When Less is More: The LLM Scaling Paradox in Context Compression

Ruishan Guo[1,*] **Yibing Liu**[2,*,†] **Guoxin Ma**[3] **Yan Wang**[2]
Yueyang Zhang[2] **Long Xia**[2] **Kecheng Chen**[4] **Zhiyuan Sun**[2] **Daiting Shi**[2]

[1]Tsinghua University [2]Baidu Inc., Beijing, China
[3]Xi'an Jiaotong University [4]City University of Hong Kong

grs24@mails.tsinghua.edu.cn lyibing112@gmail.com

https://rshnnn.github.io/When-Less-Is-More/

Abstract

Scaling up model parameters has long been a prevalent training paradigm driven by the assumption that larger models yield superior generation capabilities. However, under lossy context compression in a compressor–decoder setup, we find a ***Size-Fidelity Paradox***: increasing compressor size can lessen the faithfulness of reconstructed contexts though reconstruction error decreases. Across 27 compressor setups spanning model families, scales, and compression rates, we coin this paradox arising from two dominant factors: 1) *knowledge overwriting*: larger models increasingly replace source facts with their own prior beliefs, *e.g.*, "the white strawberry" $\rightarrow$ "the red strawberry"; and 2) *semantic drift*: larger models tend to paraphrase or restructure content instead of reproducing it verbatim, *e.g.*, "Alice hit Bob" $\rightarrow$ "Bob hit Alice". Interestingly, this paradox persists across varied settings, with mid-sized compressors often outperforming larger ones in faithful recovery. By analyzing the compressed memory via embedding geometry and reconstruction determinacy, we further reveal that compressors tend to organize memory across broader semantic subspaces, yielding more ambiguous representations prone to overwriting, drift, and weakened recovery. These findings complement existing evaluations of context compression and expose a breakdown of scaling laws when the objective shifts from plausible generation to faithful preservation.

1 Introduction

The scaling hypothesis is dominating the training principle of large language models [1–4]. It posits that increasing the number of model parameters leads to enhanced performance, which has been corroborated across a variety of model families (*e.g.*, GPT [5], LLaMA [6], and Qwen [7]) and applications(*e.g.*, recommendaton system [8, 9] and code generation [10]). However, in this paper, we identify a ***Size-Fidelity Paradox*** in context compression problem: *beyond a certain parameter scale, larger models underperform smaller ones in preserving source fidelity*.

Language models in context compression typically act as compressors, mapping natural language into a small set of memory tokens for decoding acceleration [11–15]. In line with the prevalence of scaling hypothesis, prior studies also report that larger compressors yield better compression quality across diverse tasks [16–18]. Yet, as illustrated in Fig. 1, a size-fidelity paradox emerges. When reconstructing the phrase *blue-banded bee* and *the bee shakes pollen loose from flowers*, a lite compressor (*e.g.*, Qwen-3 0.6B) produces only minor literal deviations that preserve the original

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

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

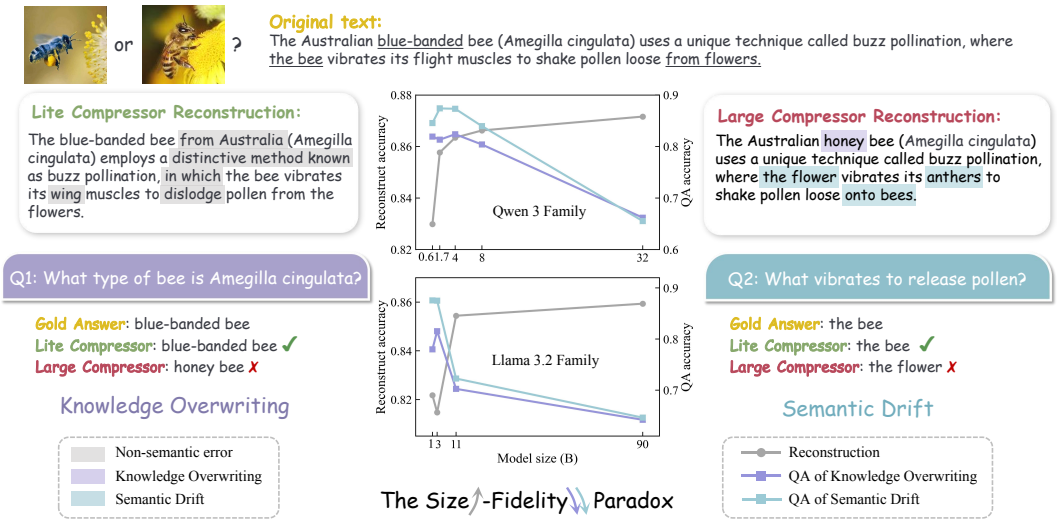


Figure 1: **The Size-Fidelity Paradox in context compression. Left & Right**: A qualitative case study illustrating the breakdown of faithfulness. The lite compressor preserves factual details (Q1, Q2), whereas large compressor produces *knowledge overwriting* by replacing "blue-banded bee" with the prior-driven "honey bee", and *semantic drift* by distorting the causal relation ("bee pollinates flower" $\rightarrow$ "flower pollinates bee"). **Center**: Quantitative results across Qwen and LLaMA show the same pattern: as compressor size increases (x-axis), surface-level *Reconstruction* scores (dashed lines) remain high, yet *QA accuracy* (solid lines) significantly degrades. This divergence shows that stronger surface reconstruction does not ensure faithful preservation of source-specific content.

meaning. In contrast, larger compressors (*e.g.*, LLaMA-3.2 90B) exhibit two critical failure modes: *1) knowledge overwriting*, where source facts (*blue-banded bee*) are replaced with the model internal knowledge, *e.g.*, *honey bee*, and 2) *semantic drift*, where reconstructions deviate from the literal source, *e.g.*, *the flower vibrates its anthers to shake pollen loose onto bees*. While somewhat dissatisfying, standard reconstruction-level evaluations fail to capture this degradation [19–21].

To this end, this paper designs two diagnostic tasks that isolate knowledge overwriting and semantic drift, offering a principled evaluation framework that reveals fidelity failures invisible to surface metrics. Through sweeping experiments across LLM families (Qwen and LLaMA), spanning 0.6B to 90B parameters at multiple compression ratios, we systematically confirm both the existence and generality of this paradox. To further understand how scaling affects compression behavior, we probe compressed memory embeddings along two axes: memory geometry (via effective rank) and reconstruction determinacy (via conditional entropy). On the geometry side, effective rank analysis reveals that larger compressors diffuse memory across more subspaces, weakening source-fact anchoring and promoting knowledge overwriting. On the determinacy side, we find that larger compressors loosen reconstruction constraints, admitting plausible alternatives that manifest as semantic drift. These analyses reveal a fundamental tradeoff: the properties that enable complex reasoning in large models also compromise the rigid fidelity required for faithful reconstruction. Our contributions are summarized as follows:

- We identify a LLM scaling paradox in context compression: larger compressor models can underperform smaller ones though reconstruction performance improves, contradicting prevailing assumptions.
- We introduce CoFiBench (Compression Fidelity Benchmark), a diagnostic benchmark comprising two complementary QA tasks to uncover the underlying performance in context compression regarding *knowledge overwriting* and *semantic drifting* problems. This complements limitations of existing compression evaluations, providing a more robust assessment.
- We analyze compressed memory embeddings through effective rank and conditional entropy, showing how scaling changes memory organization and reconstruction determinacy in ways that weaken source-specific recovery.

2 Related Work

2.1 Scaling Laws in Language Models

Scaling laws have become a central principle in modern language modeling. Extensive work [1, 2, 22, 23] shows that increasing model scale generally reduces training loss and improves performance across many generation and reasoning tasks [9, 10, 24–26], and larger models often exhibit stronger world knowledge and emergent abilities. However, recent studies have increasingly indicated that scaling does not guarantee uniformly better behavior. In particular, some work [27–30] reports that larger models can be less reliable on challenging inputs, for example by making high-confidence errors or showing weaker calibration. While these findings highlight important limitations of scaling, they mainly examine answer correctness and reliability in question answering or reasoning [31, 32], thus providing limited insight into how scaling behaves in tasks requiring faithful input reproduction.

Our work addresses this gap by studying scaling behavior in a fidelity-critical compressor–decoder setting, where the objective is to reconstruct the source text with minimal distortion rather than generate a plausible paraphrase. Notably, while larger compressors can achieve better reconstruction metrics, smaller models perform better on fidelity-oriented downstream evaluation, revealing a size–fidelity paradox. Beyond identifying this phenomenon, we further analyze why scaling causes such degradation and how scale-amplified properties can undermine literal reconstruction.

2.2 Evaluation of Long-Context Compression

Long-context compression has been widely studied to reduce the cost of long-context modeling [33, 34] and to make long inputs usable under limited context windows [18, 35–40]. Most approaches [11, 16, 41] follow a compressor–decoder paradigm, where a long sequence is encoded into a compact latent representation and a decoder reconstructs text conditioned on it. Correspondingly, prior work [11, 16] largely evaluates these methods through reconstruction-centric protocols, comparing performance across compression ratios and model scales using standard metrics(e.g., training loss, perplexity, BERTScore [42] and n-gram overlap scores such as BLEU [43] and ROUGE [44]).

While convenient, such metrics can be superficial and incomplete proxies for compression quality. They primarily reward fluency and surface similarity, but do not distinguish information genuinely recovered from the compressed representation from plausible content supplied by a model's parametric priors. As a result, a model can obtain high reconstruction scores while still changing or losing important details from the original text. Motivated by this limitation, we propose a usage-oriented evaluation perspective that assesses compression by its functional support for realistic downstream tasks, providing a more faithful measure of what the compressed representation actually preserves.

3 Preliminaries

Compressor-Decoder Architecture. The prevalent training paradigms in context compression [11, 16, 38, 41, 45–50] employs compressor-decoder setup, where one language model acts as a compressor transforming the discrete input sequence into a set of continuous latent embeddings (memory tokens); whereas, another language model performs as the decoder, targeting at reconstructing compressed representation to original text.

Formally, let $\mathcal{V}$ denote a vocabulary of tokens. Given an input sequence $\mathbf{x} = (x_1, \ldots, x_L) \in \mathcal{V}^L$, the objective of context compression is to learn a compressor f_θ, which maps the input $\mathbf{x}$ to a latent tensor $\mathbf{Z} \in \mathbb{R}^{M \times d}$:

$$\mathbf{Z} = f_\theta(\mathbf{x}), \quad \text{where } M \ll L. \tag{1}$$

Here, d represents the hidden dimension, and M is the number of memory slots. The compression rate is defined as $\rho = L/M$. The decoder g_ϕ operates as a causal language model conditioned on $\mathbf{Z}$. It estimates the probability distribution of the original sequence:

$$P_\phi(\mathbf{x}|\mathbf{Z}) = \prod_{t=1}^{L} P_\phi(x_t|x_{<t}, \mathbf{Z}). \tag{2}$$

Training objective. The training loss of compressor combines reconstruction fidelity and generative capability:

$$\mathcal{L} = \mathcal{L}_{\text{re}} + \mathcal{L}_{\text{nt}} = -\sum_{t=1}^{k} \log P_\phi(x_t|x_{<t}, \mathbf{Z}) - \sum_{t=k+1}^{n} \log P_\phi(x_t|x_{<t}, \mathbf{Z}), \tag{3}$$

here, $\mathcal{L}_{\text{re}}$ enforces faithful reconstruction of the compressed prefix $x_{1:k}$ conditioned on $\mathbf{Z}$, while $\mathcal{L}_{\text{nt}}$ applies standard autoregressive prediction to the continuation $x_{k+1:n}$.

Unlike abstractive summarization which allows paraphrasing, our objective is exact reconstruction, any deviation from the verbatim input is treated as information loss.

4 The Size-Fidelity Paradox

In this section, we characterize the **Size-Fidelity Paradox** through two dissection tasks designed to isolate distinct failure modes. Specifically, we demonstrate that larger models exhibit a heightened propensity for *knowledge overwriting*, where source facts are substituted with internal knowledge, and *semantic drift*, in which loose restructuring precedence over verbatim preservation. These results confirm that fidelity loss is a systematic consequence of scaling rather than random error.

4.1 Model Selection & Scale Range

To enable a rigorous analysis of scaling behavior under context compression, we evaluate compressor models drawn from two widely used model families, Qwen-3 [7] and LLaMA-3.2 [6], spanning nearly three orders of magnitude in parameter count, from 0.6B to 90B. For each model size, we train compressors at three compression rates ($4\times$, $16\times$, and $64\times$). In the main experiments, all compressors are evaluated with the fixed Meta-LLaMA-3.2-8B-Instruct decoder. All models are trained on high-quality text chunks from the Fineweb dataset [51], using an identical training protocol to minimize confounding factors. Tab. 1 summarizes our dataset configuration for these experiments. We provide full reproducibility details in Appendix B, including training hyperparameters, memory-token configuration, dataset preprocessing, chunking and filtering rules.

Table 1: Task and dataset statistics.

Task	Dataset	#Contexts	#QAs
Reconstruction	Fineweb	4,444,672	–
Knowledge Overwriting	FaithEval	598	598
	ConflictQA	1356	1356
Semantic Drift	Fineweb	449	8980
	FaithEval	598	11960

Before analyzing the failure modes, we first examine the standard training signal. Larger compressors descend faster and reach lower loss (Fig. 2), suggesting more efficient optimization; however, this optimization advantage does not imply faithful compression.

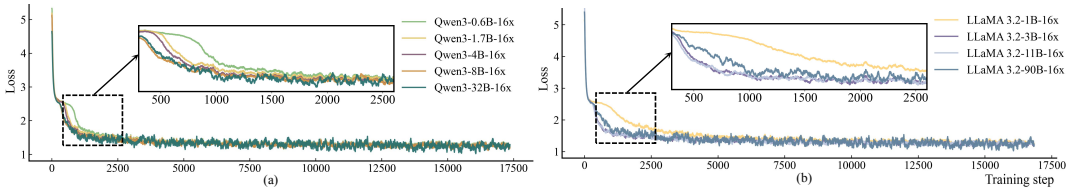


Figure 2: Training loss for Qwen (left) and LLaMA (right) compressors at $16\times$ compression.

A similar limitation appears in standard reconstruction metrics. Tab. 2 reports BLEU [43], ROUGE-L [44], chrF++ [52], and BERTScore-F1 [53] for Qwen3 compressors at $4\times$ compression on FineWeb, covering lexical, sequence-level, character-level, and semantic overlap. Across scales, these scores change only mildly, while BERTScore-F1 nearly saturates, indicating limited sensitivity to source-critical facts and relations.

Table 2: Standard reconstruction metrics for Qwen3 compressors at $4\times$ compression.

#Params	BLEU	ROUGE-L	chrF++	BERTScore	Avg.
0.6B	0.952	0.969	0.983	0.997	0.975
1.7B	0.941	0.957	0.975	0.995	0.967
4B	0.943	0.964	0.978	0.996	0.970
8B	0.940	0.955	0.969	0.995	0.965
32B	**0.968**	**0.982**	**0.991**	**0.998**	**0.985**

Table 3: **Quantitative Evaluation of the Size-Fidelity Paradox across Qwen and LLaMA Families (0.6B–90B).** BLEU represents surface reconstruction, while diagnostic QA scores measure *knowledge overwriting* and *semantic drift* across compression rates. **Bold** marks the best model within each family/compression group. Shaded Avg./Δ columns denote category averages and drops from the group best; red marks the worst QA average/drop. Full results are provided in Appendix C.

Compression Rate	Model	#Params	Reconstruction		*Knowledge Overwriting*				*Semantic Drift*			
					FaithEval	ConflictQA	Avg.	Δ	FineWeb	FaithEval	Avg.	Δ
4×	Qwen 3	0.6b	0.95	(-2.06%)	0.71	**0.95**	0.83	(-1.19%)	0.83	**0.83**	**0.83**	——
		1.7b	0.94	(-3.09%)	**0.76**	0.89	0.83	(-1.19%)	0.83	0.81	0.82	(-1.20%)
		4b	0.94	(-3.09%)	0.72	**0.95**	**0.84**	——	**0.87**	0.79	**0.83**	——
		8b	0.94	(-3.09%)	0.74	0.91	0.83	(-1.19%)	0.79	**0.83**	0.81	(-2.41%)
		32b	**0.97**	——	0.68	0.80	0.74	(-11.90%)	0.60	0.78	0.69	(-16.87%)
	LLaMA 3.2	1b	0.94	(-3.09%)	**0.77**	0.88	**0.83**	——	0.78	**0.85**	**0.82**	——
		3b	0.90	(-7.22%)	0.74	**0.91**	**0.83**	——	**0.79**	0.84	**0.82**	——
		11b	0.92	(-5.15%)	0.65	0.80	0.73	(-12.05%)	0.64	0.76	0.70	(-14.63%)
		90b	**0.97**	——	0.60	0.81	0.71	(-14.46%)	0.59	0.68	0.64	(-21.95%)
16×	Qwen 3	0.6b	0.83	(-3.49%)	0.70	**0.94**	0.82	(-1.20%)	0.75	0.84	0.80	(-3.61%)
		1.7b	**0.86**	——	**0.75**	0.88	0.82	(-1.20%)	0.79	**0.86**	**0.83**	——
		4b	**0.86**	——	0.71	**0.94**	**0.83**	——	**0.82**	0.83	**0.83**	——
		8b	0.84	(-2.33%)	0.72	0.89	0.81	(-2.41%)	0.73	0.79	0.76	(-8.43%)
		32b	0.85	(-1.16%)	0.61	0.72	0.67	(-19.28%)	0.57	0.74	0.66	(-20.48%)
	LLaMA 3.2	1b	0.82	(-4.65%)	0.72	0.84	0.78	(-4.88%)	**0.75**	**0.90**	**0.83**	——
		3b	0.81	(-5.81%)	**0.73**	**0.90**	**0.82**	——	**0.75**	**0.90**	**0.83**	——
		11b	0.85	(-1.16%)	0.63	0.77	0.70	(-14.63%)	0.61	0.73	0.67	(-19.28%)
		90b	**0.86**	——	0.55	0.74	0.65	(-20.73%)	0.56	0.63	0.60	(-27.71%)
64×	Qwen 3	0.6b	0.17	(-41.38%)	0.44	0.59	0.52	(-17.74%)	0.31	0.39	0.35	(-20.45%)
		1.7b	0.21	(-27.59%)	0.51	0.60	0.56	(-11.29%)	0.35	0.42	0.39	(-11.36%)
		4b	0.27	(-6.90%)	**0.54**	**0.70**	**0.62**	——	**0.41**	0.46	**0.44**	——
		8b	0.28	(-3.45%)	**0.54**	0.67	0.61	(-3.23%)	0.38	**0.49**	**0.44**	——
		32b	**0.29**	——	0.47	0.56	0.52	(-17.74%)	0.27	0.35	0.31	(-29.55%)
	LLaMA 3.2	1b	0.19	(-42.42%)	0.49	0.57	0.53	(-14.52%)	0.33	0.44	0.39	(-7.32%)
		3b	0.29	(-12.12%)	**0.55**	**0.68**	**0.62**	——	**0.35**	**0.47**	**0.41**	——
		11b	0.30	(-9.09%)	0.51	0.63	0.57	(-8.06%)	0.24	0.33	0.29	(-31.71%)
		90b	**0.33**	——	0.45	0.61	0.53	(-14.52%)	0.24	0.31	0.28	(-34.15%)

4.2 Dissection 1: Knowledge Overwriting

Knowledge overwriting refers to a failure mode in which a compression model prioritizes its parametric world knowledge over conflicting source context, as illustrated below, generating outputs that deviate from the compressed input and exhibit reduced factual fidelity.

> **Input** Einstein was born in **France**. *Q:* birthplace?
> **No overwrite** *Recons.* In **France**, Einstein was born. *A:* France ✓
> **Overwrite** *Recons.* Einstein was born in **Germany**. *A:* Germany ✗; world prior replaces source

Knowledge Overwriting Evaluation. To isolate this phenomenon, we construct a controlled counterfactual evaluation based on FaithEval [20] and ConflictQA [54]. Both datasets introduce explicit factual contradictions into natural contexts (e.g., replacing "Einstein was born in Germany" with "Einstein was born in France"). We retain unambiguous examples with established facts, ensuring a clear conflict between source evidence and parametric knowledge.

Each counterfactual context is compressed and paired with a question targeting the modified fact. Candidate answers include *i) the source-consistent counterfactual answer* and *ii) plausible world-knowledge alternatives*. We measure *knowledge faithfulness* as the accuracy of selecting the source-consistent answer, i.e., prioritizing compressed evidence over internal priors. Dataset construction and filtering procedures are detailed are provided in Appendix D, with an optional candidate-generation prompt included for reference.

Results and Analysis. Table 3 reveals a non-monotonic scale effect: larger compressors do not necessarily improve faithfulness and are more susceptible to parametric-prior intrusion. For example, at 64×, LLaMA-90B improves BLEU over LLaMA-3B (0.33 vs. 0.29) but drops clearly in overwriting accuracy (0.53 vs. 0.62). Notably, compressors around the 3–4B scale emerge as a particularly stable and effective regime, often achieving better faithfulness across model families.

4.3 Dissection 2: Semantic Drift

Semantic drift refers to fluent, topically coherent compression that nevertheless distorts the semantic or relational structure of the source. Unlike knowledge overwriting, it arises from structural degradation during paraphrastic compression rather than conflict with parametric knowledge, with the case below highlighting a typical scope shift.

Input	The **dwarves** are funny, except maybe **Thorin**. *Q:* exception?
No drift	*Recons.* The dwarves are funny, though **Thorin** may be an exception. *A:* Thorin ✓
Scope-drift	*Recons.* Except for **the dwarves**, Thorin is funny. *A:* the dwarves ✗; exception scope shifts

Semantic Drift Evaluation. Existing benchmarks focus on topical relevance or factual recall, and are largely insensitive to structural distortions such as role reversal or modifier loss. We therefore build a diagnostic QA dataset over FineWeb [51] and FaithEval [20], using DeepSeek-R1 [55] to generate questions that require precise relations rather than topical gist.

We cover seven drift dimensions: *main topic*, *entity list*, *predicate exactness*, *relation anchor*, *coreference*, *role binding*, and *modifier scope*, probing shifts in subjects, entities, predicates, discourse links, antecedents, roles, and qualifiers. Answerable questions require exact-substring answers from the source; the callout below illustrates how reconstruction can shift scope despite topical fluency. Tab. 3 reports the micro-averaged QA accuracy across all seven dimensions. Per-dimension results and generation/filtering details are in Appendices F, D.1, and D.3.

Results and Analysis. Tab. 3 shows the same non-monotonic scale effect for structural QA accuracy: fidelity improves up to the 3–4B range but declines for larger compressors, even as BLEU continues to increase. This divergence points to semantic drift rather than factual overwriting: larger models better preserve fluent semantic gist, but become less sensitive to fine-grained relational constraints. Overall, scaling appears to favor abstract semantic compression over precise structural preservation.

5 Generalization and Downstream Impact

5.1 Generalization Analysis

To verify that the Size–Fidelity Paradox does not arise from a particular experimental configuration, we examine its transferability through ***(1) decoder compatibility***, ***(2) cross-domain robustness***, and ***(3) loss-weight sensitivity experiments***. These settings vary the decoder, the pretraining corpus, and the training-objective balance, respectively, while keeping the main compression protocol unchanged.

5.1.1 Decoder Compatibility

In our main experiments, we use a fixed Meta-LLaMA-3.2-8B-Instruct decoder. In this section, we train compressors from the Qwen3 family (0.6B, 4B, 8B) at 16× compression and evaluate them with Qwen3 decoders of varying capacity: 0.6B and 4B. This cross-family transfer (LLaMA→Qwen) tests whether the paradox generalizes beyond architectural specifics, while the decoder size ablation tests whether smaller decoders expose or suppress the fidelity degradation we observe in larger models.

Table 4: Decoder generalizability ablation with Qwen3 compressors at 16× compression and Qwen3-0.6B/4B decoders.

Compressor	Recons.↑	QA(i)↑		QA(ii)↑	
	Fineweb	FaithEval	ConflictQA	Fineweb	FaithEval
Decoder: Qwen3-0.6B	*Compression Rate:16x*				
Qwen3-0.6b	0.62	0.59	0.63	0.58	0.66
Qwen3-4b	0.66	0.54	0.62	0.58	0.68
Qwen3-8b	0.61	0.53	0.59	0.52	0.61
Decoder: Qwen3-4B	*Compression Rate:16x*				
Qwen3-0.6b	0.76	0.65	0.80	0.62	0.75
Qwen3-4b	0.79	0.66	0.81	0.63	0.77
Qwen3-8b	0.77	0.62	0.75	0.57	0.71

Robustness to Decoder Scaling. Tab. 4 presents reconstruction accuracy and QA performance across decoder configurations. When the decoder size is reduced from 8B to 4B, we observe a slight baseline drop in reconstruction metrics. However, the paradox remains pronounced: the 8B compressor continues to exhibit significantly higher rates of knowledge overwriting and semantic drift compared to the 4B compressor.

When using the 0.6B decoder, all compressors experience substantial degradation in both reconstruction and QA tasks, reflecting the decoder's limited generation capacity. Crucially, this bottleneck narrows the performance gap between compressors—the 8B compressor's tendency toward overwriting and drift is partially suppressed by the decoder's constraints. However, the relative ordering persists: even under a weak decoder, the 8B compressor shows higher error rates on both QA tasks compared to the 0.6B and 4B compressors. This suggests that while decoder capacity modulates the severity of the paradox, it does not eliminate the underlying mechanism. We further pair Qwen3-32B with stronger decoders, includ-

ing a matched 32B decoder, in Appendix E; it still underperforms the 4B baseline on fidelity-oriented QA, suggesting that decoder under-capacity is not the main cause.

5.1.2 Cross-domain Robustness

To assess the out-of-domain generalizability of the paradox, we additionally pretrain Qwen3 compressors (at 16× compression) on ArxivCorpus (≈20.9B tokens) [56], a scientific corpus of academic abstracts, and apply the same reconstruction and diagnostic QA protocols. As shown in Tab. 5, reconstruction improves with model scale, while fidelity-oriented QA remains non-monotonic: the larger compressor underperforms the mid-sized compressor on both knowledge-overwriting and semantic-drift probes. This confirms that the scale–fidelity tradeoff persists beyond the original FineWeb pretraining domain.

Table 5: Cross-domain robustness on ArxivCorpus.

Compressor	Recons.↑	QA(i)↑		QA(ii)↑	
	ArxivCorpus	FaithEval	ConflictQA	Fineweb	FaithEval
Qwen3-0.6B	0.72	0.60	0.77	0.71	0.76
Qwen3-4B	0.76	**0.69**	**0.82**	**0.79**	**0.81**
Qwen3-8B	**0.77**	0.63	0.79	0.74	0.77

5.1.3 Loss-weight Sensitivity

A possible concern is that different model scales may require different balances between reconstruction and next-token prediction during pretraining. We therefore conduct a λ-sweep, where λ weights the next-token prediction loss$\mathcal{L}_{\mathrm{nt}}$ and $(1-\lambda)$ controls the reconstruction loss$\mathcal{L}_{\mathrm{re}}$, and evaluate Qwen3 compressors (at 4× compression) with the same reconstruction and diagnostic QA protocols.

Tab. 6 shows that the loss balance shifts the reconstruction–fidelity tradeoff, but does not restore monotonic scaling in fidelity-oriented QA. Larger compressors still fail to consistently outperform smaller or mid-sized ones on knowledge overwriting and semantic drift, suggesting that the scale–fidelity tradeoff is not merely an artifact of applying the same loss mixture to models of different sizes.

Table 6: Loss-weight sensitivity at 4× compression.

Conpressor	λ	Recons.↑	QA(i)↑		QA(ii)↑	
		FineWeb	Faith.	Conf.	FineWeb	Faith.
Qwen3-0.6B	0.1	0.90	0.69	0.89	0.77	0.75
	0.3	0.93	**0.71**	0.92	0.80	0.79
	0.5 (default)	**0.95**	**0.71**	**0.95**	**0.83**	**0.83**
	0.7	0.91	0.69	0.90	0.78	0.77
	0.9	0.87	0.66	0.85	0.74	0.71
Qwen3-4B	0.3	**0.94**	0.70	0.91	0.86	0.77
	0.5 (default)	**0.94**	**0.72**	**0.95**	**0.87**	**0.79**
	0.7	0.91	0.68	0.90	0.82	0.78

5.2 Downstream Impact on RAG-based QA and ICL

To further examine downstream effects of the scale–fidelity tradeoff, we evaluate Qwen3 compressors from 0.6B to 32B on retrieval-augmented generative question answering (RAG-based QA) and in-context learning (ICL).

Table 7: Results on RAG-based QA.

RAG-based QA	MuSiQue		QASPER	
	F1(%)↑	EM(%)↑	F1(%)↑	EM(%)↑
w/o Context	9.11	2.67	13.34	5.87
w/ Context	**33.26**	**23.91**	**46.22**	**31.90**
Qwen3-0.6b 4×	<u>24.14</u>	<u>15.08</u>	28.93	19.16
Qwen3-1.7b 4×	22.67	13.65	<u>29.46</u>	<u>19.37</u>
Qwen3-4b 4×	20.75	11.31	28.06	18.82
Qwen3-8b 4×	16.93	9.10	24.35	15.72
Qwen3-32b 4×	14.55	0.53	23.94	13.18

For RAG-based QA, each example is a (context, question, answer) triple. We compare three settings: (i) *w/o Context*, where the inference model receives only the question; (ii) *w/ Context*, where it receives the original context; and (iii) *4× Compressed Context*, where the context is compressed at 4× and reconstructed by the fixed decoder before QA. We evaluate on MuSiQue [57], which requires connected multi-hop evidence composition, and QASPER [58], which targets information-seeking QA over scientific documents, using F1 and Exact Match (EM).

For ICL, we evaluate five BBH subsets [59–61]: *Date Understanding*, *Causal Judgement*, *Logical Deduction*, *Formal Fallacies*, and *Tracking Shuffled Objects*, reporting EM accuracy over 250/187 evaluation examples for each subset. We compare (i) *Zero-shot*, (ii) *Full ICL*, and (iii) *4× Compressed ICL*; in the compressed setting, only the demonstration block is compressed and reconstructed, while the test query remains unchanged.

Tab. 7 reports the results on RAG-based QA. Across MuSiQue and QASPER, the 8B compressor performs worse than the 0.6B and 1.7B compressors, with a similar decline observed for 32B.

Table 8: Results on ICL.

ICL	BBH / EM(%)↑					
	Date.	Causal.	Logic.	Formal.	Track.	Avg.
Zero-shot	43.2	52.9	49.6	53.6	50.8	50.0
Full ICL	**62.8**	**75.4**	**70.4**	**76.0**	**67.2**	**70.4**
Qwen3-0.6b 4×	58.4	70.6	66.8	69.6	60.4	65.2
Qwen3-1.7b 4×	58.8	70.1	66.8	69.2	59.2	64.8
Qwen3-4b 4×	57.2	69.0	67.6	67.6	56.8	63.6
Qwen3-8b 4×	55.6	65.2	63.2	65.2	54.8	60.8
Qwen3-32b 4×	55.2	64.2	63.6	64.0	57.6	60.9

Tab. 8 reports the results on BBH ICL tasks, where compressing the demonstrations with larger compressors again leads to lower accuracy, especially for the 8B and 32B models. Overall, the downstream results align with our diagnostic findings: when task performance depends on source-specific facts, relations, or demonstrations, larger compressors can be less reliable despite stronger reconstruction signals. This suggests that compressor scale should be considered together with the fidelity requirements of the target application.

6 Beyond Scale: How Compression Changes with Model Size?

Having established the Size-Fidelity Paradox, we now look beyond parameter count. Fidelity in context compression depends on what the memory embeddings $\mathbf{Z} \in \mathbb{R}^{m \times d}$ (defined in Sec. 3) preserve about the source, and how tightly they constrain reconstruction.

We follow the same diagnostic logic for two probes: first identify how the probe changes with scale, then use fixed-scale training dynamics to interpret whether the compression objective favors high or low values, and finally test whether the probe predicts the failure mode implied by its meaning. Effective rank probes memory organization and *knowledge overwriting*; conditional entropy probes reconstruction determinacy and *semantic drift*.

6.1 How Scale Reorganizes Memory: Source Alignment vs. Semantic Dispersion

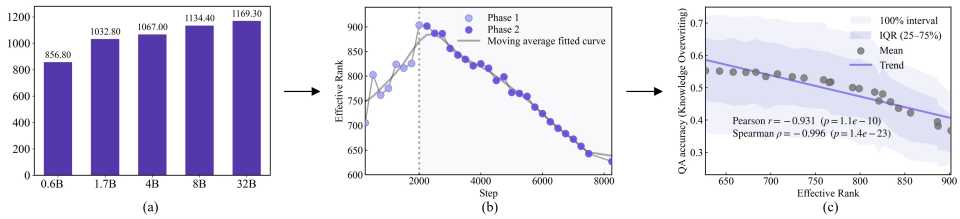


Figure 3: **Effective-rank diagnosis.** (a) Scaling broadens the active memory subspace; (b) fixed-scale training compresses it after early expansion; (c) higher post-peak rank predicts lower knowledge-faithfulness QA accuracy.

Knowledge overwriting is a source-anchoring failure: when the compressed context conflicts with parametric knowledge, reconstruction may recover a plausible prior fact rather than the source fact. If this failure is tied to memory geometry, the key question is whether the active subspace remains aligned with the particular input instead of expanding into nearby semantic alternatives.

We use effective rank as a probe of how broadly the compressed memory uses its latent space. Given the batch-stacked memory matrix $\mathbf{Z}_{\text{batch}}$, we compute

$$\text{erank}(\mathbf{Z}) = \exp\left(-\sum_{i=1}^{r} p_i \log p_i\right), \qquad p_i = \frac{\sigma_i}{\sum_{j=1}^{r} \sigma_j}, \tag{4}$$

where σ_i are the singular values of $\mathbf{Z}_{\text{batch}}$. Higher effective rank means that representation mass spreads across more active directions (Appendix G.1 reports erank/d across scales). Such spread can encode richer semantic neighborhoods, but faithful compression requires source-aligned directions.

The scale sweep gives the first clue: effective rank increases with compressor size, so larger compressors activate a broader memory subspace (Fig. 3(a)). This helps explain why scaling can improve surface reconstruction while weakening fidelity. A broader subspace may preserve the source gist, yet keep prior-compatible alternatives close enough to compete with rare or counterfactual facts.

Training dynamics reveal the sign of this probe. With model size fixed, rank first expands then contracts (Fig. 3(b)): early optimization opens semantic directions to separate inputs, while later optimization prunes toward a source-sufficient structure. Thus, high rank is not inherently desirable; after basic reconstruction is learned, compression favors a compact, source-aligned subspace.

This interpretation predicts the failure link: if rank measures semantic breadth, excessive rank should hurt when source evidence must override prior knowledge. In the post-peak regime, higher effective rank correlates with lower knowledge-faithfulness QA accuracy (Fig. 3(c)). The rank-intervention results in Appendix G.2 supports this structure: moderate spectral truncation improves fidelity by trimming the low-energy tail, whereas aggressive truncation removes source-bearing directions and hurts reconstruction. The useful bottleneck is not the smallest one, but one whose leading subspace preserves source evidence without inviting unnecessary variation.

6.2 Memory Constraints: Source Recovery vs. Plausible Alternatives

Semantic drift is an under-constraint failure: reconstruction remains topical and fluent, but changes roles, modifiers, or causal relations. This suggests that memory supports plausible continuations without forcing the source-specific one, which conditional entropy probes under the same decoder.

Given $\mathbf{Z} = f_\theta(\mathbf{x})$, the decoder defines $P_\phi(\mathbf{x} \mid \mathbf{Z}) = \prod_{t=1}^{L} P_\phi(x_t \mid x_{<t}, \mathbf{Z})$. At each reconstruction step, we compute the teacher-forced token entropy $H_t(\mathbf{Z}) = -\sum_{v \in \mathcal{V}} P_\phi(v \mid x_{<t}, \mathbf{Z}) \log P_\phi(v \mid x_{<t}, \mathbf{Z})$, and average it over tokens to obtain $H(\mathbf{Z})$. By conditioning on the gold prefix, this entropy abstracts away sampling-induced variation and directly reflects how tightly the memory constrains the next source token. Low entropy indicates determinate reconstruction, whereas high entropy means that multiple plausible continuations remain competitive.

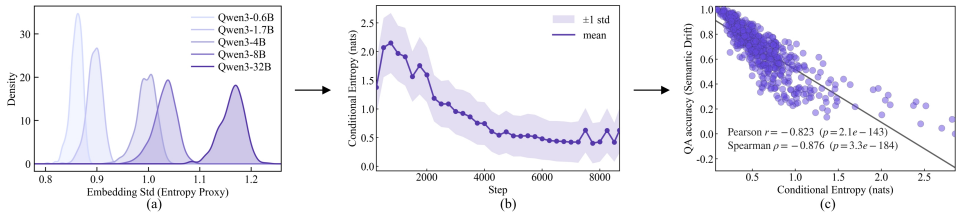


Figure 4: **Conditional-entropy diagnosis.** (a) Larger compressors yield higher conditional entropy, leading to less determinate reconstruction; (b) fixed-scale training reduces conditional entropy; (c) higher conditional entropy predicts lower semantic-drift QA accuracy.

The scale trend identifies conditional entropy as the probe: larger compressors show higher entropy than mid-sized ones (Fig. 4(a); Appendix H.1). Training again gives the sign. For a fixed model, entropy steadily decreases as optimization sharpens the source token path (Fig. 4(b)). Scaling therefore counteracts training, leaving more plausible alternatives under the same bottleneck.

This ambiguity is what semantic drift exposes. Higher conditional entropy correlates with lower semantic-drift QA accuracy (Fig. 4(c)): when memory does not sharply select the source continuation, role bindings can shift, modifiers can attach to the wrong entity, and causal relations can be smoothed into generic associations. The output remains coherent, but no longer preserves the fine-grained structure required by faithful compression.

7 Discussion: Hallucination or Fidelity Failure?

Table 9: **Fidelity failures vs. hallucination.**

	Hallucination	*Knowledge overwrite*	*Semantic drift*
Source	The drug reduced tumor growth in mice. Query: Was it tested in humans?	The white strawberry was sliced for the tart.	The doctor called the nurse after the patient fainted.
Output	Yes, the drug succeeded in a 2023 human clinical trial.	The red strawberry was sliced for the tart.	The nurse called the doctor after the patient fainted.

Although knowledge overwriting and semantic drift can superficially resemble hallucination, they describe a different failure regime. **Hallucination** is typically an output-level grounding error, where a model introduces content unsupported by evidence or inconsistent with external truth. Context compression instead defines a **lossy communication** problem: the source $\mathbf{x}$ must be transmitted through a compact memory $\mathbf{Z}$ and reconstructed as $\hat{\mathbf{x}}$. A reconstruction is erroneous whenever it fails to preserve source-specific facts or relations, even if the substituted content is fluent, plausible, or factually correct in the world. Thus, the Size-Fidelity Paradox is not simply that larger models hallucinate more under compression; rather, scaling changes how source information is represented and resolved, improving plausible reconstruction while weakening **source-faithful recovery**.

8 Conclusion

This work has demonstrated that the prevailing intuition behind parameter scaling does not straightforwardly transfer to lossy context compression. Across experiments ranging from 0.6B to 90B model parameters, we have identified a *Size-Fidelity Paradox*: larger compressors achieve lower training loss yet produce less faithful compressed representations. To pinpoint how fidelity erodes, we have introduced two diagnostic QA tasks that separately measure knowledge overwriting and semantic drift, while isolating the effects of model scale. By holding parameter count fixed and probing the compressed representations, we have demonstrated that scaling changes how source information is organized in memory. Larger compressors tend to use broader semantic memory structures and can leave more plausible reconstructions available, weakening source-specific recovery. These diagnostics reveal fidelity degradations that standard reconstruction evaluations overlook, establishing a principled guideline for assessing context compression effectiveness. More broadly, this work challenges the universality of scaling laws, highlighting domains where larger models may require fundamentally different design principles to achieve desired behaviors.

9 Acknowledgements and Disclosures

This work was supported by the AI Search Team at Baidu. We thank the team members for their valuable discussions and feedback throughout this research. We also thank the anonymous reviewers for their constructive comments, which helped improve the clarity and presentation of this work. The authors' affiliations with Baidu are disclosed in the author list. The authors declare no other competing interests.

References

[1] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[2] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[3] Hanyu Lai, Xiao Liu, Junjie Gao, Jiale Cheng, Zehan Qi, Yifan Xu, Shuntian Yao, Dan Zhang, Jinhua Du, Zhenyu Hou, et al. A survey of post-training scaling in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2771–2791, 2025.

[4] Ayan Sengupta, Yash Goel, and Tanmoy Chakraborty. How to upscale neural networks with scaling law? a survey and practical guidelines. *arXiv preprint arXiv:2502.12051*, 2025.

[5] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[6] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[7] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, and Chenxu Lv. Qwen3 technical report. 2025.

[8] Buyun Zhang, Liang Luo, Yuxin Chen, Jade Nie, Xi Liu, Daifeng Guo, Yanli Zhao, Shen Li, Yuchen Hao, Yantao Yao, et al. Wukong: Towards a scaling law for large-scale recommendation. *arXiv preprint arXiv:2403.02545*, 2024.

[9] Hongru Cai, Yongqi Li, Ruifeng Yuan, Wenjie Wang, Zhen Zhang, Wenjie Li, and Tat-Seng Chua. Exploring training and inference scaling laws in generative retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1339–1349, 2025.

[10] Wenxuan Yang, Qingqv Wei, Chenxi Ma, Weimin Tan, and Bo Yan. Scaling laws for data-efficient visual transfer learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2968–2976, 2025.

[11] Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. *arXiv preprint arXiv:2307.06945*, 2023.

[12] Zongqian Li, Yinhong Liu, Yixuan Su, and Nigel Collier. Prompt compression for large language models: A survey. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7182–7195, 2025.

[13] Chengzhengxu Li, Xiaoming Liu, Zhaohan Zhang, Shengchao Liu, Guoxin Ma, Yu Lan, Cong Wang, and Chao Shen. Can reasoning path still be effective as input? bridging post-reasoning to chain-of-thought compression. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 39014–39034, 2026.

[14] Jie He, Richard He Bai, Sinead Williamson, Jeff Z Pan, Navdeep Jaitly, and Yizhe Zhang. Clara: Bridging retrieval and generation with continuous latent reasoning. *arXiv preprint arXiv:2511.18659*, 2025.

[15] Yurii Kuratov, Mikhail Arkhipov, Aydar Bulatov, and Mikhail Burtsev. Cramming 1568 tokens into a single vector and back again: Exploring the limits of embedding space capacity. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19323–19339, 2025.

[16] Yuhong Dai, Jianxun Lian, Yitian Huang, Wei Zhang, Mingyang Zhou, Mingqi Wu, Xing Xie, and Hao Liao. Pretraining context compressor for large language models with embedding-based memory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28715–28732, 2025.

[17] Xiaoqiang Lin, Aritra Ghosh, Bryan Kian Hsiang Low, Anshumali Shrivastava, and Vijai Mohan. Refrag: Rethinking rag based decoding. *arXiv preprint arXiv:2509.01092*, 2025.

[18] Gabriele Berton, Jayakrishnan Unnikrishnan, Son Tran, and Mubarak Shah. Compllm: Compression for long context q&a. *arXiv preprint arXiv:2509.19228*, 2025.

[19] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*, 2021.

[20] Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. Faitheval: Can your language model stay faithful to context, even if" the moon is made of marshmallows". *arXiv preprint arXiv:2410.03727*, 2024.

[21] Yongchao Long, Xian Wu, Yingying Zhang, Xianbin Wen, Yuxi Zhou, and Shenda Hong. Copy-paste to mitigate large language model hallucinations. *arXiv preprint arXiv:2510.00508*, 2025.

[22] Yangjun Ruan, Chris J Maddison, and Tatsunori B Hashimoto. Observational scaling laws and the predictability of langauge model performance. *Advances in Neural Information Processing Systems*, 37:15841–15892, 2024.

[23] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

[24] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024.

[25] Hanyu Lin, Min Cai, Jiawei Wen, and Haodi Zhang. Tyler: Typed latent reasoning for language models–when to think, what to compute, and how much to allocate. *arXiv preprint arXiv:2606.16360*, 2026.

[26] Guoxin Ma, Yibing Liu, Chengzhengxu Li, Yu Liang, Yan Wang, Yueyang Zhang, Kecheng Chen, Zhaohan Zhang, Zhiyuan Sun, and Daiting Shi. Thinking as compression: Your reasoning model is secretly a context compressor. *arXiv preprint arXiv:2605.28713*, 2026.

[27] Jason Wei, Najoung Kim, Yi Tay, and Quoc Le. Inverse scaling can become u-shaped. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15580–15591, 2023.

[28] Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68, 2024.

[29] Pengtao Ma, Ziliang Zhou, Ciyu Ruan, Haoyang Wang, Kaiyuan Li, Zihang Gong, Wenhua Ding, Chen Gao, Jingao Xu, and Xinlei Chen. Eventprune: Cascaded event-assisted token pruning for efficient first-person dynamic spatial reasoning. *arXiv preprint arXiv:2605.19506*, 2026.

[30] Dong Chen, Fei Gao, Shuo Zhang, Yueting Zhuang, Siliang Tang, Qidong Liu, Hua Wang, Xin Yang, and Mingliang Xu. Improving large models with small models: Lower costs and better performance. *Neural Networks*, page 108276, 2025.

[31] Shibei Meng, Binxin Yang, Yuan Liu, Jiexuan Zhang, Zhengyao Lv, Hubery Yin, and Qiang Xu. Versusq: Pairwise margin reasoning for generalizable video quality assessment. *arXiv preprint arXiv:2605.21130*, 2026.

[32] Chubin Chen, Sujie Hu, Jiashu Zhu, Meiqi Wu, Jintao Chen, Yanxun Li, Nisha Huang, Chengyu Fang, Jiahong Wu, Xiangxiang Chu, et al. Taming preference mode collapse via directional decoupling alignment in diffusion reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12775–12786, 2026.

[33] Jiajun Han, Guoxin Ma, Liangsheng Xu, Ruiqi Yu, Yingdi Guo, Zongbing Wang, Xinhua Ye, Chenyan Hua, Xin Xie, Pengcheng Yang, et al. Wegr: Unified generative retrieval for industrial search systems in wechat. *arXiv preprint arXiv:2610.16979*, 2026.

[34] Xingyilang Yin, Chengzhengxu Li, Jiahao Chang, Chi-Man Pun, and Xiaodong Cun. Mllm-4d: Towards visual-based spatial-temporal intelligence. *arXiv preprint arXiv:2603.00515*, 2026.

[35] Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 6342–6353, 2023.

[36] Ming Xia, Guoxin Ma, Zhiqin Zhou, and Dongmin Huang. Unleash the potential of long semantic ids for generative recommendation. *arXiv preprint arXiv:2602.13573*, 2026.

[37] Howard Yen, Tianyu Gao, and Danqi Chen. Long-context language modeling with parallel context encoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2588–2610, 2024.

[38] Zongqian Li, Yixuan Su, and Nigel Collier. 500xcompressor: Generalized prompt compression for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25081–25091, 2025.

[39] Ying Yang, Zhengyao Lv, Yujia Zeng, Tianlin Pan, Haofan Wang, Yueming Lyu, Binxin Yang, Hubery Yin, Chen Li, Jing Lyu, et al. Stableworld: Towards stable and consistent long interactive video generation. *arXiv preprint arXiv:2601.15281*, 2026.

[40] Sujie Hu, Huaisong Zhang, Pujing Yang, Qian Wang, Xiaomin Li, Xinhang Leng, Chen Li, Jing Lyu, Xiu Li, and Ling Pan. Blame errant words, not the canvas: Defect-aware spatial credit assignment for visual text generation. *arXiv preprint arXiv:2610.20794*, 2026.

[41] Chengzhengxu Li, Xiaoming Liu, Zhaohan Zhang, Shaochu Zhang, Shengchao Liu, Guoxin Ma, Yu Lan, and Chao Shen. Upfront chain-of-thought: A cooperative framework for chain-of-thought compression. *arXiv preprint arXiv:2510.08647*, 2025.

[42] Weronika Łajewska, Momchil Hardalov, Laura Aina, Neha Anna John, Hang Su, and Lluís Màrquez. Understanding and improving information preservation in prompt compression for llms. *arXiv preprint arXiv:2503.19114*, 2025.

[43] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[44] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[45] Jesse Mu, Xiang Li, and Noah Goodman. Learning to compress prompts with gist tokens. *Advances in Neural Information Processing Systems*, 36:19327–19352, 2023.

[46] Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. Adapting language models to compress contexts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3829–3846, 2023.

[47] David Rau, Shuai Wang, Hervé Déjean, and Stéphane Clinchant. Context embeddings for efficient answer generation in rag. *arXiv preprint arXiv:2407.09252*, 2024.

[48] Jun Gao, Qi Lv, Zili Wang, Tianxiang Wu, Ziqiang Cao, and Wenjie Li. Uniicl: An efficient icl framework unifying compression, selection, and generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 500–510, 2025.

[49] Sijun Tan, Xiuyu Li, Shishir G Patil, Ziyang Wu, Tianjun Zhang, Kurt Keutzer, Joseph E Gonzalez, and Raluca Ada Popa. Lloco: Learning long contexts offline. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17605–17621, 2024.

[50] Jun Gao, Ziqiang Cao, and Wenjie Li. Selfcp: Compressing over-limit prompt via the frozen large language model itself. *Information Processing & Management*, 61(6):103873, 2024.

[51] Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

[52] Maja Popović. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618, 2017.

[53] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[54] Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *Proceedings of ICLR*, 2024.

[55] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[56] Colin B Clement, Matthew Bierbaum, Kevin P O'Keeffe, and Alexander A Alemi. On the use of arxiv as a dataset. *arXiv preprint arXiv:1905.00075*, 2019.

[57] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 2022.

[58] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. 2021.

[59] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

[60] Chengzhengxu Li, Xiaoming Liu, Yichen Wang, Duyi Li, Yu Lan, and Chao Shen. Dialogue for prompting: A policy-gradient-based discrete prompt generation for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18481–18489, 2024.

[61] Chengzhengxu Li, Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Chen Liu, Yu Lan, and Chao Shen. Concentrate attention: Towards domain-generalizable prompt optimization for language models. *Advances in Neural Information Processing Systems*, 37:3391–3420, 2024.

A Limitations

Our study focuses on fidelity-critical context compression in a memory-token compressor–decoder setting. Although we evaluate two model families, multiple compression rates, decoder configurations, training domains, loss weights, and downstream tasks, the conclusions are still tied to this class of compression architectures and training objectives. Other compression paradigms, such as retrieval-based compression, extractive compression, or architectures with explicit copy mechanisms, may exhibit different scaling behavior.

Our rank and entropy analyses are intended as diagnostic probes of the compressed memory $\mathbf{Z}$, rather than a complete causal account of the paradox. Effective rank helps characterize how source information is organized in memory, while conditional entropy measures how tightly the memory constrains reconstruction under teacher forcing. These quantities reveal consistent associations with knowledge overwriting and semantic drift, but they do not by themselves specify a complete intervention for restoring fidelity in large compressors. The rank-intervention results further suggest that fidelity depends on the learned source-aligned structure of the memory, not only on scalar properties such as rank.

Finally, our diagnostic QA tasks are designed to stress source-specific fidelity, especially counterfactual facts and fine-grained semantic relations. This makes them sensitive to failures that standard reconstruction metrics overlook, but they are not meant to replace all downstream evaluations. Different applications may tolerate different levels of paraphrase, abstraction, or factual substitution. A practical compressor should therefore be selected and trained according to the fidelity requirements of the target use case. Developing large-compressor training methods that explicitly strengthen source anchoring while preserving efficiency remains an important direction for future work.

B Reproducibility Details

B.1 Training Hyperparameters

In distributed training, we set the global batch size to 512, computed as 2 examples per GPU $\times$ 4 gradient-accumulation steps $\times$ 16 GPUs. For the multimodal LLaMA-3.2-11B and 90B variants, we remove the vision encoders and retain only the text-based language model backbones, ensuring a consistent text-only setting across scales. Detailed pretraining parameters and experimental settings are summarized in Table 10.

B.2 Memory-Token Configuration

All experiments use 256-token compression segments. For a compression rate r, each segment is compressed into

$$m = 256/r$$

memory embeddings. Therefore, the $4\times$, $16\times$, and $64\times$ settings use 64, 16, and 4 memory embeddings per segment, respectively.

For each input segment, we append m learnable memory query tokens to the compressor input and take the final-layer hidden states at these memory-token positions as the compressed representation.

Table 10: Pretraining hyperparameters and distributed setup.

Hyperparameter	Setting
Optimizer	AdamW
Learning rate	1×10^{-4}
Learning-rate scheduler	Cosine
Global batch size	512
Warmup steps	300
Max gradient norm	0.5
Epochs	1
Precision	bfloat16
Trainable modules	Compressor and memory converter
Frozen modules	Decoder
Distributed setup	2×8 A100 GPUs
DeepSpeed strategy	ZeRO-3 for 32B compressors; ZeRO-1 for smaller compressors
Evaluation interval	Every 250 optimization steps
Checkpoint interval	Every 750 optimization steps
Model selection metric	Reconstruction validation loss

When a context contains multiple segments, each segment is compressed independently. The projected memory embeddings are then concatenated and wrapped with memory-boundary embeddings before being passed to the fixed decoder.

B.3 Dataset Preprocessing

We preprocess FineWeb [51] and ArxivCorpus [56] with the same pipeline. Raw documents are first tokenized with the compressor tokenizer without adding extra special tokens. We then recover the text span corresponding to each selected segment and tokenize it again with the decoder tokenizer, so that the compressor input and decoder supervision are aligned to their respective vocabularies.

Each pretraining example contains two aligned text spans: the current segment and the immediately following segment. The current segment is used for reconstruction supervision, while the following segment is used for continuation supervision. This construction supports the two pretraining objectives used in the main experiments: reconstructing the compressed text and predicting future context from the compressed prefix.

B.4 Chunking and Filtering Rules

Documents are chunked into consecutive 256-token segments according to the compressor tokenizer. We keep only documents with at least two full segments, since continuation supervision requires both a current segment and its following segment. Formally, a document with fewer than 512 compressor tokens is discarded.

For each remaining document with token sequence length L, we generate training pairs from adjacent chunks:

$$(C_j, C_{j+1}), \quad j = 0, \ldots, \lfloor L/256 \rfloor - 2.$$

Here C_j is the segment to be compressed and reconstructed, and C_{j+1} is the continuation target. Any trailing tokens that do not form a complete 256-token segment are discarded. The same chunking and filtering rules are used across all compressor scales and compression rates, ensuring that performance differences are not caused by different data construction procedures.

C Complete Quantitative Results

D Diagnostic QA Generation Prompts and Filtering Criteria

This section details the prompt engineering strategies used to elicit diagnostic samples, as well as the filtering procedures used to obtain the final datasets. The prompts are designed to strictly enforce the

Table 11: **Complete Quantitative Evaluation of the Size-Fidelity Paradox across Qwen and LLaMA Families (0.6B–90B).** We report surface reconstruction metrics and two diagnostic QA evaluations under varying compression rates. **Bold** marks the best model within each family/compression group. Shaded Avg./Δ columns show category averages and relative percentage drops from the group best; red marks the worst QA Avg./drop.

Compression Rate	Model	#Params	Reconstruction						*Knowledge Overwriting*				*Semantic Drift*			
			BLEU	ROUGE-L	chrF++	BERTScore	Avg.	Δ	FaithEval	ConflictQA	Avg.	Δ	FineWeb	FaithEval	Avg.	Δ
4×	Qwen 3	0.6b	0.95	0.97	0.98	0.997	0.97	(-1.02%)	0.71	**0.95**	0.83	(-1.19%)	0.83	**0.83**	**0.83**	——
		1.7b	0.94	0.96	0.98	0.995	0.97	(-2.04%)	**0.76**	0.89	0.83	(-1.19%)	0.83	0.81	0.82	(-1.20%)
		4b	0.94	0.96	0.98	0.996	0.97	(-2.04%)	0.72	**0.95**	**0.84**	——	**0.87**	0.79	**0.83**	——
		8b	0.94	0.96	0.97	0.995	0.97	(-2.04%)	0.74	0.91	0.83	(-1.19%)	0.79	**0.83**	0.81	(-2.41%)
		32b	**0.97**	**0.98**	**0.99**	**0.998**	**0.98**	——	0.68	0.80	0.74	(-11.90%)	0.60	0.78	0.69	(-16.87%)
	LLaMA 3.2	1b	0.94	0.96	0.97	0.995	0.97	(-2.02%)	**0.77**	0.88	**0.83**	——	0.78	**0.85**	**0.82**	——
		3b	0.90	0.93	0.95	0.992	0.94	(-4.04%)	0.74	**0.91**	**0.83**	——	**0.79**	0.84	**0.82**	——
		11b	0.92	0.94	0.97	0.993	0.96	(-3.03%)	0.65	0.80	0.73	(-12.05%)	0.64	0.76	0.70	(-14.63%)
		90b	**0.97**	**0.99**	**0.99**	**0.999**	**0.99**	——	0.60	0.81	0.71	(-14.46%)	0.59	0.68	0.64	(-21.95%)
16×	Qwen 3	0.6b	0.83	0.86	0.89	0.986	0.89	(-3.26%)	0.70	**0.94**	0.82	(-1.20%)	0.75	0.84	0.80	(-3.61%)
		1.7b	**0.86**	0.89	0.93	0.988	0.92	(-1.09%)	**0.75**	0.88	0.82	(-1.20%)	0.79	**0.86**	**0.83**	——
		4b	**0.86**	**0.90**	**0.95**	0.988	**0.92**	——	0.71	**0.94**	**0.83**	——	**0.82**	0.83	**0.83**	——
		8b	0.84	0.87	0.90	0.986	0.90	(-3.26%)	0.72	0.89	0.81	(-2.41%)	0.73	0.79	0.76	(-8.43%)
		32b	0.85	0.87	0.92	**0.991**	0.91	(-2.17%)	0.61	0.72	0.67	(-19.28%)	0.57	0.74	0.66	(-20.48%)
	LLaMA 3.2	1b	0.82	0.85	0.89	0.983	0.89	(-5.38%)	0.72	0.84	0.78	(-4.88%)	**0.75**	**0.90**	**0.83**	——
		3b	0.81	0.85	0.88	0.982	0.88	(-5.38%)	**0.73**	**0.90**	**0.82**	——	**0.75**	**0.90**	**0.83**	——
		11b	0.85	0.89	0.92	0.986	0.91	(-2.15%)	0.63	0.77	0.70	(-14.63%)	0.61	0.73	0.67	(-19.28%)
		90b	**0.86**	**0.91**	**0.96**	**0.993**	**0.93**	——	0.55	0.74	0.65	(-20.73%)	0.56	0.63	0.60	(-27.71%)
64×	Qwen 3	0.6b	0.17	0.31	0.44	0.872	0.45	(-22.41%)	0.44	0.59	0.52	(-17.74%)	0.31	0.39	0.35	(-20.45%)
		1.7b	0.21	0.39	0.49	0.880	0.49	(-13.79%)	0.51	0.60	0.56	(-11.29%)	0.35	0.42	0.39	(-11.36%)
		4b	0.27	0.44	0.58	0.894	0.55	(-5.17%)	**0.54**	**0.70**	**0.62**	——	**0.41**	0.46	**0.44**	——
		8b	0.28	0.46	**0.59**	0.925	0.56	(-1.72%)	**0.54**	0.67	0.61	(-3.23%)	0.38	**0.49**	**0.44**	——
		32b	**0.29**	**0.49**	**0.59**	**0.930**	**0.58**	——	0.47	0.56	0.52	(-17.74%)	0.27	0.35	0.31	(-29.55%)
	LLaMA 3.2	1b	0.19	0.30	0.38	0.878	0.44	(-30.65%)	0.49	0.57	0.53	(-14.52%)	0.33	0.44	0.39	(-7.32%)
		3b	0.29	0.50	0.62	0.919	0.58	(-6.45%)	**0.55**	**0.68**	**0.62**	——	**0.35**	**0.47**	**0.41**	——
		11b	0.30	0.53	0.63	0.927	0.60	(-4.84%)	0.51	0.63	0.57	(-8.06%)	0.24	0.33	0.29	(-31.71%)
		90b	**0.33**	**0.55**	**0.67**	**0.944**	**0.62**	——	0.45	0.61	0.53	(-14.52%)	0.24	0.31	0.28	(-34.15%)

generation of questions that probe the model's reliance on parametric priors versus faithful context reconstruction. Task-specific filters further remove overly easy samples, focusing the evaluation on cases where source fidelity is more likely to be challenged.

D.1 Semantic Drift QA Generation Prompt

To quantify *Semantic Drift*, we require questions that demand precise structural preservation rather than mere topical understanding. The prompt for this task directs the offline model to generate questions focusing on fine-grained semantic dimensions—such as role binding, modifier scope, and predicate exactness. Crucially, the prompt imposes a constraint that the ground-truth answer must be an exact substring of the original text, ensuring that any subtle paraphrasing or relational distortion in the reconstructed context is penalized.

> ***Semantic Drift QA Generation***
> *You are an expert dataset writer building a QA suite to detect TOPIC/ENTITY DRIFT after compression+reconstruction.*
>
> ***GOAL***
> *Generate questions that will fail if the reconstructed text:*
> *- shifts the main topic, scope or subject,*
> *- replaces specific entities with generic categories (over-generalization),*
> *- changes the stated actions or relationships, such as rewriting key verbs/relations in a way that changes meaning (relationship drift),*
> *- drops list items or swaps roles/relationships (who did what to whom),*
> *- breaks cross-sentence relations (cause/effect, contrast, concession, example, and progression are weakened, reversed, or rewritten).*
>
> ***IMPORTANT:*** *We are NOT adding distractor sentences. We want high-coverage,*

high-precision QA over the given CONTEXT.

HARD RULES
- Output EXACTLY 20 QA items as JSON only (no extra text).
- 15 items must be answerable from CONTEXT.
- For answerable items, the answer MUST be an exact substring from CONTEXT (verbatim, case/punctuation preserved).
- Provide evidence_span: a short quote from CONTEXT that contains the answer verbatim.
- Answers must be SPECIFIC: do NOT use pronouns-only answers (it/they/this/that/he/she). Prefer named entities or concrete noun phrases.
- 5 items must be unanswerable but plausible and closely related (to detect drift-to-generic or invented details).
- For these, set answer="UNANSWERABLE" and evidence_span="".

COVERAGE & DIVERSITY REQUIREMENTS (for the 15 answerable)
- Make sure answers cover the entire context:
- ≥4 answers from the first third of CONTEXT,
- ≥4 from the middle third,
- ≥4 from the last third.
- Include the following question types (minimum counts):
1) main_topic (≥1): asks for the central subject/entity or the discussed domain/category.
2) entity_list (≥2): asks for lists or multiple items mentioned (answer is the exact list span).
3) predicate_exact (≥1): asks for the EXACT verb/relational phrase used.
- The gold answer must include the predicate words, not just an entity.
4) relation_anchor (≥1): asks about explicit relations: reason, contrast, cause/effect, "instead/however", "to do X, they do Y".
- The answer should be the specific clause/phrase that states the relation (short span).
5) coreference_link (≥2): asks what a pronoun/this/that/these refers to, answered by the antecedent noun phrase in CONTEXT.
6) role_binding (≥2): asks who performs an action or what is affected (forces correct role assignment).
7) modifier_scope (≥2): asks for extracting key modifiers that often get generalized away (category qualifiers, degree/adverb, constraints).

QUALITY CONSTRAINTS
- Keep each question unambiguous with a single correct answer span.
- Avoid trivial questions that can be answered by many spans (e.g., "What is mentioned?").
- Avoid yes/no questions unless the answer span is a clear "yes/no" phrase in CONTEXT (rare).

OUTPUT JSON FORMAT
```
{
  "qas": [
    {
      "id": "T01",
      "type": "answerable" | "unanswerable",
      "probe": "main_topic" | "entity_list" | "predicate_exact" | "relation_anchor" | "coreference_link" | "role_binding" | "modifier_scope",
      "question": "...",
      "answer": "...",
      "evidence_span": "..."
    }
  ]
}
```

CONTEXT
{CONTEXT}

SELF-CHECK (must satisfy before output)
- For every answerable item: answer string appears exactly in CONTEXT.
- For every unanswerable item: CONTEXT does not explicitly contain the answer.
- Ensure all probe-type minimum counts are met.

D.2 Knowledge Overwriting QA Generation Prompt (optional)

The primary objective of the *Knowledge Overwriting* task is to evaluate whether the compressor retains source-specific counterfactuals or succumbs to its own internal world knowledge. Consequently, the prompt instructs the model to identify conflict-inducing facts within the text (e.g., a counterfactual birth location or an inverted causal relationship) and generate questions where the correct answer must be derived exclusively from the provided context, directly contradicting common priors.

Knowledge Overwriting QA Generation
You are a high-precision question writer for measuring whether a reconstructed text stays faithful to the given context even when the context contradicts common knowledge.

TASK
Given CONTEXT, generate QA items that:
- Anchor on statements that may conflict with typical world knowledge, stereotypes, or common textbook facts.
- The gold answers must follow CONTEXT (contextual truth), not real-world truth.

RULES
- Write questions in the same language as CONTEXT.
- Produce exactly 12 QA items in JSON (no extra text).
- 9 items must be answerable from CONTEXT with EXACT-SUBSTRING answers (verbatim).
- 3 items must be unanswerable (plausible "world-knowledge" questions that readers might assume, but CONTEXT does not state).
- Among the 9 answerable items:
- At least 5 must be marked "world_conflict": true
** world_conflict means: a typical educated reader might expect a different answer from general knowledge,*
BUT the answer is explicitly given in CONTEXT.
- At least 3 must include named entities / numbers / dates if present.
- At least 3 must test causal or procedural relations (X leads to Y) to detect "corrective rewriting".
- For each answerable item, provide:
- answer (verbatim substring)
- evidence_span (short quote containing answer)
- world_conflict (true/false)
- conflict_note: 1 short sentence explaining what common belief it might contradict (do NOT correct it).

OUTPUT FORMAT (JSON)

```
{
  "qas": [
    {
      "id": "K1",
      "type": "answerable" | "unanswerable",
      "world_conflict": true | false,
      "conflict_note": "...",
      "question": "...",
      "answer": "...",
      "evidence_span": "..."
    }
```

```
    ]
}

CONTEXT
{CONTEXT}

SELF-CHECK
- Do not "fix" the context. Gold answers must reflect CONTEXT exactly.
- For answerable: answer must be an exact substring in CONTEXT.
- For unanswerable: ensure CONTEXT does not contain the answer.
```

D.3 Filtering Criteria for Diagnostic QA

We apply task-specific filtering to remove overly easy samples and focus the evaluation on cases that better expose fidelity failures. Filtering is performed with Qwen3-0.6B, 1.7B, 4B, 8B, and 32B compressors at 16× compression rate. Each source context is compressed and reconstructed with the fixed decoder, and Exact Match (EM) is computed against the source-consistent gold answer after standard normalization.

Knowledge Overwriting QA. Each context corresponds to one diagnostic triple $(q, a^{\text{gold}}, \mathcal{A}^{\text{alt}})$, where a^{gold} is the source-consistent answer and $\mathcal{A}^{\text{alt}}$ contains plausible prior-based alternatives. We discard a candidate if all five Qwen3 compressors achieve EM > 0.8. This removes cases that are faithfully recovered across scales and retains samples more likely to reveal knowledge overwriting.

Semantic Drift QA. Each context is paired with 20 QA pairs covering the semantic dimensions in Sec. 4.3. For each compressor, we average EM over the 20 QA pairs for that context, and discard the context if all five compressors achieve average EM > 0.8. This filters out structurally easy contexts and keeps examples that better test fine-grained relational preservation.

E Additional Decoder Compatibility with Large Decoders

Table 12: Additional decoder compatibility with large decoders. We evaluate Qwen3 compressors at 16× compression. "Recons." denotes reconstruction accuracy on FineWeb. Numbers in parentheses indicate changes relative to the decoder configuration specified in the row label.

Compressor	Decoder	Reconstruction	QA(i): *Knowledge Overwriting*		QA(ii): *Semantic Drift*	
		FineWeb	FaithEval	ConflictQA	FineWeb	FaithEval
Qwen3-4B	Qwen3-4B	**0.79**	**0.66**	**0.81**	**0.63**	**0.77**
Qwen3-32B	Qwen3-4B (*reference*)	0.80	0.59	0.69	0.52	0.68
Qwen3-32B	LLaMA-3.2-8B (vs. 4B)	0.85 (+0.05)	0.61 (+0.02)	0.72 (+0.03)	0.57 (+0.05)	0.74 (+0.06)
Qwen3-32B	Qwen3-32B (vs. 8B)	**0.86** (+0.01)	0.63 (+0.02)	0.74 (+0.02)	0.55 (-0.02)	0.70 (-0.04)

A possible concern is that the fidelity degradation of very large compressors may come from a mismatch between a high-capacity compressor and a weaker decoder. To test this, we pair Qwen3-32B with stronger decoders at 16× compression, including LLaMA-3.2-8B and a matched Qwen3-32B decoder. As shown in Tab. 12, increasing decoder capacity improves reconstruction and knowledge-overwriting accuracy to some extent. However, Qwen3-32B paired with Qwen3-32B still remains clearly below the Qwen3-4B compressor with a Qwen3-4B decoder on all fidelity-oriented QA metrics. Moreover, increasing the decoder from LLaMA-3.2-8B to Qwen3-32B does not improve semantic-drift QA and slightly decreases it on both FineWeb and FaithEval. These results suggest that stronger decoders can smooth or improve reconstruction, but they do not eliminate the fidelity degradation of very large compressors.

Table 13: **Per-dimension semantic-drift QA results on FineWeb and FaithEval.** Each block corresponds to one dataset. The row #QA pairs reports the number of generated QA pairs for each semantic-drift dimension and the total number across all dimensions; the remaining rows report QA accuracy (%) for each dimension and the corresponding overall micro-average.

Compress. Rate	Model	#Params	Per-dimension ***Semantic-Drift*** QA results							
			Main topic	Entity list	Predicate exact	Relation anchor	Coreference link	Role binding	Modifier scope	**Average**
(i) Dataset: FineWeb										
	#QA pairs		1029	1447	1091	1104	1741	1180	1388	8980(all)
4×	Qwen 3	0.6b	0.7839	**0.9031**	0.9105	0.7399	0.7624	0.8412	0.8787	0.8311
		1.7b	**0.7968**	0.9021	0.8587	0.7734	0.7726	0.8496	0.8838	0.8341
		4b	0.8649	0.8970	**0.9294**	**0.8074**	**0.8282**	**0.8729**	**0.9039**	**0.8708**
		8b	0.7594	0.7794	0.8053	0.7284	0.7445	0.8651	0.8630	0.7914
		32b	0.6050	0.6603	0.6165	0.5353	0.5444	0.6569	0.6114	0.6028
	LLaMA 3.2	1b	**0.7925**	0.7333	0.8375	0.6954	0.7383	0.8137	**0.8561**	0.7786
		3b	0.7500	**0.8299**	**0.8537**	**0.7168**	**0.7463**	**0.8667**	0.7784	**0.7904**
		11b	0.6150	0.6734	0.6331	0.5982	0.6019	0.6805	0.7010	0.6439
		90b	0.5713	0.6355	0.5884	0.5283	0.5410	0.6224	0.6381	0.5896
16×	Qwen 3	0.6b	0.7140	0.8283	0.7721	0.6757	0.7135	0.7618	0.7926	0.7531
		1.7b	0.8095	**0.8674**	**0.8665**	0.7327	0.7523	0.7964	0.7344	0.7919
		4b	**0.8226**	0.8409	0.7943	**0.7432**	**0.7839**	**0.8653**	**0.9012**	**0.8226**
		8b	0.7397	0.7984	0.7725	0.6750	0.6649	0.7274	0.7379	0.7288
		32b	0.5954	0.5709	0.5953	0.5142	0.5429	0.6136	0.5917	0.5731
	LLaMA 3.2	1b	0.7163	0.8058	0.7312	**0.6857**	**0.7082**	0.7563	**0.8143**	0.7476
		3b	**0.7495**	**0.8211**	**0.7723**	0.6710	0.7035	**0.7684**	0.7565	**0.7488**
		11b	0.5710	0.6658	0.6465	0.5497	0.5595	0.5989	0.6611	0.6082
		90b	0.5701	0.5963	0.5783	0.5211	0.5168	0.5824	0.6162	0.5677
64×	Qwen 3	0.6b	0.2998	0.3436	0.3423	0.2823	0.2988	0.3441	0.2908	0.3141
		1.7b	0.3305	0.3543	0.3749	0.3170	0.3288	0.3643	0.3804	0.3499
		4b	**0.4079**	0.4016	**0.4297**	**0.3856**	**0.3764**	**0.4487**	**0.4529**	**0.4130**
		8b	0.3768	**0.4086**	0.3881	0.3485	0.3524	0.4109	0.3511	0.3756
		32b	0.2728	0.2929	0.2861	0.2491	0.2467	0.2849	0.2911	0.2741
	LLaMA 3.2	1b	0.3251	0.3165	0.3638	0.3087	0.3144	**0.3610**	0.3436	0.3319
		3b	**0.3534**	**0.3559**	**0.3818**	**0.3250**	**0.3298**	0.3421	**0.3624**	**0.3491**
		11b	0.2484	0.2556	0.2658	0.2183	0.2229	0.2334	0.2674	0.2440
		32b	0.2479	0.2503	0.2486	0.2205	0.2228	0.2458	0.2404	0.2387
(ii) Dataset: FaithEval										
	#QA pairs		1207	1973	1363	1495	2005	2019	1898	11960(all)
4×	Qwen 3	0.6b	0.7855	**0.9046**	**0.9120**	0.7414	0.7639	0.8427	**0.8656**	**0.8328**
		1.7b	0.7902	0.8213	0.8786	0.7457	0.7363	**0.8826**	0.8248	0.8119
		4b	0.7929	0.8221	0.7992	0.7265	0.7437	0.7558	0.8558	0.7856
		8b	**0.8385**	0.8373	0.8578	**0.7744**	**0.7727**	0.8492	0.8570	0.8262
		32b	0.7773	0.8351	0.8556	0.6962	0.7306	0.7458	0.8285	0.7806
	LLaMA 3.2	1b	0.8452	**0.9084**	**0.8809**	**0.7955**	0.7757	**0.8633**	0.9051	**0.8544**
		3b	**0.8509**	0.8998	0.8013	0.7881	**0.7962**	0.8214	**0.9233**	0.8428
		11b	0.7463	0.7645	0.8236	0.7163	0.7199	0.8180	0.7593	0.7641
		90b	0.6475	0.7041	0.6638	0.6292	0.6331	0.7112	0.7361	0.6788
16×	Qwen 3	0.6b	**0.8736**	**0.9060**	0.8498	0.7541	0.7924	0.9251	0.7792	0.8414
		1.7b	0.8281	0.9048	**0.9292**	**0.8080**	**0.8003**	**0.9362**	0.8209	**0.8622**
		4b	0.7894	0.8595	0.8647	0.7417	0.7923	0.8972	**0.8690**	0.8349
		8b	0.8114	0.8143	0.8159	0.7228	0.7197	0.8336	0.7850	0.7855
		32b	0.7359	0.7687	0.8080	0.6933	0.6782	0.7009	0.8082	0.7401
	LLaMA 3.2	1b	**0.9285**	0.8395	**0.9387**	0.8417	0.8462	0.9322	**0.9647**	0.8967
		3b	0.8502	**0.9518**	0.8850	**0.8448**	**0.8471**	**0.9621**	0.9414	**0.9031**
		11b	0.7478	0.7839	0.7571	0.6882	0.6657	0.7093	0.7906	0.7339
		90b	0.6341	0.6750	0.5954	0.5822	0.5946	0.6454	0.6799	0.6325
64×	Qwen 3	0.6b	0.4038	0.4183	0.3918	0.3623	0.3587	0.3776	0.4205	0.3903
		1.7b	0.3977	0.4363	0.4513	0.3888	0.3783	0.4491	0.4287	0.4194
		4b	0.4659	0.4744	0.4664	0.4258	0.4318	0.4589	**0.5084**	0.4622
		8b	**0.4671**	**0.5323**	**0.5400**	**0.4482**	**0.4661**	**0.5342**	0.4447	**0.4914**
		32b	0.3419	0.3396	0.3693	0.3263	0.3229	0.3738	0.3697	0.3493
	LLaMA 3.2	1b	0.4147	**0.4621**	0.4604	0.4125	0.3963	0.4582	0.4688	0.4403
		3b	**0.4766**	0.4490	**0.4634**	**0.4254**	**0.4437**	**0.5148**	**0.4988**	**0.4686**
		11b	0.3288	0.3388	0.3310	0.3012	0.2963	0.3558	0.3524	0.3301
		90b	0.3043	0.3325	0.3453	0.2932	0.2929	0.3366	0.2937	0.3141

F Per-Dimension Results of Semantic Drift QA

G Additional Analysis of Memory Geometry

G.1 Effective Rank and Hidden Dimension

We further examine whether the increase in effective rank is merely a consequence of larger hidden dimensions. For this analysis, memory tokens are stacked as observations and the effective rank is computed over the hidden dimension. As shown in Tab. 14, effective rank increases with scale, but the normalized ratio erank/d decreases. Thus, larger compressors do not simply occupy the

Table 14: Effective rank normalized by hidden size.

Model	Hidden size d	eRank	eRank/d
Qwen3-0.6B	1024	856.8	0.8455
Qwen3-1.7B	2048	1032.8	0.5043
Qwen3-4B	2560	1067.0	0.4168
Qwen3-8B	4096	1134.4	0.2770
Qwen3-32B	5120	1169.3	0.2284
LLaMA-3.2-1B	2048	975.2	0.4762
LLaMA-3.2-3B	3072	1072.6	0.3492
LLaMA-3.2-11B	4096	1149.3	0.2810
LLaMA-3.2-90B	8192	1194.4	0.1458

Table 15: Rank intervention on Qwen3-8B memory embeddings at 16× compression.

Compressor / Intervention	Reconstruction	QA(i): *Knowledge Overwriting*	QA(ii): *Semantic Drift*
	FineWeb	ConflictQA	FineWeb
Qwen3-4B	**0.8622**	**0.9414**	**0.8226**
Qwen3-8B *(reference)*	0.8419	0.8931	0.7288
Qwen3-8B *(90%)*	0.8493 (+0.0074)	0.9104 (+0.0173)	0.7452 (+0.0164)
Qwen3-8B *(75%)*	0.7933 (-0.0486)	0.7802 (-0.1129)	0.6685 (-0.0603)
Qwen3-8B *(50%)*	0.7276 (-0.1143)	0.7313 (-0.1618)	0.5841 (-0.1447)

expanded hidden space uniformly. Instead, they form broader yet still relatively low-dimensional memory structures, consistent with the view that scaling changes how source information is organized rather than only increasing representational size.

G.2 Rank Intervention and Spectral Structure

To further probe the role of memory geometry, we apply post-hoc truncated-SVD projection to Qwen3-8B memory embeddings at 16× compression. Tab. 15 and Fig. 5-6 shows a non-monotonic pattern. Moderate truncation improves reconstruction and both fidelity-oriented QA scores, while stronger truncation sharply degrades them. This suggests that part of the spectral tail corresponds to diffuse semantic variation, but the leading subspace still contains source-bearing evidence that must be preserved.

H Additional Analysis of Reconstruction Entropy

H.1 A Bottleneck View of Conditional Entropy

A principled way to interpret the entropy trend is to view context compression as source transmission under a fixed bottleneck. Let X denote the source, $Z = f_\theta(X)$ the compressed memory, and $S = h(X)$ a coarse semantic abstraction of the source. Since S is determined by X,

$$I(X;Z) = I(S;Z) + I(X;Z \mid S).$$

Under a fixed memory budget, increasing the amount of semantic information preserved in Z does not guarantee that source-specific residual details are equally preserved. The remaining uncertainty satisfies

$$H(X \mid Z) = H(S \mid Z) + H(X \mid S, Z), \quad H(X \mid S, Z) = H(X \mid S) - I(X;Z \mid S).$$

Thus, when Z captures broad semantic content but under-specifies source-specific details, multiple reconstructions can remain compatible with the same memory. The teacher-forced conditional entropy used in Sec. 6.2 provides a token-level probe of this ambiguity. Higher entropy indicates that the memory leaves more plausible continuations available, which is precisely the condition under which fine-grained semantic drift becomes more likely.

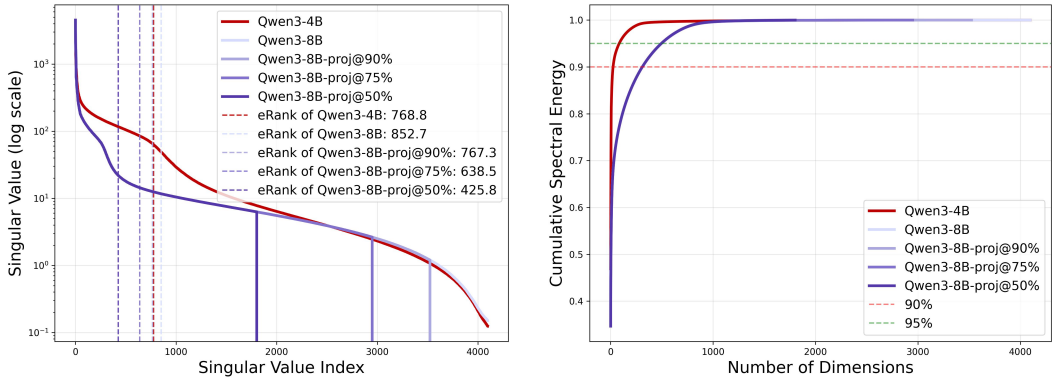


Figure 5: **Spectral evidence for a source-sufficient regime.** Mild truncation removes low-energy tail directions, whereas stronger truncation cuts into source-bearing spectral regions. **(a) Singular-value spectra.** The natural 8B compressor has a flatter, longer-tailed spectrum than the natural 4B model. A 90% projection mainly trims the tail, while 75%/50% projections cut into the shoulder near the spectral elbow. **(b) Cumulative spectral energy.** The 90% projection preserves most leading spectral energy, whereas stronger projections discard increasingly larger portions of the spectrum, explaining the transition from improved fidelity to source-signal loss.

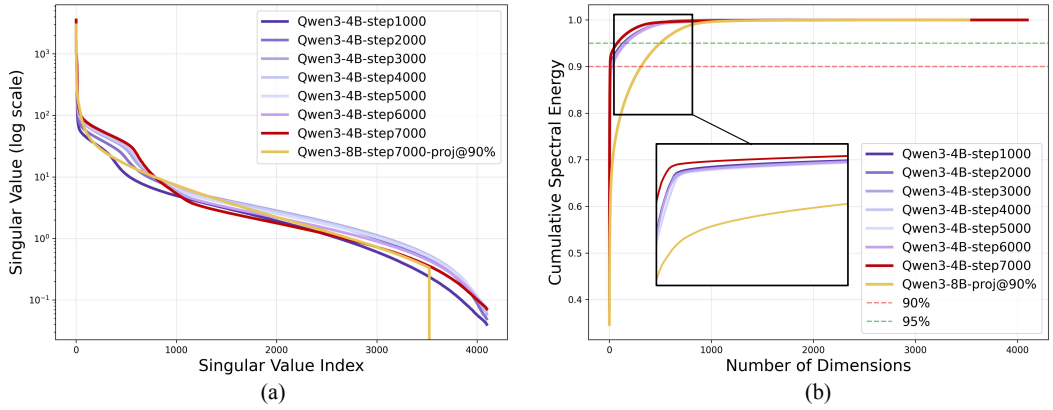
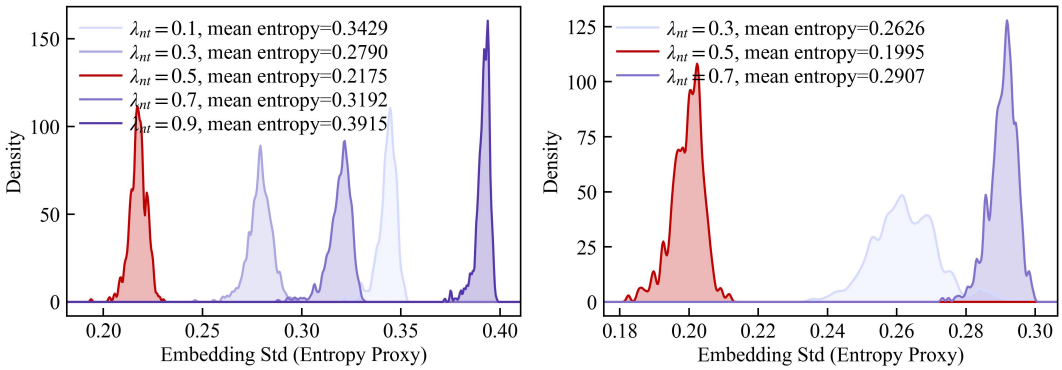


Figure 6: **Training forms the source-aligned spectral structure. (a) Singular-value spectra across 4B training.** As training proceeds, the 4B compressor evolves toward a smoother and more stable low-rank spectrum, suggesting that the source-sufficient structure is formed during optimization rather than imposed afterward. The 90%-projected 8B approaches the late-stage 4B in spectral concentration, but remains a truncation of a different spectrum. **(b) Cumulative spectral energy across 4B training.** The 4B model progressively concentrates more spectral energy into a compact leading subspace, consistent with convergence toward a source-aligned memory structure. Thus, similar $eRank$ can reflect similar spectral concentration without guaranteeing the same represented subspace or reconstruction fidelity.

H.2 Entropy under Loss-Weight Sweep

We also examine conditional entropy under the loss-weight sweep in Sec. 5.1.3. Fig. 7 shows the entropy distributions for Qwen3-0.6B and Qwen3-4B at $4\times$ compression. In both cases, entropy is lowest near the balanced setting $\lambda = 0.5$, where reconstruction and next-token prediction jointly shape the memory. Moving away from this balance broadens the conditional distribution, suggesting that reconstruction determinacy depends on how the objective allocates capacity between source recovery and general language modeling.



(a) Compressor: Qwen3-0.6b

(b) Compressor: Qwen3-4b

Figure 7: **Entropy distributions under λ-sweep.** Shown are the embedding-entropy distributions of compressors trained with different λ values ($\lambda L_{nt} + (1-\lambda)L_{re}$) for Qwen3-0.6B(a) and Qwen3-4B(b) at 4×; in both cases, the mean entropy is lowest around $\lambda = 0.5$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Sec. 1 for the main claims and Sec. 4 for the detailed contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix A for the discussion on the limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: See Sec. 3 for assumptions and proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Appendix B for reproducibility details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We include our data in the supplementary material, covering the contexts and QA pairs for both *knowledge overwriting* and *semantic drift*. The code follows the PCC compressor–decoder paradigm [16], with reproducibility instructions provided in Appendix B.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We present our experiment settings and implementation details in Sec. 4.1 and Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide the the statistical significance of the experiments in Sec. 5.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We introduce the compute resources need for our work in Appendix B.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impacts in Sec. 5.2.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We describe the safeguards in Sec. 4.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly credited and cited licenses of existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provided proper documents for our released data.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not include crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: This study is based on large language models, and the use of such models is discussed in detail in Sec. 4.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.