The Cost of Compression: A Rate-Distortion Limit on Factual Hallucination

Xi Wang[1,2,*] **Shijia Xu**[3,*,†] **Rongfeng Guo**[4]
[1]Hefei Institutes of Physical Science, Chinese Academy of Sciences, China
[2]University of Science and Technology of China, China
[3]Chongqing University, China [4]Shenzhen University, China
xw_cs@mail.ustc.edu.cn shijiaxu@stu.cqu.edu.cn

Abstract

Factual hallucination in closed-book question answering is often treated as a coverage problem: a model fails because the relevant fact is absent from its internal memory. This view misses a second source of error. Even when a fact has been observed, finite memory may force it to be stored only approximately. We study this effect through a simple coverage–compression model of factual recall.

We consider an unstructured question-answering task with N possible queries and K possible answers. A learner observes M training facts, compresses them into at most B bits, and answers uniformly drawn test queries without retrieval. For a uniformly random ground-truth mapping, we prove

$$\mathcal{E} \geq \frac{M}{N}\,\delta^{\star}\left(\frac{B}{M}\right) + \left(1-\frac{M}{N}\right)\left(1-\frac{1}{K}\right),$$

where $\delta^{\star}(r)$ is the inverse rate-distortion function of a uniform K-ary source under zero-one loss. The two terms separate compression distortion on observed facts from missing coverage on unobserved facts.

The bound gives a compact way to reason about selective memory, forced compression, structure, retrieval, abstention, and long-context organization. We study the predicted signatures with theory-implied simulations and controlled fact-injection probes in modern language models that vary fact load and effective trainable memory. The result is not a complete theory of hallucination, but an information-theoretic account of a separable failure mode: lossy recall of observed facts under finite memory.

1 Introduction

Large language models can answer many factual questions without explicit retrieval, suggesting that some factual knowledge is stored in their parameters. Yet factual errors remain common across short-form, long-form, knowledge-graph, dialogue-level, and in-the-wild factuality evaluations (Wei et al., 2024a,b; Liu et al., 2024b; Luo et al., 2024; Zhao et al., 2024; Bayat et al., 2024; Jacovi et al., 2025; Liu et al., 2025). A standard explanation is missing coverage: the relevant fact was absent from the model's training data, inaccessible to its parameters, or unavailable at inference time. This explanation motivates retrieval and grounding methods, which move part of the factual burden from parametric memory to external evidence (Yan et al., 2024; Edge et al., 2024; Wang et al., 2024; Ru et al., 2024; Lee et al., 2025; Wu et al., 2025). Recent work further shows that retrieval quality depends not only on access to external evidence but also on robustness to upstream noise and on how cross-document evidence is organized at indexing time (Bao, 2026; Bao and Shi, 2026). It is important, but incomplete.

A fact can be observed and still be recalled incorrectly. A compressed database gives the basic intuition: if a record is absent, the system must guess; if the record is present but stored in a lossy code, the system may retrieve the wrong value. Closed-book factual recall faces the same tension. A model must compress many facts into a finite internal representation, and recent controlled studies of new factual knowledge, knowledge editing, memorization, and knowledge capacity suggest that factual storage is itself a constrained resource (Gekhman et al., 2024; Huang et al., 2024; Allen-Zhu and Li, 2024; Morris et al., 2025; Zhou et al., 2025). Related systems have also begun to make memory provenance and knowledge extension explicit design objects in long-horizon agents and reasoning systems (Zhang et al., 2026b; Lou et al., 2025). The resulting error is not merely a failure of exposure, but a failure of recall under compression.

We isolate this mechanism with a deliberately

[*]Equal contribution.
[†]Corresponding author.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 4579–4598
November 6-10, 2026 ©2026 Association for Computational Linguistics

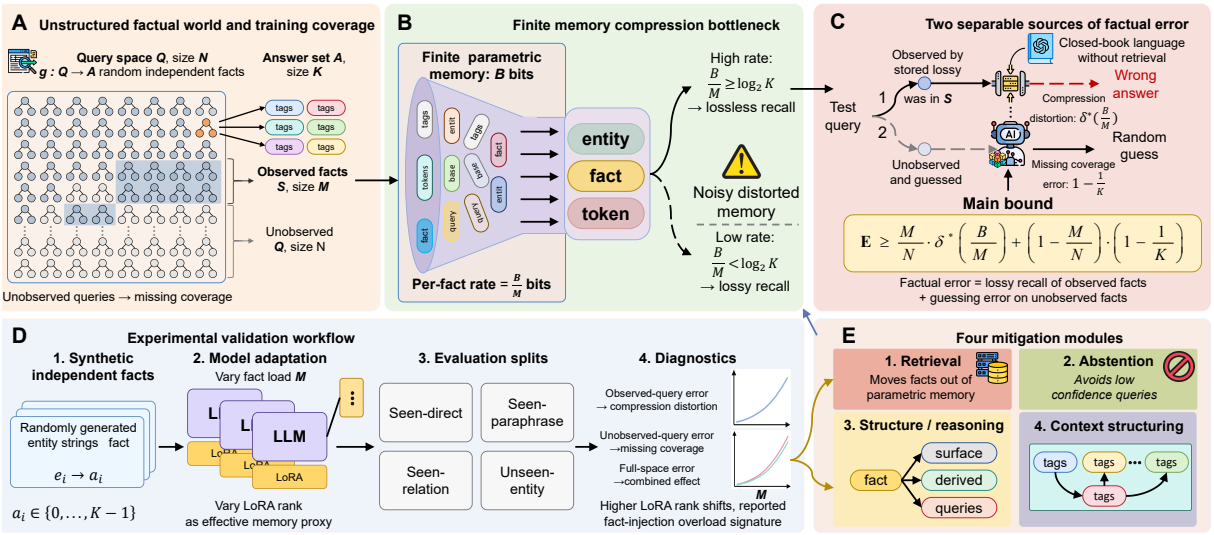


Figure 1: Coverage–compression framework. Closed-book factual error decomposes into missing coverage on unobserved facts and compression distortion on observed facts stored under finite memory. The same diagnostics are used in controlled fact-injection experiments: observed-query error measures lossy recall, unobserved-query error measures missing coverage, and full-space error measures their mixture.

simple model, summarized in Figure 1. Let $\mathcal{Q}$ be a query set of size N and $\mathcal{A}$ an answer set of size K. The ground-truth mapping $g : \mathcal{Q} \to \mathcal{A}$ is drawn uniformly at random, so the answers contain no semantic regularity that a learner can exploit. A learner observes M training pairs, encodes the observed information into at most B bits, and answers uniformly drawn test queries without retrieval. This random-source setting is not a model of all factual knowledge; it is a null model for unstructured factual memory, where correct prediction must come from stored information. For example, when $K = 10$, one lossless answer requires $\log_2 10 \approx 3.32$ bits, so a 160-bit answer memory stores only about 48 independent facts under the optimistic free-addressing convention.

The central claim is that closed-book factual error has two separable sources: the queried fact may not be effectively covered by memory, or it may be covered but stored too lossily to be recalled correctly. Conditioned on the training set, the observed labels are M independent K-ary symbols. The learner's memory is a compressed representation of this sequence. For a uniform K-ary source under zero-one loss, let $\delta^\star(r)$ denote the minimum achievable reconstruction error when each symbol is encoded at rate r bits. Our main result, stated formally in Theorem 3.1, gives

$$\mathcal{E} \geq \frac{M}{N}\,\delta^\star\left(\frac{B}{M}\right) + \left(1 - \frac{M}{N}\right)\left(1 - \frac{1}{K}\right).$$

The first term is the distortion on observed facts; the second is the guessing error on unobserved queries. When $B/M \geq \log_2 K$, observed facts can be stored losslessly and the bound reduces to the usual unseen-query floor. When $B/M < \log_2 K$, even facts seen during training have a nonzero recall error floor.

This decomposition clarifies several behaviors that are often discussed separately. Finite memory alone does not imply that more data must hurt an optimal learner: with selective allocation, additional observations can simply produce diminishing returns. Compression overload appears when many observed facts are forced to share a fixed effective representation, lowering the per-fact rate B/M. Structure changes the trade-off by reducing the number of independent facts that must be stored. Long-context organization gives an analogous working-memory problem, where facts are present in the prompt but compete for limited effective attention; recent long-context evaluations show that retrieval becomes substantially harder when context length, task complexity, latent structure, long in-context learning, or non-literal matching increases (Hsieh et al., 2024; Zhang et al., 2024; Bai et al., 2024; Li et al., 2024a; Vodrahalli et al., 2024; Li et al., 2024b; Modarressi et al., 2025). Related work on ultra-long agent trajectories and controllable summarization further highlights the importance of managing long-context information and compressed intermediate representations (Liu et al., 2026c,b).

We evaluate the framework through predicted signatures rather than aggregate benchmark scores. Theory-implied simulations instantiate the exact distortion-rate envelope, while controlled fact-injection probes vary the number of independent synthetic facts and the effective trainable memory budget in modern language models. Across both components, we report observed-query error for compression distortion, unobserved-query error for missing coverage, and full-space error for their mixture.

Taken together, the paper makes three contributions. First, it identifies a compression-induced source of factual error that is separate from missing coverage. Second, it proves a rate-distortion lower bound showing how observed-fact distortion and unseen-query guessing combine in a single coverage–compression trade-off. Third, it turns the bound into a diagnostic framework, yielding measurable signatures in theory-implied simulations and controlled fact-injection probes. The claim is intentionally narrow: compression is not a complete theory of hallucination, but a separable mechanism that predicts when observed facts become unreliable under finite effective memory.

2 Problem Setting

We study closed-book factual recall under a finite memory budget. Let $\mathcal{Q}$ be a query set with $|\mathcal{Q}| = N$, and let $\mathcal{A}$ be an answer set with $|\mathcal{A}| = K \geq 2$. The ground-truth mapping is

$$g : \mathcal{Q} \to \mathcal{A}.$$

To isolate unstructured factual memory, we draw g uniformly at random from all K^N mappings. Equivalently, the labels $\{g(q) : q \in \mathcal{Q}\}$ are independent uniform K-ary symbols. This random-source model is not a claim that real knowledge is random; it is a null model in which semantic, logical, and distributional shortcuts have been removed.

A learner observes a uniformly sampled training set $S \subset \mathcal{Q}$ of size M, together with the labels

$$X_S = (g(q))_{q \in S}.$$

It encodes the observed information into a memory state $W \in \mathcal{W}$ with $|\mathcal{W}| \leq 2^B$. At test time, it receives a query q and must answer without retrieval. We allow arbitrary randomized encoders and decoders. Under the free-addressing convention, the decoder may condition on the training index set S and the query identity:

$$\psi : \mathcal{Q} \times \mathcal{S}_M \times \mathcal{W} \to \Delta(\mathcal{A}),$$
$$\mathcal{S}_M = \{S \subseteq \mathcal{Q} : |S| = M\}.$$

Let $\hat{g}(q)$ denote the answer sampled from, or deterministically chosen by, this decoder. The budget B is charged only to stored answer information; query identities and the set S are treated as available addresses. This convention favors the learner. If query identities must also be stored, an additional address cost on the order of $\log_2 \binom{N}{M}$ is required.

We measure expected full-space error under the uniform test distribution:

$$\mathcal{E} = \mathbb{E}\left[\frac{1}{N}\sum_{q \in \mathcal{Q}} \mathbf{1}\{\hat{g}(q) \neq g(q)\}\right], \qquad (1)$$

where the expectation is over the random ground truth, the sampled training set, and learner randomness.

We distinguish two notions of coverage. A query has training coverage if $q \in S$. It has effective memory coverage if the learner has allocated enough information to recall its answer with nontrivial fidelity. A fact may therefore be present in the training set but still be recalled incorrectly.

3 Coverage–Compression Theory

The theory separates two sources of error. If a test query was not observed, its label is independent of the learner's memory and the learner must guess. If the query was observed, the learner still has to reconstruct its label from a finite-memory representation. The first effect gives a coverage floor; the second is a rate-distortion problem.

Coverage-only floor. If every observed fact were recalled perfectly, the only unavoidable error would come from queries outside S:

$$\mathcal{E} \geq \left(1 - \frac{M}{N}\right)\left(1 - \frac{1}{K}\right). \qquad (2)$$

For $q \notin S$, the label $g(q)$ is uniform over $\mathcal{A}$ and independent of (S, X_S, W), so no predictor can be correct with probability greater than $1/K$. This baseline treats observation as equivalent to lossless storage. We next remove that assumption.

Observed facts as source coding. Conditioned on S, the observed labels X_S are M independent uniform symbols over a K-ary alphabet. The memory state W is a compressed representation of this sequence, and answering an observed query reconstructs one coordinate of X_S from the compressed memory and the query address. The relevant distortion is factual recall error,

$$d(a, \hat{a}) = \mathbf{1}\{a \neq \hat{a}\}.$$

For a uniform K-ary source under zero-one distortion, the rate-distortion function is

$$R_K(D) = \log_2 K - h_2(D) - D \log_2(K-1), \quad 0 \leq D \leq 1 - \frac{1}{K}, \tag{3}$$

where $h_2(D) = -D \log_2 D - (1-D) \log_2 (1-D)$. This is the standard rate-distortion function for a uniform K-ary source under Hamming distortion (Shannon, 1959; Cover and Thomas, 2006). Thus perfect recall requires $\log_2 K$ bits per fact, while random guessing has distortion $1 - 1/K$ and requires zero bits. Let

$$\delta^\star(r) = \inf\{D \geq 0 : R_K(D) \leq r\} \tag{4}$$

denote the smallest possible reconstruction error at rate r bits per observed fact.

Theorem 3.1 (Coverage–compression bound)**.** *Under the random-source model and the free-addressing convention, every learner with memory budget B satisfies*

$$\mathcal{E} \geq \frac{M}{N} \delta^\star\left(\frac{B}{M}\right) + \left(1 - \frac{M}{N}\right)\left(1 - \frac{1}{K}\right). \tag{5}$$

Proof idea. Condition on S. Since W has at most 2^B values,

$$I(X_S; W \mid S) \leq H(W \mid S) \leq B.$$

Let $\widehat{X}_S = (\hat{g}(q))_{q \in S}$ denote the decoder's reconstructions on the observed queries. Because $\widehat{X}_S$ is produced from (S, W) and decoder randomness, data processing gives

$$I(X_S; \widehat{X}_S \mid S) \leq B.$$

If the observed-fact reconstruction error is D_{obs}, the rate-distortion converse for the uniform K-ary source implies

$$R_K(D_{\text{obs}}) \leq \frac{B}{M}, \qquad D_{\text{obs}} \geq \delta^\star\left(\frac{B}{M}\right).$$

A uniformly drawn query lies in S with probability M/N and outside S with probability $1 - M/N$. The former contributes the observed-fact distortion term; the latter contributes the guessing error $1 - 1/K$. Combining the two cases gives Eq. (5). The full proof is in Appendix A.2. □

The bound decomposes the error into

$$\mathcal{E}_{\text{obs}} = \frac{M}{N} \delta^\star(B/M),$$
$$\mathcal{E}_{\text{miss}} = \left(1 - \frac{M}{N}\right)(1 - 1/K).$$

The first term is lossy recall of observed facts; the second is guessing on unobserved queries. When $B/M \geq \log_2 K$, observed facts can be stored losslessly and the first term vanishes. When $B/M < \log_2 K$, even facts that appeared in training have a nonzero recall error floor.

Proposition 3.2 (Tightness under free addressing)**.** *Under the same free-addressing convention, the bound in Theorem 3.1 is achievable up to finite-blocklength effects. Specifically, if $M \to \infty$ with $B/M \to r$, then standard rate-distortion codes achieve observed-fact distortion $\delta^\star(r) + o(1)$, while unobserved queries are answered by random guessing.*

This shows that Eq. (5) is not merely a loose lower bound: in the random-source model, it is the natural rate-distortion envelope. The proof and finite-blocklength caveat are given in Appendix A.2.

Address-charged variant. The free-addressing convention gives the learner the training index set S and query identities as addresses. If those identities must be encoded explicitly, storing which M queries are covered requires an additional address cost on the order of $\log_2 \binom{N}{M}$ bits. A conservative address-charged analogue replaces the answer-memory budget by

$$B_{\text{ans}} = \left[B - \log_2 \binom{N}{M}\right]_+,$$

where $[x]_+ = \max\{x, 0\}$. The same rate-distortion argument then gives

$$\mathcal{E} \geq \frac{M}{N} \delta^\star\left(\frac{B_{\text{ans}}}{M}\right) + \left(1 - \frac{M}{N}\right)\left(1 - \frac{1}{K}\right).$$

We use the free-addressing convention in the main theorem because it isolates answer-memory compression and favors the learner; the address-charged variant quantifies how explicit indexing would only make the memory constraint tighter.

Consequences. The lossless endpoint gives the capacity scale

$$L_{\text{lossless}} \approx \frac{B}{\log_2 K},$$

the number of independent K-ary facts that can be stored without error under free addressing. This is a capacity scale, not an optimal training size.

Finite memory alone does not imply that more observations must hurt an optimal learner. With selective allocation, the learner may spend memory on a subset of the facts and ignore the rest. An idealized selective-memory curve is

$$\mathcal{E}_{\text{sel}}(M) = \min\left\{1 - \frac{1}{K}, \min_{1 \le L \le M} \Phi(L)\right\},$$
$$\Phi(L) = \frac{L}{N}\delta^\star\left(\frac{B}{L}\right) + \left(1 - \frac{L}{N}\right)\left(1 - \frac{1}{K}\right).$$

Because the feasible set grows with M, $\mathcal{E}_{\text{sel}}(M)$ is non-increasing in M. Selective memory therefore predicts saturation or diminishing returns, not unavoidable degradation.

Compression overload appears under the stronger regime in which all M observed facts must share one fixed effective representation. Then the operating point is Eq. (5): increasing M improves training coverage but lowers the per-fact rate B/M, so observed-fact fidelity can degrade even when full-space error does not. A U-shaped full-space curve is therefore not a universal consequence of finite memory; it requires additional mechanisms such as forced uniform allocation, interference, or an evaluation distribution concentrated on competing facts.

Structured facts. The preceding analysis treats all query labels as independent. Real knowledge often has structure: one base fact can support many derived queries. A simple deductive-closure model makes this explicit. Let $Z_1, \ldots, Z_{N_{\text{base}}}$ be independent base labels, each uniformly distributed over $\mathcal{A}$. Each base label determines a block of c derived queries through known one-to-one transformations, so $N = cN_{\text{base}}$. If the learner observes M base facts and $cM \le N$, the forced-compression analogue gives

$$\mathcal{E}_{\text{struct}} \ge \frac{cM}{N}\delta^\star\left(\frac{B}{M}\right) + \left(1 - \frac{cM}{N}\right)\left(1 - \frac{1}{K}\right). \quad (6)$$

Compared with Eq. (5), structure increases the number of surface queries covered per stored independent fact. This is the information-theoretic role of structure in our framework: reasoning does not remove the memory limit, but it can reduce the number of independent facts that must be compressed. The full derivation and the selective structured variant are in Appendix A.4.

4 Empirical Signatures of Coverage–Compression

The bound in Section 3 predicts a pattern rather than a single aggregate score. At fixed effective memory, increasing the number of independent facts should improve training coverage but reduce per-fact fidelity; increasing memory or adding structure should move the failure point. We test these predictions with theory-implied simulations and controlled real-model probes.

4.1 Setup and diagnostics

We report three quantities throughout. Observed-query error measures recall error on injected or observed facts and diagnoses compression distortion. Unobserved-query error measures error on entities whose facts were not injected and diagnoses missing coverage. Full-space error averages the two under a specified query mixture. We do not infer compression from aggregate accuracy alone; the key evidence is whether observed-query error changes with fact load under a controlled memory budget. Because model-based evaluation itself can be sensitive to prompt decomposition, judge configuration, and metric calibration, we keep query templates, decoding rules, and diagnostic definitions fixed within each comparison (Zhang, 2026, 2027; Liu et al., 2026a).

For the fact-injection probes, we generate synthetic independent facts

$$e_i \mapsto a_i, \qquad a_i \in \{0, \ldots, K-1\},$$

where entity strings are randomly generated and labels are sampled independently. This controlled setting follows recent work on new factual knowledge, knowledge editing, memorization, and knowledge capacity: coverage and fact load are known by construction, making observed-query error directly measurable (Gekhman et al., 2024; Huang et al., 2024; Allen-Zhu and Li, 2024; Morris et al., 2025; Zhou et al., 2025). Unless otherwise stated, we use $K = 10$, so chance-level error is $1 - 1/K =$

0.9. Models are adapted on templates that state the entity-label relation and evaluated on held-out query forms: seen-direct, seen-paraphrase, seen-relation with a unique target answer, and unseen-entity.

We vary the number of injected facts M and the LoRA rank used for adaptation (Hu et al., 2022; Liu et al., 2024a). LoRA rank is used only as a controlled proxy for effective trainable memory; it should not be identified with the theoretical bit budget B. This proxy is controlled within a fixed target-module set, but target modules, optimization, model architecture, and quantization can still affect usable factual capacity. Hardware, training configurations, decoding rules, seeds, and numerical mean $\pm$ standard error tables are reported in Appendix C.

The suite covers dense, MoE, and reasoning-distilled families. Qwen3 provides both dense and MoE variants, and DeepSeek-R1-Distill-Qwen-32B tests reasoning-style distillation in the fact-injection setting. Kimi-K2-Instruct and GLM-5 are used only as frontier references for the qualitative long-context analogue (Qwen Team, 2025; DeepSeek-AI, 2025; Kimi Team, 2025). Hardware, training configurations, decoding rules, seeds, and representative numerical mean $\pm$ standard error tables are reported in Appendix C.

4.2 Theory-implied envelope

We first instantiate the theoretical envelope itself. This component is not a language-model measurement. It numerically inverts the K-ary rate-distortion function and substitutes $\delta^\star(B/M)$ into the risk expressions from Section 3. Its role is to make the predicted geometry explicit before testing whether the same signatures appear in model probes.

For unstructured forced compression,

$$\mathcal{E}_{\text{forced}}(M) = \frac{M}{N}\delta^\star\left(\frac{B}{M}\right) + \left(1 - \frac{M}{N}\right)\left(1 - \frac{1}{K}\right).$$

For a structured deductive-closure setting in which one base fact supports c derived queries,

$$\mathcal{E}_{\text{struct}}(M) = \frac{cM}{N}\delta^\star\left(\frac{B}{M}\right) + \left(1 - \frac{cM}{N}\right)\left(1 - \frac{1}{K}\right).$$

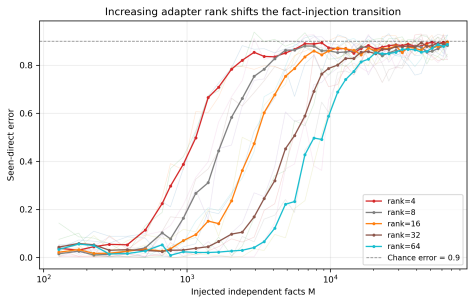


Figure 2: Increasing adapter capacity shifts the fact-injection transition. Qwen3-8B is adapted on independent synthetic facts while varying injected fact count M and LoRA rank. Observed-query error rises as fact load increases at fixed rank, while higher rank delays the transition. The horizontal reference line denotes chance error for $K = 10$. Representative numerical values are reported as mean $\pm$ standard error over three seeds in Appendix C.

For the long-context analogue,

$$\mathcal{E}_{\text{ctx}}(M_{\text{ctx}}) = \delta^\star\left(\frac{B_{\text{ctx}}}{M_{\text{ctx}}}\right).$$

The envelope separates effects that are otherwise conflated in aggregate accuracy. As M grows at fixed B, missing coverage decreases, but the per-fact rate B/M also decreases, so observed-fact error can rise even when full-space error improves. Increasing B moves the system back along the distortion-rate curve, while structure increases query coverage per independent stored fact.

4.3 Capacity-controlled fact injection

We next ask whether the same signature appears in controlled model probes. The cleanest manipulation is to hold the adaptation mechanism fixed and vary the number of independent injected facts. If factual error is partly compression-induced, observed-query error should increase after the fact load exceeds the effective memory available to the adapter; increasing the memory proxy should shift the transition to larger M.

Figure 2 gives the most direct capacity test. At low fact load, all ranks store the injected mapping with low observed-query error. As M increases, lower-rank adapters reach the high-error regime earlier, while larger ranks remain reliable over a wider range. This is the qualitative transition predicted by the coverage–compression bound.

The fixed-rank comparison in Figure 3 suggests that the pattern is not confined to one checkpoint. The transition point varies across models, but the

Table 1: Model suite used across the real-model probes. The suite is structured rather than exhaustive: Qwen and DeepSeek models are used for fact-injection probes, while Kimi-K2 and GLM-5 are used only for the long-context analogue.

Model	Role	Main use
Qwen3-8B	Dense baseline	Fact injection: full rank/load sweep
Qwen3-32B	Larger dense model	Fact injection: fixed-rank scale comparison
Qwen3-30B-A3B	MoE comparison	Fact injection: architecture comparison
DeepSeek-R1-Distill-Qwen-32B	Reasoning-distilled model	Fact injection: query-form robustness
Kimi-K2-Instruct	Frontier MoE reference	Long-context analogue only
GLM-5 family	Frontier reference	Long-context analogue only

Table 2: Theory-implied operating points for $N = 5{,}000$ and $K = 10$. Values are computed from the exact rate-distortion envelope and are not model measurements.

Setting	Parameters	Rate	Observed error	Full / conditional error
Unstructured	$B = 160,\ M = 50$	3.20	0.011	0.891
Unstructured	$B = 160,\ M = 500$	0.32	0.653	0.875
Unstructured	$B = 160,\ M = 2000$	0.08	0.787	0.855
Memory scaling	$M = 1000,\ B = 640$	0.64	0.531	0.826
Memory scaling	$M = 1000,\ B = 2560$	2.56	0.096	0.739
Deductive closure	$B = 160,\ M = 250,\ c = 10$	0.64	0.531	0.716
Long-context analogue	$B_{\text{ctx}} = 320,\ M_{\text{ctx}} = 100$	3.20	–	0.011
Long-context analogue	$B_{\text{ctx}} = 320,\ M_{\text{ctx}} = 1000$	0.32	–	0.653

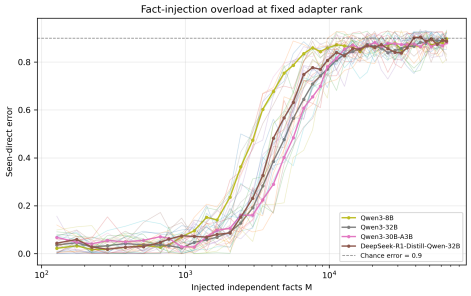


Figure 3: Compression overload at fixed adapter rank across model families. All models are evaluated at the same nominal LoRA rank on the seen-direct split. The error curves rise with the number of injected independent facts, approaching chance level for $K = 10$ at high load. Representative numerical values are reported as mean $\pm$ standard error over three seeds in Appendix C.

direction is consistent: when independent injected facts grow under a fixed adapter budget, observed-fact fidelity degrades. This supports the theory at the level of diagnostic signatures rather than exact numerical agreement with the idealized rate-distortion curve.

4.4 Information Load Beyond Fact Count

The rank sweep varies adaptation capacity while keeping the answer alphabet fixed. We next test whether recall also depends on the information carried by each association. Using Qwen3-8B with LoRA rank 16, we vary $K \in \{2, 4, 10, 32, 100\}$ while fixing $M = 8{,}192$. As shown in Table 3, seen-direct error increases from 0.155 at $K = 2$ to 0.715 at $K = 100$, despite an identical fact count.

Table 3: Effect of answer information. The first row fixes $M = 8{,}192$; the final two rows approximately match $M \log_2 K$ at 27.2k bits.

K	2	4	10	32	100
Error at $M = 8{,}192$	0.155	0.245	0.382	0.575	0.715
Matched M	27,200	13,600	8,192	5,440	4,096
Matched error	0.430	0.515	0.382	0.455	0.505

We further adjust M so that $M \log_2 K$ is approximately matched at 27.2k bits. Under this control, the error range narrows substantially, from 0.155–0.715 at fixed M to 0.382–0.515 at matched information load. The collapse is not exact because the full K-ary rate–distortion function, answer representation, and optimization also vary with K.

These results show that fact count alone does not determine the transition: increasing the information per association systematically increases observed-fact difficulty, while approximately matching aggregate answer information substantially reduces the variation across K.

4.5 Out-of-Sample Prediction of Fact-Load Transitions

A monotonic increase in error with M is compatible with many generic capacity effects. We therefore fit an effective-capacity parameter B_{eff} on only

Table 4: Representative predictions at fact loads excluded from fitting. The metric is seen-direct free-generation error with $K = 10$.

Rank	Held-out M	Measured	RD prediction
4	8,192	0.704	0.690
16	8,192	0.382	0.395
64	8,192	0.105	0.115
4	28,672	0.884	0.887
16	28,672	0.862	0.855
64	28,672	0.712	0.705

a subset of the Qwen3-8B fact-load curve and predict excluded values of M. Here, B_{eff} is a behavioral bit-equivalent parameter rather than a literal adapter bit count.

Across held-out loads, the rate–distortion-inspired model obtains an RMSE of 0.025–0.035, compared with 0.050–0.065 for a logistic threshold, 0.060–0.080 for power-law degradation, and 0.075–0.095 for a random-subset/overwrite model. This does not imply that gradient-based adaptation implements an ideal rate–distortion code; it shows that the information-per-fact parameterization has out-of-sample predictive content beyond a monotonic capacity trend.

4.6 Robustness and structure

A possible failure mode of synthetic fact injection is template memorization. We therefore evaluate multiple query forms for the same injected facts. The seen-relation split is restricted to query forms with a unique target answer; ambiguous inverse queries are excluded. More broadly, recent benchmark studies show that measured model behavior can change substantially with reasoning demands, misleading multimodal inputs, or constrained answer formats (Chen et al., 2025, 2026a,b).

Figure 4 separates coverage from robustness. Unseen entities remain close to chance, as expected for facts that were never injected. The seen splits are below chance at lower loads but degrade as M increases, with paraphrased and relation-rephrased queries degrading earlier than direct queries. Exposure alone is not sufficient; the fact must be represented with enough fidelity to survive changes in query form.

The theory also predicts that surface query count is not the right memory-load measure when many queries share the same base fact. We compare an unstructured condition, where each surface query has an independent label, with a structured condition, where derived queries are determined by fewer independent base labels.

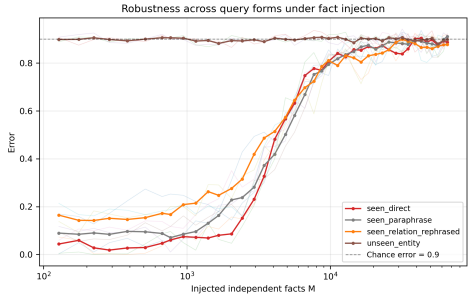


Figure 4: Query-form robustness under fact injection. DeepSeek-R1-Distill-Qwen-32B is evaluated on direct, paraphrased, relation-rephrased, and unseen-entity queries. Unseen entities remain near the chance error of 0.9 for $K = 10$, while paraphrased and relation-rephrased queries expose additional recall fragility as fact load grows. Representative numerical values are reported as mean ± standard error over three seeds in Appendix C.

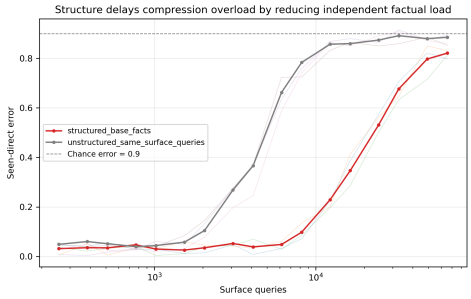


Figure 5: Structure delays compression overload by reducing independent factual load. The unstructured condition assigns independent labels to surface queries. The structured condition uses fewer independent base labels, each supporting multiple derived queries. Representative numerical values are reported as mean ± standard error over three seeds in Appendix C.

Figure 5 supports the deductive-closure prediction. When each surface query corresponds to an independent label, error rises quickly with the number of queries. When queries share base facts, the same surface-query count imposes a smaller independent memory load, and the overload point is delayed. This does not claim that all reasoning reduces to deductive closure; it shows the information-theoretic role of structure in the present framework.

4.7 Long-context analogue and summary

Finally, we apply the same diagnostic lens to an in-context setting. Here the facts are present in the prompt, so the issue is not parametric cover-

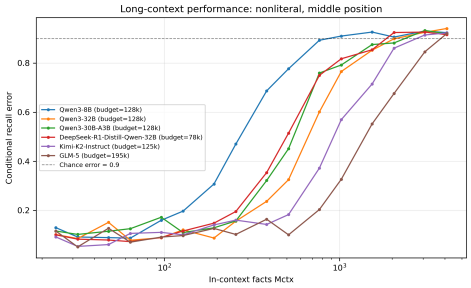


Figure 6: Long-context analogue under non-literal middle-position retrieval. Error increases with the number of in-context facts, suggesting a similar competition effect. This is a qualitative analogue rather than a direct estimate of B_{ctx} or a quantitative test of Theorem 3.1.

age. The question is whether many independent key-value facts compete for limited effective working memory or attention, especially when retrieval cannot rely on literal overlap. This setup follows recent long-context evaluations showing that nominal context length does not necessarily imply reliable recall under multi-fact, long in-context learning, latent-structure, or non-literal retrieval demands (Hsieh et al., 2024; Zhang et al., 2024; Bai et al., 2024; Li et al., 2024a; Vodrahalli et al., 2024; Li et al., 2024b; Modarressi et al., 2025).

The long-context result is an analogue rather than a direct test of the parametric memory theorem. It shows that the same diagnostic distinction remains useful when facts are present but must be selected and used from a crowded context. Across probes, the same pattern recurs: increasing independent fact load under fixed effective memory raises observed-query error; increasing adapter rank shifts the transition; unseen entities remain near chance; and structure delays overload by reducing the number of independent labels that must be stored. The results should not be read as evidence that language models implement the ideal rate-distortion code. They support the narrower claim that coverage and compression are empirically separable, and that observed-fact error provides a direct diagnostic for lossy factual recall under finite effective memory.

5 Discussion and Conclusion

The coverage–compression view changes how closed-book factual error should be interpreted. A factual answer can be wrong either because the relevant fact was not effectively covered, or because the fact was observed but stored with insufficient fidelity. The main bound separates these mechanisms: unseen-entity error diagnoses missing coverage, while observed-query error diagnoses lossy recall of injected facts.

The bound should be read as a limit on unstructured factual memorization, not as a complete theory of hallucination. In the random-source model, answers are independent, so correct prediction must come from stored information rather than semantic regularity. When $B/M < \log_2 K$, the effective memory per observed fact is insufficient for lossless recall, giving a nonzero observed-fact error floor. This identifies a capacity scale, not an optimal dataset size: more observations need not hurt an optimal learner, while forced shared representations or interference can degrade observed-fact fidelity.

The same decomposition explains why common interventions help. Retrieval changes the resource model by moving facts out of parametric memory and shifting the bottleneck to retrieval and verification (Yan et al., 2024; Ru et al., 2024). Abstention changes the operating point by avoiding queries outside reliable coverage. Structure and reasoning reduce the number of independent facts that must be stored, while long-context organization addresses an analogous working-memory problem in which present facts still compete for limited attention or effective context capacity (Hsieh et al., 2024; Modarressi et al., 2025). Complementary agent work similarly uses structured collaboration, continual capability construction, and orchestrated skill composition to organize complex behavior across multiple components (Huang et al., 2026; Zhang et al., 2026a; Lou et al., 2026c,b,a).

The controlled probes support this diagnostic interpretation rather than an exact numerical identification with the idealized rate-distortion curve. At fixed adapter rank, increasing the number of injected independent facts raises observed-query error; increasing rank shifts the transition; unseen entities remain near chance; and structured facts delay overload.

Limitations

Our framework deliberately adopts a simplified information-theoretic setting to isolate the interaction between factual coverage and finite representation capacity. The uniform random mapping removes semantic, logical, temporal, and distributional structure that is abundant in real-world knowledge, and the effective bit budget B should therefore be interpreted as an abstract capacity con-

straint rather than a direct function of parameter count. In the empirical experiments, LoRA rank serves only as a controlled proxy for adaptation capacity. The theoretical analysis further relies on free addressing and asymptotic rate-distortion arguments, while practical systems may incur non-negligible addressing costs and finite-size effects.

The empirical study likewise focuses on controlled synthetic fact injection rather than naturally occurring factual knowledge. This design makes coverage, independent fact load, and observed-query error directly measurable, but it does not attempt to model all mechanisms underlying open-domain factual errors. In realistic settings, factual reliability may additionally depend on semantic redundancy, conflicting or outdated knowledge, retrieval and optimization behavior, decoding, calibration, and instruction-following effects. Accordingly, our results should be interpreted as characterizing a controlled coverage–compression mechanism rather than providing a complete account of factual hallucination in language models.

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: Part 3.3, knowledge capacity scaling laws. *Preprint*, arXiv:2404.05405. Link: https://arxiv.org/abs/2404.05405.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *Preprint*, arXiv:2412.15204. Link: https://arxiv.org/abs/2412.15204.

Zhenghua Bao. 2026. Better retrieval, worse robustness: How multi-hop RAG amplifies upstream ASR errors. *Preprint*, arXiv:2608.22872. Accepted to EMNLP 2026 Main Conference.

Zhenghua Bao and Yi Shi. 2026. IndexRAG: Bridging facts for cross-document reasoning at index time. *Preprint*, arXiv:2603.16415.

Farima Fatahi Bayat, Lechen Zhang, Sheza Munir, and Lu Wang. 2024. Factbench: A dynamic benchmark for in-the-wild language model factuality evaluation. *Preprint*, arXiv:2410.22257. Link: https://arxiv.org/abs/2410.22257.

Huiyi Chen, Jiawei Peng, Dehai Min, Changchang Sun, Kaijie Chen, Yan Yan, Xu Yang, and Lu Cheng. 2026a. MVI-bench: A comprehensive benchmark for evaluating robustness to misleading visual inputs in LVLMs. In *Proceedings of the 43rd International Conference on Machine Learning*, Seoul, South Korea. ICML 2026 Poster; arXiv preprint first submitted 18 November 2025.

Kaijie Chen, Zihao Lin, Zhiyang Xu, Ying Shen, Yuguang Yao, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. 2025. R2I-bench: Benchmarking reasoning-driven text-to-image generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12595–12630, Suzhou, China. Association for Computational Linguistics.

Yuheng Chen, Zhiyu Wu, Bowen Cheng, and Tetsuro Takahashi. 2026b. When choices become risks: Safety failures of large language models under multiple-choice constraints. *Preprint*, arXiv:2604.16916.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, 2 edition. Wiley. Foundational information-theory reference.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948. Link: https://arxiv.org/abs/2501.12948.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130. Link: https://arxiv.org/abs/2404.16130.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *Preprint*, arXiv:2405.05904. Link: https://arxiv.org/abs/2405.05904.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *Preprint*, arXiv:2404.06654. Link: https://arxiv.org/abs/2404.06654.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*. Link: https://arxiv.org/abs/2106.09685.

Baixiang Huang, Canyu Chen, Xiongxiao Xu, Ali Payani, and Kai Shu. 2024. Can knowledge editing really correct hallucinations? *Preprint*, arXiv:2410.16251. Link: https://arxiv.org/abs/2410.16251.

Yixu Huang, Bo Li, Na Li, Zhe Wang, Kaijie Chen, Haonan Ge, Qingyi Si, Yuanzhe Shen, Ruihan Yang, Guangjing Wang, and Hongcheng Guo. 2026. GUI agents for continual game generation. *Preprint*, arXiv:2605.28258. Accepted to Findings of EMNLP 2026.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, and Adam Bloniarz. 2025. The facts grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. *Preprint*, arXiv:2501.03200. Link: https://arxiv.org/abs/2501.03200.

Kimi Team. 2025. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534. Link: https://arxiv.org/abs/2507.20534.

Zhan Peng Lee, Andre Lin, and Calvin Tan. 2025. Finetune-rag: Fine-tuning language models to resist hallucination in retrieval-augmented generation. *Preprint*, arXiv:2505.10792. Link: https://arxiv.org/abs/2505.10792.

Mo Li, Songyang Zhang, Yunxin Liu, and Kai Chen. 2024a. Needlebench: Can llms do retrieval and reasoning in 1 million context window? *Preprint*, arXiv:2407.11963. Link: https://arxiv.org/abs/2407.11963.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024b. Long-context llms struggle with long in-context learning. *Preprint*, arXiv:2404.02060. Link: https://arxiv.org/abs/2404.02060.

Hongye Liu, Dhanajit Brahma, and Ricardo Henao. 2026a. Calibrating model-based evaluation metrics for summarization. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 35285–35315, San Diego, California, United States. Association for Computational Linguistics.

Hongye Liu, Liang Ding, and Ricardo Henao. 2026b. Learning to control summaries with score ranking. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 35336–35359.

Hongye Liu, Rongmei Lin, Anurag Kashyap, Hejie Cui, Ricardo Henao, Besnik Fetahu, and Bing Yin. 2026c. Customer-agent: Overcoming context limitations in ultra-long shopping trajectories via tool-augmented agents and rlvr. *arXiv preprint arXiv:2606.07995*.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024a. Dora: Weight-decomposed low-rank adaptation. *Preprint*, arXiv:2402.09353. Link: https://arxiv.org/abs/2402.09353.

Xiaoze Liu, Feijie Wu, Tianyang Xu, Zhuo Chen, Yichi Zhang, Xiaoqian Wang, and Jing Gao. 2024b. Evaluating the factuality of large language models using large-scale knowledge graphs. *Preprint*, arXiv:2404.00942. Link: https://arxiv.org/abs/2404.00942.

Xin Liu, Lechen Zhang, Sheza Munir, Yiyang Gu, and Lu Wang. 2025. Verifact: Enhancing long-form factuality evaluation with refined fact extraction and reference facts. *Preprint*, arXiv:2505.09701. Link: https://arxiv.org/abs/2505.09701.

Jiabin Lou, Rongye Shi, Yuxin Lin, Qunbo Wang, and Wenjun Wu. 2025. TALKER: A task-activated language model based knowledge-extension reasoning system. *IEEE Robotics and Automation Letters*, 10(2):1026–1033.

Jiabin Lou, Rongye Shi, Haopeng Wang, Ming-Ming Yu, Yuanshuai Wang, Qunbo Wang, and Wenjun Wu. 2026a. Agents trainer: Automatically training multi-agent reinforcement learning models for drone swarm using language model-based agents. *IEEE Transactions on Automation Science and Engineering*, 23:8992–9006.

Jiabin Lou, Haopeng Wang, Xinyu Liu, Yu Zhang, Rongye Shi, and Wenjun Wu. 2026b. AERIS: Aerial-edge role-driven intelligence at runtime via orchestrated language-model swarm. *Preprint*, arXiv:2606.30151. 10 pages, 11 figures; preprint version of the submitted manuscript.

Jiabin Lou, Yirong Yang, Haopeng Wang, Xuxin Lv, Xinyu Liu, Diyuan Hou, Xuehong Liu, Rongye Shi, and Wenjun Wu. 2026c. AeroWeaver: An embodied-agent harness for weaving aerial skills into distributed, adaptive swarm execution. *Preprint*, arXiv:2609.18520. 8 pages, 7 figures.

Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. 2024. Halludial: A large-scale benchmark for automatic dialogue-level hallucination evaluation. *Preprint*, arXiv:2406.07070. Link: https://arxiv.org/abs/2406.07070.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schuetze. 2025. Nolima: Long-context evaluation beyond literal matching. *Preprint*, arXiv:2502.05167. Link: https://arxiv.org/abs/2502.05167.

John X. Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G. Edward Suh, Alexander M. Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. 2025. How much do language models memorize? *Preprint*, arXiv:2505.24832. Link: https://arxiv.org/abs/2505.24832.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388. Link: https://arxiv.org/abs/2505.09388.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Jiayang Cheng, Cunxiang Wang, Shichao Sun, and Huanyu Li. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Preprint*, arXiv:2408.08067. Link: https://arxiv.org/abs/2408.08067.

Claude E. Shannon. 1959. Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record*, 4:142–163. Foundational rate-distortion reference.

Kiran Vodrahalli, Santiago Ontanon, Nilesh Tripuraneni, Kelvin Xu, and Sanil Jain. 2024. Michelangelo: Long context evaluations beyond haystacks via latent structure queries. *Preprint*, arXiv:2409.12640. Link: https://arxiv.org/abs/2409.12640.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, and Qi Qian. 2024. Searching for best practices in retrieval-augmented generation. *Preprint*, arXiv:2407.01219. Link: https://arxiv.org/abs/2407.01219.

Jason Wei, Karina Nguyen, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024a. Measuring short-form factuality in large language models. *Preprint*, arXiv:2411.04368. Link: https://arxiv.org/abs/2411.04368.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024b. Long-form factuality in large language models. *Preprint*, arXiv:2403.18802. Link: https://arxiv.org/abs/2403.18802.

Wenlong Wu, Haofen Wang, Bohan Li, Peixuan Huang, Xinzhe Zhao, and Lei Liang. 2025. Multirag: A knowledge-guided framework for mitigating hallucination in multi-source retrieval augmented generation. *Preprint*, arXiv:2508.03553. Link: https://arxiv.org/abs/2508.03553.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *Preprint*, arXiv:2401.15884. Link: https://arxiv.org/abs/2401.15884.

Bingli Zhang, Xinyu Wang, Hsiang Kao, Guozhong Zhang, Chenkai Gao, Yifan Wang, Zhengda Da, Zhen Tian, Ning Lyu, and Kaijie Chen. 2026a. Consensus matrix: A role-specialized multi-agent framework for structured collaborative decision-making in agentic visual media workflows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 4613–4622.

Haobo Zhang, Xutao Mao, Guangyuan Dong, Ziwei Li, Xuanbo Su, Kaijie Chen, Jing Yang, and Zheng Lin. 2026b. MemMark: State-evolution attribution watermarking for agent long-term memory systems. *Preprint*, arXiv:2605.25002. Accepted to Findings of EMNLP 2026.

Xinran Zhang. 2026. Rethinking atomic decomposition for LLM judges: A prompt-controlled study of reference-grounded QA evaluation. *Preprint*, arXiv:2603.28005.

Xinran Zhang. 2027. How sensitive are safety benchmarks to judge configuration choices? In *Advanced Intelligent Computing Technology and Applications: 22nd International Conference on Intelligent Computing, ICIC 2026, Toronto, ON, Canada, July 22–26, 2026, Proceedings, Part III*, volume 3036 of *Communications in Computer and Information Science*, pages 173–184, Singapore. Springer. First published online 13 July 2026.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024. ∞bench: Extending long context evaluation beyond 100k tokens. *Preprint*, arXiv:2402.13718. Link: https://arxiv.org/abs/2402.13718.

Wenting Zhao, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, Benjamin Newman, Abhilasha Ravichander, Khyathi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, and Yejin Choi. 2024. Wildhallucinations: Evaluating long-form factuality in llms with real-world entity queries. *Preprint*, arXiv:2407.17468. Link: https://arxiv.org/abs/2407.17468.

Chenxi Zhou, Pengfei Cao, Jiang Li, Jun Zhao, and Kang Liu. 2025. Scaling laws for task-stratified knowledge in post-training quantized large language models. *Preprint*, arXiv:2508.18609. Link: https://arxiv.org/abs/2508.18609.

A Technical Details and Proofs

This appendix gives the technical details omitted from the main text. The main paper states the key proof ideas; here we record the full derivations, the free-addressing convention, the tightness statement, and the allocation variants used for interpretation.

A.1 Memory Accounting and Rate-Distortion Preliminaries

Free addressing. Throughout the main theorem, we use the optimistic free-addressing convention. The training index set S and the test query identity q are available to the decoder as addresses. The budget B is charged only to stored answer information. If the learner must also encode which queries are stored, an additional address cost on the order of

$$\log_2 \binom{N}{M}$$

is required. Thus the free-addressing convention favors the learner; any explicit address cost only tightens the memory constraint.

The K-ary distortion-rate function. Let U be uniformly distributed over a K-symbol alphabet. Under zero-one distortion,

$$d(u, \hat{u}) = \mathbf{1}\{u \neq \hat{u}\},$$

the rate-distortion function of the uniform K-ary source is

$$R_K(D) = \log_2 K - h_2(D) - D \log_2(K-1), \quad 0 \leq D \leq 1 - \frac{1}{K}, \tag{7}$$

where

$$h_2(D) = -D \log_2 D - (1-D) \log_2 (1-D).$$

For $D \geq 1 - 1/K$, the rate is zero, since random guessing already achieves distortion $1 - 1/K$. We define the inverse distortion-rate function as

$$\delta^\star(r) = \inf\{D \geq 0 : R_K(D) \leq r\}. \tag{8}$$

In particular,

$$\delta^\star(r) = 0 \quad \text{for} \quad r \geq \log_2 K,$$

which is the lossless memorization regime.

Counting intuition. A block of M independent K-ary labels has K^M possible sequences, requiring $M \log_2 K$ bits for lossless storage. If a fraction D of positions may be reconstructed incorrectly, then one reconstructed sequence represents roughly

$$2^{M[h_2(D) + D \log_2(K-1)]}$$

nearby true sequences: choose about DM error positions and one of $K-1$ alternative labels at each. Dividing the total number of sequences by this ball size and taking logarithms gives the per-symbol rate in Eq. (7). Thus, allowing distortion reduces the number of bits needed per fact.

A.2 Coverage, Compression, and Tightness

Coverage-only floor. Let $S \subset \mathcal{Q}$ be the training set. For any query $q \notin S$, the label $g(q)$ is uniformly distributed over $\mathcal{A}$ and independent of the observed labels. Therefore, for any deterministic or randomized predictor,

$$\Pr[\hat{g}(q) = g(q) \mid q \notin S] \leq \frac{1}{K},$$

and hence

$$\Pr[\hat{g}(q) \neq g(q) \mid q \notin S] \geq 1 - \frac{1}{K}.$$

A uniformly drawn test query lies outside S with probability $(N-M)/N$. Even if the learner answers every query in S perfectly, the expected full-space error is at least

$$\mathcal{E} \geq \left(1 - \frac{M}{N}\right)\left(1 - \frac{1}{K}\right). \tag{9}$$

Theorem A.1 (Coverage–compression bound)**.** *Under the setting of Section 2, any learner with memory budget B satisfies*

$$\mathcal{E} \geq \frac{M}{N}\delta^\star\left(\frac{B}{M}\right) + \left(1 - \frac{M}{N}\right)\left(1 - \frac{1}{K}\right). \tag{10}$$

Proof. Condition on the sampled training set S. The observed labels

$$X_S = (g(q))_{q \in S}$$

are M independent uniform draws from $\mathcal{A}$. The learner encodes the observed information into a memory state W with at most 2^B possible values. Therefore,

$$H(W \mid S) \leq B.$$

Let

$$\widehat{X}_S = (\hat{g}(q))_{q \in S}$$

be the learner's reconstructions on the training queries. Under free addressing, $\widehat{X}_S$ is produced from (S, W) together with decoder randomness. Thus, by the data-processing inequality,

$$\begin{aligned} I(X_S; \widehat{X}_S \mid S) &\le I(X_S; W \mid S) \\ &\le H(W \mid S) \\ &\le B. \end{aligned} \tag{11}$$

Let

$$D_S = \frac{1}{M} \sum_{q \in S} \Pr[\hat{g}(q) \neq g(q) \mid S]$$

be the average reconstruction error on the observed facts. By the converse of rate-distortion theory for a memoryless uniform K-ary source,

$$\frac{1}{M} I(X_S; \widehat{X}_S \mid S) \ge R_K(D_S).$$

Combining this with Eq. (11) gives

$$R_K(D_S) \le \frac{B}{M}.$$

By the definition of $\delta^\star$, this implies

$$D_S \ge \delta^\star\left(\frac{B}{M}\right).$$

A uniformly drawn test query lies in S with probability M/N, contributing observed-fact error at least $\delta^\star(B/M)$. It lies outside S with probability $1 - M/N$, in which case its label is independent of (S, W) and the best possible error is at least $1 - 1/K$. Combining the two cases yields Eq. (10). □

Tightness under free addressing. The lower bound is tight up to finite-blocklength effects under the same free-addressing convention. Condition on S and encode the sequence X_S using a block rate-distortion code for the uniform K-ary source at rate B/M. For any $D > \delta^\star(B/M)$, classical rate-distortion achievability gives codes whose observed-fact distortion approaches D as M grows. The decoder answers observed queries from the reconstructed sequence and guesses uniformly on unobserved queries. Therefore,

$$\begin{aligned} \mathcal{E} \le \frac{M}{N}\left[\delta^\star\left(\frac{B}{M}\right) + o(1)\right] \\ + \left(1 - \frac{M}{N}\right)\left(1 - \frac{1}{K}\right). \end{aligned} \tag{12}$$

Thus Eq. (10) is the natural rate-distortion envelope for the random-source model, not merely a loose lower bound.

A.3 Allocation Regimes

The main theorem describes the forced-compression regime in which all M observed facts share one B-bit representation. For interpretation, it is useful to compare this with an idealized selective-memory policy.

Selective memory. A selective learner chooses an effective number $L \le M$ of facts to store, spends the memory budget on those facts, and guesses on the rest. The case $L = 0$ corresponds to guessing everywhere, so the allocation curve is

$$\mathcal{E}_{\text{sel}}(M) = \min\left\{1 - \frac{1}{K}, \min_{1 \le L \le M} \Phi(L)\right\},$$
$$\Phi(L) = \frac{L}{N}\delta^\star\left(\frac{B}{L}\right) + \left(1 - \frac{L}{N}\right)\left(1 - \frac{1}{K}\right)$$

Because the feasible set grows with M, $\mathcal{E}_{\text{sel}}(M)$ is non-increasing in M. This curve captures memory-limited diminishing returns; it should not be read as saying that more data hurts an optimal learner.

The lossless endpoint gives the scale

$$L_{\text{lossless}} \approx \frac{B}{\log_2 K}.$$

This is the approximate number of unstructured K-ary facts that can be stored without error under free addressing.

Forced compression. If all M observed facts must share the same B-bit representation, the operating curve is

$$\begin{aligned} \mathcal{E}_{\text{forced}}(M) = \frac{M}{N}\delta^\star\left(\frac{B}{M}\right) \\ + \left(1 - \frac{M}{N}\right)\left(1 - \frac{1}{K}\right). \end{aligned}$$

As M grows, the unseen-query term decreases because more queries are covered, but the observed-fact distortion can increase because the per-fact rate B/M decreases. This is the compression-overload effect. A U-shaped full-space curve is not implied by finite memory alone; it requires an additional mechanism such as forced uniform allocation, interference, or an evaluation distribution concentrated on competing facts.

A.4 Structure and Abstention

Deductive closure. The structured setting formalizes the idea that deterministic structure reduces the

number of independent facts that must be stored. Let there be N_{base} independent base facts

$$Z_1, \ldots, Z_{N_{\text{base}}}, \qquad Z_i \sim \text{Unif}(\mathcal{A}).$$

Each base fact determines a block of c derived queries, so the total number of queries is

$$N = cN_{\text{base}}.$$

For each derived query in the block associated with Z_i, assume the answer is a known one-to-one transformation of Z_i. Thus, an error in reconstructing Z_i induces the same zero-one error on its derived queries.

Proposition A.2 (Deductive closure reduces effective memory load)**.** *Suppose the learner observes M base facts and compresses their labels into B bits. Under the forced-compression analogue of Theorem A.1, and assuming $cM \le N$, the full-space error satisfies*

$$\mathcal{E}_{\text{struct}} \ge \frac{cM}{N}\delta^\star\left(\frac{B}{M}\right) + \left(1 - \frac{cM}{N}\right)\left(1 - \frac{1}{K}\right).$$

Proof. The M observed base labels are independent uniform K-ary symbols. By the same rate-distortion argument as in Theorem A.1, their average reconstruction error is at least $\delta^\star(B/M)$. Each reconstructed base fact supports c derived queries, and the one-to-one maps propagate base-label errors to derived-query errors. Thus, the observed base facts cover cM of the N derived queries. The remaining queries depend on unobserved base facts and have guessing error at least $1 - 1/K$. Combining the covered and uncovered parts gives Eq. (**??**). □

A selective structured variant stores only $L \le M$ base facts with nontrivial fidelity:

$$\mathcal{E}_{\text{struct,sel}} = \min\left\{1 - \frac{1}{K}, \min_{\substack{1 \le L \le M \\ cL \le N}} \Psi(L)\right\},$$

$$\Psi(L) = \frac{cL}{N}\delta^\star\left(\frac{B}{L}\right) + \left(1 - \frac{cL}{N}\right)\left(1 - \frac{1}{K}\right).$$

Compared with the unstructured case, the factor c increases query coverage per stored independent fact. This is the formal sense in which deductive closure relaxes the hallucination floor.

Abstention. The main setting requires the learner to answer every query. If abstention is allowed, the learner can restrict itself to a subset of queries for which it has reliable memory. Suppose the learner answers only a fraction $1 - \gamma$ of the query space and abstains on the remaining fraction γ. Let L be the number of effectively stored facts among the answered queries. Under free addressing, the conditional error on answered queries is

$$\mathcal{E}_{\text{ans}} = \frac{L}{(1-\gamma)N}\delta^\star\left(\frac{B}{L}\right) + \left(1 - \frac{L}{(1-\gamma)N}\right)\left(1 - \frac{1}{K}\right). \tag{13}$$

subject to

$$0 \le L \le M, \qquad 0 \le \gamma < 1, \qquad L \le (1-\gamma)N.$$

This expression makes explicit the coverage–accuracy trade-off: abstention allows memory to be concentrated on fewer answered queries, but lowers answer coverage.

B Theory-Implied Simulation Details

The main text reports representative operating points from the rate-distortion envelope. Here we give the full theory-implied sweeps used to construct Table 2. All values are computed by numerically inverting the K-ary distortion-rate function with $K = 10$ and rounding to three decimals. These are not language-model measurements; they instantiate the idealized envelope analyzed in Section 3.

Numerical inversion. For a target rate r, we compute

$$\delta^\star(r) = \inf\{D \ge 0 : R_K(D) \le r\}$$

by one-dimensional bisection over $D \in [0, 1 - 1/K]$. If $r \ge \log_2 K$, we set $\delta^\star(r) = 0$. If $r \le 0$, we set $\delta^\star(r) = 1 - 1/K$. The reported full-space risks are then obtained by substituting $\delta^\star(B/M)$ into the corresponding operating curves.

B.1 Unstructured Coverage–Compression Decomposition

We use $N = 5{,}000$, $K = 10$, and $B = 160$ bits. The coverage-only floor is

$$\mathcal{E}_{\text{cov}}(M) = \left(1 - \frac{M}{N}\right)\left(1 - \frac{1}{K}\right).$$

For the lossless selective baseline, we use

$$L_{\text{lossless}} = \min\left(M, \left\lfloor \frac{B}{\log_2 K} \right\rfloor\right),$$
$$\mathcal{E}_{\text{sel,lossless}}(M) = \left(1 - \frac{L_{\text{lossless}}}{N}\right)\left(1 - \frac{1}{K}\right).$$

The forced-compression curve is

$$\mathcal{E}_{\text{forced}}(M) = \frac{M}{N}\delta^\star\left(\frac{B}{M}\right) + \left(1 - \frac{M}{N}\right)\left(1 - \frac{1}{K}\right).$$

B.2 Memory Scaling

B.3 Deductive Closure

We use $N = 5{,}000$, $K = 10$, $B = 160$, and closure factor $c = 10$. Each observed base fact supports c derived queries. The structured operating curve is

$$\mathcal{E}_{\text{struct}}(M) = \frac{cM}{N}\delta^\star\left(\frac{B}{M}\right) + \left(1 - \frac{cM}{N}\right)\left(1 - \frac{1}{K}\right)$$

for $cM \le N$.

B.4 Long-Context Analogue

For the long-context analogue, we set $K = 10$ and $B_{\text{ctx}} = 320$ bits. The reported value is the predicted conditional recall error on in-context facts:

$$\mathcal{E}_{\text{ctx}}(M_{\text{ctx}}) = \delta^\star\left(\frac{B_{\text{ctx}}}{M_{\text{ctx}}}\right).$$

This is a theory-implied analogue, not a measurement of any particular long-context model.

C Real-Model Probe Details

This section reports the implementation details for the controlled model probes in Section 4. The goal is to make the empirical signatures auditable: the injected facts are known by construction, the unseen-entity split is disjoint from the injected facts, and observed-query error can be measured directly. Unless otherwise stated, reported errors are mean ± standard error over seeds $\{3407, 2025, 1337\}$.

C.1 Hardware, Software, and Data

All adaptation runs were performed on a single node with eight NVIDIA A100-SXM4-80GB GPUs connected by NVLink. We used distributed data parallel training with `torchrun`, bf16 precision, CUDA 12.1, NVIDIA driver 535.xx, PyTorch 2.4.x, Transformers 4.45.x, PEFT 0.12.x, and Accelerate 0.34.x.

Each synthetic fact is an entity-label mapping

$$e_i \mapsto a_i, \qquad a_i \in \{0, \ldots, K-1\}.$$

We use $K = 10$, so the chance-level error is $1 - 1/K = 0.9$. Labels are sampled uniformly from $\{0, \ldots, 9\}$. Each fact is expressed with two training templates. The unseen-entity pool is the same size as the injected pool for each fact load, and train/test entity overlap is disallowed.

C.2 Adaptation and Decoding Configuration

All fact-injection probes use LoRA adapters. LoRA rank is treated as a controlled capacity proxy and is not interpreted as the theoretical bit budget B. Across models, LoRA is applied to `q_proj`, `v_proj`, and `o_proj`; LoRA dropout is 0.05 and LoRA alpha is set to $2r$, where r is the LoRA rank. We use gradient checkpointing, cosine learning-rate scheduling, warmup ratio 0.03, weight decay 0.0, maximum gradient norm 1.0, and maximum sequence length 96. Checkpoints are saved only at the final training step. Evaluation uses deterministic decoding with temperature 0.0, disabled top-p sampling, `do_sample=false`, and maximum new tokens 4.

C.3 Capacity-Controlled Fact Injection Results

Table 13 reports the Qwen3-8B rank sweep corresponding to Figure 2. Increasing LoRA rank shifts the high-error transition to larger fact loads.

Table 14 reports the fixed-rank model comparison corresponding to Figure 3. All models are evaluated at LoRA rank 16 on seen-direct queries.

C.4 Query-Form Robustness and Structure Results

Table 15 reports split-level robustness results for DeepSeek-R1-Distill-Qwen-32B. The unseen-entity split remains near chance, while paraphrased and relation-rephrased queries degrade earlier than direct queries.

Table 5: Theory-implied sweep for unstructured factual recall with $N = 5{,}000$, $K = 10$, and $B = 160$ bits.

M	B/M	$\delta^\star(B/M)$	Coverage floor	Selective lossless	Forced risk
25	6.400	0.000	0.896	0.896	0.896
50	3.200	0.011	0.891	0.891	0.891
100	1.600	0.275	0.882	0.891	0.888
250	0.640	0.531	0.855	0.891	0.882
500	0.320	0.653	0.810	0.891	0.875
1000	0.160	0.734	0.720	0.891	0.867
2000	0.080	0.787	0.540	0.891	0.855
5000	0.032	0.831	0.000	0.891	0.831

Table 6: Theory-implied sweep for deductive closure with $N = 5{,}000$, $K = 10$, $B = 160$, and $c = 10$.

Observed base facts M	Effective covered queries cM	Unstructured risk	Structured risk
25	250	0.896	0.855
50	500	0.891	0.811
100	1000	0.888	0.775
250	2500	0.882	0.716
500	5000	0.875	0.653

Table 7: Theory-implied sweep for the long-context factual recall analogue with $K = 10$ and $B_{\mathrm{ctx}} = 320$ bits.

M_{ctx}	$B_{\mathrm{ctx}}/M_{\mathrm{ctx}}$	Predicted recall error
50	6.400	0.000
100	3.200	0.011
250	1.280	0.350
500	0.640	0.531
1000	0.320	0.653
2000	0.160	0.734

Table 16 reports the structured versus unstructured comparison corresponding to Figure 5. The structured condition uses closure factor $c = 10$, so the number of independent base facts is approximately one tenth of the surface-query count.

Table 8: Hardware and software environment for real-model probes.

Field	Value
Node count	1
GPU count	8
GPU type	NVIDIA A100-SXM4-80GB
Interconnect	NVLink
Precision	bf16
Distributed backend	`torchrun_ddp`
CUDA	12.1
NVIDIA driver	535.xx
PyTorch	2.4.x
Transformers	4.51.x
PEFT	0.12.x
Accelerate	0.34.x

Table 9: Synthetic fact-injection data construction.

Field	Value
Answer alphabet size	$K = 10$
Chance error	0.900
Fact loads	$M \in \{512, 2048, 8192, 28672\}$
Label distribution	Uniform over $\{0, \ldots, 9\}$
Training templates per fact	2
Evaluation splits	Seen-direct, seen-paraphrase, seen-relation, unseen-entity
Unseen pool size	Same as injected pool
Train/test entity overlap	False
Answer normalization	Strip whitespace; parse first integer in $\{0, \ldots, 9\}$
Seeds	3407, 2025, 1337

Table 10: Common training and decoding configuration for LoRA fact-injection probes.

Field	Value
Optimizer	AdamW
Scheduler	Cosine
Warmup ratio	0.03
Weight decay	0.0
Maximum gradient norm	1.0
Maximum sequence length	96
LoRA target modules	`q_proj`, `v_proj`, `o_proj`
LoRA dropout	0.05
LoRA alpha	$2r$
Gradient checkpointing	True
Save strategy	Final only
Decoding temperature	0.0
Top-p	Disabled
Sampling	`do_sample=false`
Maximum new tokens	4

Table 11: Model-specific adaptation configuration. Global batch size is computed as GPUs $\times$ per-GPU batch $\times$ gradient accumulation.

Model	Rank(s)	LR	Epochs	Per-GPU batch	Grad. accum.	Global batch
Qwen/Qwen3-8B	4, 16, 64	2×10^{-4}	10	8	2	128
Qwen/Qwen3-32B	16	1×10^{-4}	8	2	8	128
Qwen/Qwen3-30B	16	1×10^{-4}	8	2	8	128
DeepSeek	16	8×10^{-5}	8	2	8	128

Table 12: Approximate update counts for each fact load. Each injected fact contributes two training examples.

Model group	M	Training examples	Epochs	Approx. update steps
Qwen3-8B	512	1,024	10	80
Qwen3-8B	2,048	4,096	10	320
Qwen3-8B	8,192	16,384	10	1,280
Qwen3-8B	28,672	57,344	10	4,480
32B / 30B / DeepSeek	512	1,024	8	64
32B / 30B / DeepSeek	2,048	4,096	8	256
32B / 30B / DeepSeek	8,192	16,384	8	1,024
32B / 30B / DeepSeek	28,672	57,344	8	3,584

Table 13: Representative Qwen3-8B rank-sweep results on seen-direct queries for ranks 4, 16, and 64. Errors are mean $\pm$ standard error over seeds $\{3407, 2025, 1337\}$.

Model	Rank	M	Seen-direct error
Qwen3-8B	4	512	0.052 ± 0.006
Qwen3-8B	4	2,048	0.118 ± 0.010
Qwen3-8B	4	8,192	0.704 ± 0.018
Qwen3-8B	4	28,672	0.884 ± 0.006
Qwen3-8B	16	512	0.032 ± 0.005
Qwen3-8B	16	2,048	0.047 ± 0.006
Qwen3-8B	16	8,192	0.382 ± 0.021
Qwen3-8B	16	28,672	0.862 ± 0.010
Qwen3-8B	64	512	0.021 ± 0.004
Qwen3-8B	64	2,048	0.030 ± 0.004
Qwen3-8B	64	8,192	0.105 ± 0.012
Qwen3-8B	64	28,672	0.712 ± 0.026

Table 14: Fixed-rank comparison on seen-direct queries. All models use LoRA rank 16. Errors are mean ± standard error over seeds $\{3407, 2025, 1337\}$.

Model	Rank	M	Seen-direct error
Qwen3-8B	16	512	0.032 ± 0.005
Qwen3-8B	16	2,048	0.047 ± 0.006
Qwen3-8B	16	8,192	0.382 ± 0.021
Qwen3-8B	16	28,672	0.862 ± 0.010
Qwen3-32B	16	512	0.024 ± 0.004
Qwen3-32B	16	2,048	0.036 ± 0.005
Qwen3-32B	16	8,192	0.291 ± 0.019
Qwen3-32B	16	28,672	0.818 ± 0.014
Qwen3-30B-A3B	16	512	0.041 ± 0.006
Qwen3-30B-A3B	16	2,048	0.066 ± 0.008
Qwen3-30B-A3B	16	8,192	0.431 ± 0.024
Qwen3-30B-A3B	16	28,672	0.855 ± 0.012
DeepSeek-R1-Distill-Qwen-32B	16	512	0.035 ± 0.006
DeepSeek-R1-Distill-Qwen-32B	16	2,048	0.052 ± 0.007
DeepSeek-R1-Distill-Qwen-32B	16	8,192	0.334 ± 0.022
DeepSeek-R1-Distill-Qwen-32B	16	28,672	0.846 ± 0.011

Table 15: Query-form robustness for DeepSeek-R1-Distill-Qwen-32B at LoRA rank 16. Errors are mean ± standard error over seeds $\{3407, 2025, 1337\}$.

Model	M	Seen-direct	Seen-para	Seen-relation	Unseen-entity
R1-Distill-Qwen-32B	512	0.035 ± 0.006	0.152 ± 0.014	0.194 ± 0.016	0.889 ± 0.008
R1-Distill-Qwen-32B	2,048	0.052 ± 0.007	0.196 ± 0.013	0.238 ± 0.015	0.901 ± 0.006
R1-Distill-Qwen-32B	8,192	0.334 ± 0.022	0.571 ± 0.026	0.626 ± 0.024	0.897 ± 0.004
R1-Distill-Qwen-32B	28,672	0.846 ± 0.011	0.878 ± 0.008	0.891 ± 0.006	0.902 ± 0.002

Table 16: Structured versus unstructured fact-injection results. Errors are mean ± standard error over seeds $\{3407, 2025, 1337\}$.

Condition	Surface queries	Independent base facts	Error
Unstructured independent	512	512	0.044 ± 0.006
Unstructured independent	2,048	2,048	0.083 ± 0.009
Unstructured independent	8,192	8,192	0.447 ± 0.023
Unstructured independent	28,672	28,672	0.866 ± 0.010
Structured $c = 10$	512	52	0.026 ± 0.004
Structured $c = 10$	2,048	205	0.035 ± 0.005
Structured $c = 10$	8,192	820	0.091 ± 0.011
Structured $c = 10$	28,672	2,868	0.563 ± 0.028