ_{dev}	MAE(z)
DeepSeek	$+0.971$	0.100	$+0.934$	0.077	$+0.999$	0.009
GLM	-0.007	0.520	$+0.910$	0.098	$+0.989$	0.036
Qwen	$+0.578$	0.405	$+0.920$	0.103	$+0.943$	0.100
Nex-Pro	$+0.852$	0.254	$+0.917$	0.102	$+0.975$	0.062
Nex-Mini	$+0.453$	0.491	$+0.662$	0.458	$+0.556$	0.776
GPT-5.5	$+0.957$	0.138	$+0.944$	0.074	$+0.982$	0.054
Claude Opus 5	$+0.804$	0.330	$+0.942$	0.102	$+0.975$	0.070

The two frontier models individually exceed the 0.3 threshold (pooled $\rho_{\text{dev}} = +0.364$), consistent with their higher R1/R3 ratio (Appendix I). The preregistered analysis pools all seven models and gives the conservative estimate.

Applied to the NHL replication, the same rule gives $\rho_{\text{dev}} = +0.510$ with permutation $p < 0.0002$, which falls in the *depth effect survives* band. We report both preregistered verdicts as they stand. Neither verdict changes under any leave-one-fact-out or leave-one-model-family-out refit (Appendix I). The verdicts rest on raw tier correlations; the reconstruction-controlled analysis of Section 5.4 is a separate analysis and does not change either verdict.

G Robustness checks

Redundancy within a statistic. Mention counts also vary across players for the same statistic. At the instance level (MLB: 657 instances; NHL: 630), a regression of absolute normalized error on $\log(1 + \text{mentions in the player's own article})$ with statistic fixed effects, $\log_{10}$ pageviews, and $|y - B_0|/s_j$ gives a coefficient of -0.032 in MLB and -0.023 in NHL; a permutation null that shuffles mention counts within statistic and season gives one-sided $p = 0.005$ and 0.001. More mentions of a statistic in a player's article go with lower error for that player, holding popularity fixed.

Redundancy measurement scope. Wikipedia prose captures how frequently a statistic is discussed in narrative text. Because narrative conventions differ across sports, the measure is designed to order tiers within a domain. In both MLB and NHL, mean mentions per article differ by more than 40-fold between R2 and R1 (Tables 11 and 12).

The format example does not anchor answers. Across 450 paired calls, removing the prompt's format example ("such as 0.250") changes d=1 MAE from 0.0078 to 0.0085, and the paired-bootstrap interval $[-0.00069, +0.00195]$ includes zero. The point estimate indicates higher error without the example, so the example does not pull answers toward the league mean.

Source structure in the NHL domain. In the NHL data, six statistics (goals, assists, points, penalty minutes, power-play goals, shots) belong to the NHL API's *summary* report, while the three low-tier

Table 10: POST-CUTOFF per-model normalized error, with 2025 refusal and bare-zero rates shown separately. All error columns are computed on the common subset of items answered by at least five models (60, 60, and 53 items in 2023, 2024, and 2025), so the persistence baseline is shared across models, as in Section 5.5. Boldface in the 2025 error column marks models worse than persistence. n_{2025} is the number of the model's sixty 2025 items that were scored and is set in bold where it is below ten; GLM's 2025 error rests on seven items. The Mean row averages the seven models; the 2025 error pooled over items is 0.837 (Section 5.5).

	Model			B_{persist}			2025 non-answer		
Model	2023	2024	2025	2023	2024	2025	refusal	null	n_{2025}
DeepSeek	0.045	0.077	**1.737**	0.888	0.950	0.736	25.0%	16.7%	35
GLM	0.156	0.093	**1.021**	0.888	0.950	0.736	23.3%	65.0%	**7**
Qwen	0.524	0.595	**1.097**	0.888	0.950	0.736	35.0%	3.3%	37
Nex-Pro	0.240	0.251	**1.059**	0.888	0.950	0.736	0.0%	5.0%	57
Nex-Mini	0.704	0.790	**1.092**	0.888	0.950	0.736	0.0%	0.0%	60
GPT-5.5	0.001	0.001	0.167	0.888	0.950	0.736	0.0%	0.0%	60
Claude Opus 5	0.085	0.095	0.236	0.888	0.950	0.736	0.0%	0.0%	60
Mean	0.251	0.272	0.916	0.888	0.950	0.736	11.9%	12.9%	45.1

Table 11: Fact-level redundancy measurement, MLB (65/65 Wikipedia articles retrieved). Counts for "double play" include defensive mentions, so the GIDP count is overstated. Mention counts overlap between R3 and R2, and the main contrast separates R1 from both upper tiers (tier-mean ratios R2/R1 $\approx 45\times$ and R3/R1 $\approx 72\times$).

Tier	Fact	Mention rate	Mentions/article	ρ_{dev}	perm. p
R3	avg	76.9%	2.48	+0.908	<0.0005
R3	hr	98.5%	33.00	+0.923	<0.0005
R3	ops	64.6%	2.08	+0.859	<0.0005
R2	rbi	98.5%	15.17	+0.837	<0.0005
R2	sb	70.8%	4.35	+0.911	<0.0005
R2	bb	80.0%	3.89	+0.814	<0.0005
R1	hbp	10.8%	0.14	+0.183	0.001
R1	sf	13.8%	0.18	+0.102	0.049
R1	gidp	9.2%	0.20	+0.192	0.001
Tier means (mentions/article): R3 12.52			R2 7.81		R1 0.17 (monotone).

statistics (blocked shots, takeaways, giveaways) belong to its *real-time* report. The MLB experiment isolates redundancy because all nine facts come from one row of one source. The tier ordering holds in both domains, and within the NHL low tier, where all three facts come from the same report, reconstructibility accounts for the ordering of recall (Section 5.4).

NHL pooled results and the prominence strata. MLB draws its two prominence strata from overlapping plate-appearance bands, whereas NHL contrasts the top twenty scorers with the lowest-ice-time skaters, which produces a wider prominence spread. The NHL two-factor correlations are 0.329 (star) and 0.297 (bench) at R1 against a pooled 0.510, 0.749 and 0.654 at R2, and 0.759 and 0.556 at R3. Applying the Freedman–Lane control within each stratum under the same conventions ($\text{NPERM} = 5000$, unit-level block permutation within season, per-fact residualization), the R2-to-R1 drop holds in both strata, $0.471 \to 0.252$ in the high-prominence stratum and $0.461 \to 0.046$ in the low-prominence stratum, against $0.511 \to 0.209$ pooled. The drop is at least as large in the low-prominence stratum, so the stratification does not produce it. The residual low-tier signal comes from prominent players, at $+0.252$ ($p < 0.0002$) for stars and $+0.046$ ($p = 0.181$) for bench players, where the 95th percentile of the block null is 0.081.

Prominence as a covariate. Across all 73 units with pageview data, prominence correlates with d=1 error ($r = -0.383$, $p < 0.001$); the effect concentrates among stars ($r = -0.413$, $p = 0.012$)

Table 12: Fact-level redundancy measurement, NHL (64/64 player articles retrieved). Mention rate is the fraction of articles containing the term; ρ_{dev} is the per-fact deviation correlation against B_0. Because "goals", "points", and "shots" are common English words, the R3 and R2 counts include non-statistical uses and are overstated. The hockey senses of "takeaway" and "giveaway" are essentially absent from prose.

Tier	Fact	Mention rate	Mentions/article	ρ_{dev}	perm. p
R3	goals	95.3%	24.50	$+0.925$	<0.0005
R3	assists	85.9%	13.02	$+0.936$	<0.0005
R3	points	92.2%	21.16	$+0.963$	<0.0005
R2	pim	15.6%	0.30	$+0.617$	<0.0005
R2	ppg	17.2%	0.25	$+0.798$	<0.0005
R2	shots	42.2%	0.95	$+0.923$	<0.0005
R1	blk	3.1%	0.03	$+0.288$	<0.0005
R1	tka	0.0%	0.00	$+0.645$	<0.0005
R1	gva	0.0%	0.00	$+0.642$	<0.0005

Table 13: Fact-level redundancy measurement, country statistics (25 countries). Mention rate (the fraction of country articles containing the term) and mentions per article are counted with the same Wikipedia text-matching procedure as in the sports domains. ρ_{dev}: pooled deviation correlation across seven models under the protocol of Table 2 (unit-scale errors excluded); ρ_s: Spearman correlation on the same records; n: pooled sample size.

Tier	Fact	Mention rate	Mentions/article	ρ_{dev}	ρ_s	n
R3	population	1.00	12.84	$+1.000$	$+0.999$	175
R3	GDP	0.92	3.32	$+0.979$	$+0.864$	171
R2	life expectancy	0.88	1.56	$+0.973$	$+0.958$	175
R2	unemployment	0.44	0.64	$+0.865$	$+0.844$	175
R1	inflation	0.04	0.20	$+0.909$	$+0.845$	175
R1	Gini coefficient	0.33	0.75	$+0.795$	$+0.763$	168
Tier means (mentions/article): R3 8.08		R2 1.10	R1 0.48.			

while bench players show no graded relationship ($p = 0.554$). Its association with lower error is confined to R3 and reverses at R1 (Appendix D).

H In-sample computation of the tier baseline

In Equal-Depth and its NHL replication, B_0 is the fact-season median across the probed players, so it is computed from the same true values against which the models are scored. This construction cannot produce the tier ordering because all three tiers share it, and recomputing B_0 leave-one-player-out raises every tier-level correlation (Appendix L). The cutoff and depth claims use baselines that do not depend on the scored values.

I Model capability and the lowest-tier drop

The size of the R1 drop (Section 5.3) varies across models and is associated with model capability, most clearly in NHL. In ascending order, the per-model R1/R3 ratios in MLB are 0.069 (GLM), 0.105 (Qwen), 0.190 (Nex-Pro), 0.192 (DeepSeek), 0.302 (Nex-Mini), 0.348 (Claude Opus 5), and 0.403 (GPT-5.5), and in NHL they are 0.365 (Qwen), 0.396 (Nex-Mini), 0.571 (GLM), 0.587 (Nex-Pro), 0.638 (DeepSeek), 0.743 (Claude Opus 5), and 0.821 (GPT-5.5).

Two checks support reading this as a capability ordering. First, capability is measured outside the probed answers. Correlating R1 retention with a model's own R3 score would share noise, because both come from the same answers, so we use each model's normalized error on the Post-Cutoff 2023 season, which comes from a different experiment with a different player sample and fact set.

Table 14: Fact-level redundancy measurement, corporate finance (25 S&P 500 companies). Counts use the same text-matching procedure, and ρ_{dev} and Spearman ρ_s are computed under the protocol of Table 2 (unit-scale errors excluded). R1 facts receive zero Wikipedia mentions. The debt-to-equity correlation includes one company with negative stockholders' equity (truth -12.9), for which three of the seven models answer with a positive value.

Tier	Fact	Mention rate	Mentions/article	ρ_{dev}	ρ_s	n
R3	revenue	0.84	3.08	$+0.871$	$+0.879$	175
R3	net income	0.32	0.44	$+0.867$	$+0.851$	174
R2	operating income	0.28	1.44	$+0.732$	$+0.843$	173
R2	total assets	0.04	0.08	$+0.997$	$+0.935$	173
R1	current ratio	0.00	0.00	$+0.821$	$+0.790$	168
R1	debt-to-equity	0.00	0.00	$+0.347$	$+0.691$	175
Tier means (mentions/article): R3 1.76 R2 0.76 R1 0.00.						

The correlation is -0.38 in MLB and -0.90 in NHL, negative because lower error means higher capability; the association is clear in NHL and weaker in MLB. Second, the R1 advantage is not explained by reconstruction. In MLB paired tests against the reconstruction baseline of Section 5.4, the two frontier models exceed the fitted B_{recon} at R1 by $\Delta\rho_{\mathrm{dev}} = +0.464$ and $+0.398$ ($p < 0.001$); of the other five, three exceed it at $p < 0.05$ (Nex-Mini $+0.283$, $p = 0.001$; DeepSeek $+0.234$, $p = 0.003$; Nex-Pro $+0.233$, $p = 0.012$) and two do not (GLM, Qwen).

Stability of the capability ordering. Bootstrapped over models, these correlations are negative in 88.5% (MLB) and 99.9% (NHL) of resamples, and they are negative in every leave-one-model-out refit in both domains.

Leave-one-out refits. Dropping each fact in turn keeps the MLB R1 correlation in $[0.148, 0.187]$ around 0.163, with permutation $p \leq 0.001$ throughout, and the NHL R1 correlation in $[0.437, 0.649]$ around 0.510. No single-fact removal changes either domain's preregistered verdict, including the removal of hit-by-pitch, sacrifice flies, or grounded-into-double-play in MLB. Dropping each independent model family in turn moves pooled R1 within $[0.140, 0.184]$ (MLB) and $[0.474, 0.558]$ (NHL) and leaves every tier ordering intact; the two Nex models share weights and are dropped as a single family. Bootstrapping the tier-level correlations over $(\text{player}, \text{season})$ units gives 95% intervals that are disjoint between R1 and R3 in both domains, with MLB R1 $[0.111, 0.228]$ against R3 $[0.848, 0.924]$ and NHL R1 $[0.439, 0.573]$ against R3 $[0.923, 0.956]$.

Cross-fitted baselines. Recomputing B_0 and its normalizing SD leave-one-player-out, including inside every permutation, *raises* every tier-level correlation (MLB R1 $0.163 \rightarrow 0.205$; NHL R1 $0.510 \rightarrow 0.571$). The in-sample baseline used in the main text is therefore conservative, and the tier contrasts do not depend on scoring against a median that includes the scored unit.

J Repeat sampling and answer dispersion

The main-text results use one sample per cell at `temperature`$= 1.0$. To measure sampling variability, we re-sampled a preregistered subset (40 cells per tier in MLB, drawn by fixed seed from cell identifiers alone, three additional draws each) on all seven models and measured dispersion within each (model, cell). The results below use the two frontier models, as in Section 5.6; the seven-model results are given at the end of the section.

Answer consistency also follows redundancy. Across four samples of the same cell, the fraction of cells returning an *identical* number is 85.0% at R3, 67.5% at R2, and 5.0% at R1, with mean within-cell SD of 0.024, 0.049, and 0.442 in z units. At high redundancy the models are nearly deterministic at temperature one, while at low redundancy their answers vary from sample to sample.

Averaging the four samples, which removes most of this noise, changes the correlations by $+0.002$ (R3), -0.001 (R2), and $+0.103$ (R1). The low tier gains the most, as expected if single-sample noise attenuates it, but the gain is small relative to the tier gap. On this subset the tier-level correlations are $0.990/0.970/0.470$ with one sample and $0.993/0.969/0.573$ with four. Across all seven models

(280 cells per tier), identical answers occur in 50.0%, 32.9%, and 6.1% of cells, and averaging changes R1 by +0.002. The R1 drop therefore reflects what the models know and does not arise from single-sample decoding.

K ALTERNATIVE DETECTOR SCORES

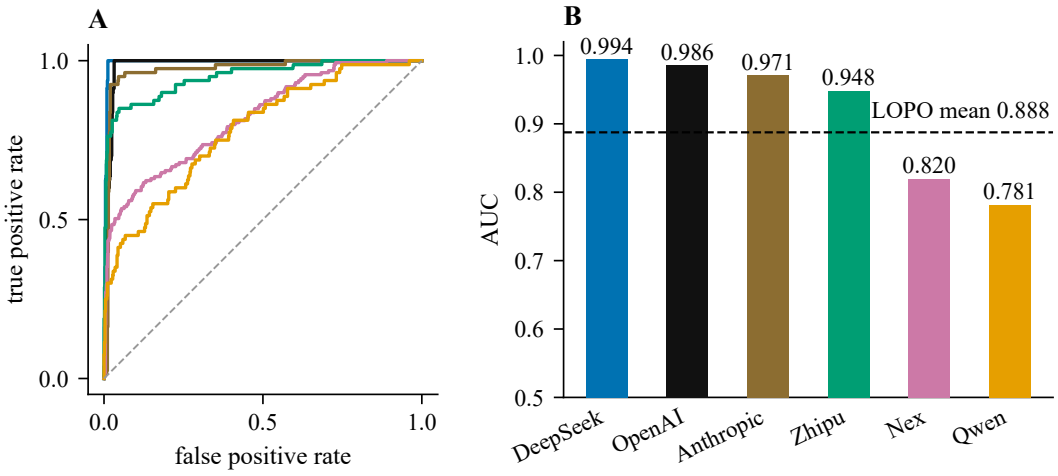

Figure 3: **Memorization detector operating characteristic** (depth-labeled variant). Redundancy-labelled pooled AUCs are reported in Section 5.6. **A.** Pooled ROC per model family (score $= |a-y|$; positive class d=1, negative d=2+3); colors as in B. **B.** Pooled AUC per model family; the dashed line is the leave-one-player-out mean.

Failure mode at the redundancy boundary. Neither frontier model declines any of the 930 low-redundancy items they were asked. Some fabrications fall far outside the physical range of the statistic. For gidp and hbp, DeepSeek returned bare integers such as 73, 100, and 130, whereas no probed value of these statistics exceeds 23; these answers are retained in the analysis (Appendix B). Larger models have been reported to predict with useful calibration whether they can answer (Kadavath et al., 2022); at this boundary the models show no such calibration.

To test whether simpler behavioral signals account for the detector's discrimination, we apply the leave-one-player-out AUC protocol of Section 5.6 to three candidate scores on the same released records, without new model calls.

Table 15: Leave-one-player-out AUC for candidate detector scores (positive class d=1, negative d=2+d=3, n = 65 players). †The last row is a pooled AUC over the 159 R3 and R1 MLB cells repeat-sampled on the two frontier models, with redundancy tiers as classes and is not directly comparable to the other rows; the repeat subset gives most players too few cells for a per-player estimate.

Score	Mean	Median	Min
$\|a - y\|$ (deviation from truth; the probe's score)	**0.888**	0.891	0.751
Refusal indicator	0.500	0.500	0.500
$\|a - B\|$ (single-response distance from baseline)	0.427	0.455	0.000
Dispersion over four samples (no ground truth)	**0.952**†	—	—

The refusal score is at chance because the models almost never refuse within their training period (Section 5.3). The baseline-distance score, which also needs no ground truth, is slightly *below* chance, because fabricated answers depart from the baseline as much as recalled ones do, so whether an answer departs from the typical value carries no information about memorization. Among scores computed from a *single* response, only distance from the truth discriminates; neither refusal behavior nor departure from the typical value is informative on its own. Four responses to the same item

change this. Their dispersion, which also needs no ground truth, separates the top tier from the bottom tier at AUC 0.952 on the 159 R3 and R1 MLB cells repeat-sampled on the two frontier models, against 0.957 for the ground-truth score on the same cells, and the rule that all four answers are identical gives TPR 0.850 at FPR 0.051 and precision 0.944. Separation is weak where the tiers are close (R3 vs. R2, AUC 0.590) and strong where they differ (R2 vs. R1, 0.934). We report the pooled value because the repeat subset gives most players only four cells, so per-player leave-one-player-out AUCs would rest on two-against-two comparisons.

L ATTENUATION SIMULATION AND THE RECONSTRUCTION BASELINE

Reconstructibility specification. The multiple correlation R for each statistic is computed as the Pearson correlation between season-standardized (z) values and fitted values from an ordinary least-squares regression of that statistic on the three R3 statistics (batting average, home runs, OPS) of the same (entity, period) unit. The regression is fit on all $N = 75$ equal-depth units (combining 2023 and 2024 seasons). Units with a player appearing in both seasons contribute two observations; excluding duplicates ($N = 65$) changes the six non-headline R values by at most 0.06. No cross-validation is applied because R serves as a descriptive measure of linear association among published statistics, not as a predictive score. Tier averages of the resulting values are reported in Section 5.4.

Attenuation simulation. To test whether the R1 drop is a mechanical consequence of the low tier's number scale, we simulate what the probe would measure if R1 recall were exactly as good as R3 recall. For each model we resample its own R3 deviation residuals (in z units) onto R1 true values, invert the normalization, round to integers (R1 facts are counts), and recompute the tier correlation; 1,000 draws, seed fixed. A perfect-recall oracle scores $\rho_{\text{dev}} = 1.000$ on R1, so integer-valued targets lose no information to quantization.

Table 16: Equal-noise simulation of the R1 ρ_{dev} the probe would measure if R1 recall matched R3 recall. The observed R1 ρ_{dev} is 0.163.

Simulation variant	median	$p_{2.5}$	min	observed quantile
R3 residual pool, rounded	0.913	0.902	0.894	0.000
R3 residual pool, unrounded	0.907	0.891	0.882	0.000
HR-only residual pool (integer counts), rounded	0.935	0.924	0.918	0.000

The observed 0.163 lies below the minimum of every variant, and rounding *raises* the simulated correlation ($0.907 \rightarrow 0.913$) instead of attenuating it. The same holds for each fact (simulated medians $\approx$0.91 against observed 0.10–0.19). Range restriction and granularity therefore cannot account for the R1 drop.

Leave-one-player-out B_0. B_0 in EQUAL-DEPTH is the fact-season median across the probed players, so each unit contributes to the baseline it is scored against. Recomputing B_0 (and the normalizing SD) from the other probed units of the same season (37 in 2023 and 36 in 2024) for every unit removes that dependence. Over 4,700 valid records the tier correlations are $+0.905$ (R3), $+0.871$ (R2), and $+0.205$ (R1), against $+0.894$, $+0.855$, and $+0.163$ in sample. All three correlations rise, by at most 0.042, and the ordering is unchanged. The in-sample baseline slightly lowers the correlations, so the values reported in the main text are conservative.

Historical reconstruction baseline. A leave-one-player-out ridge regression (Hoerl & Kennard, 1970; Stone, 1974) predicts each statistic from the previous-season record. Models exceed it at R3 ($\rho_{\text{dev}} = 0.893$ vs. 0.514; paired $\Delta = +0.385$ on the same records, $p < 0.0002$) and R2 (0.853 vs. 0.664). At R1 the baseline has no predictive value (-0.064), so low-tier answers do not come from extrapolating the previous season.

Reconstruction baseline details. The reconstruction family is *linear* and has three members: the fitted ridge baseline B_{recon} on the previous season's row, a ridge fit on the previous-season value of the same statistic (B_{prev}), and the previous-season row plus current-season $\log_{10}$ pageviews. B_{recon} is a leave-one-player-out ridge regression on the previous season's full statistical row, with λ selected by an inner leave-one-out grid (median $\lambda = 10$). The family is built for MLB only; the partial-correlation control covers NHL. It is evaluated on the 73 of 75 units that have a previous

season (71 for the pageview variant, since one player lacks pageview data). The two excluded units are rookies (Joey Wiemer, Jordan Walker 2023), and all model figures in Section 5.4 are recomputed on the same 73-unit subset (R3 ρ_{dev} 0.893 on the 73-unit subset vs. 0.894 on the full set). Per-model paired tests confirm the pooled result at R3 for all seven models; at R2, five of seven are significant (Qwen, $\Delta = -0.029$, and Nex-Mini, $\Delta = +0.024$, are not). At R1 the models separate. The two frontier models exceed the baseline at $\Delta = +0.464$ and $+0.398$ ($p < 0.001$), three of the remaining five exceed it by smaller margins (Nex-Mini $+0.283$, $p = 0.001$; DeepSeek $+0.234$, $p = 0.003$; Nex-Pro $+0.233$, $p = 0.012$), and two do not. At R1 the pooled models reach 0.156 against 0.208 for B_{prev}, and their MAE(z) of 1.097 is worse than every variant, so at R1 the models do not beat a fit to last season's value of the same statistic. The pooled R1 correlation exceeds the fitted B_{recon} by $+0.218$; B_{recon} is itself anti-correlated at R1 ($\rho_{\text{dev}} = -0.064$; -0.386 on hit-by-pitch).