BINDING BREAKS FIRST: DIAGNOSING CONTEXTUAL HALLUCINATIONS UNDER KV-CACHE COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Key–value (KV) cache compression reduces the memory cost of long-context inference, but can also degrade context-dependent reasoning and increase hallucinations that contradict the provided context. Existing evaluations mainly measure how much performance drops, leaving unclear what information is actually lost or corrupted by compression. We propose BINDPROBE, a diagnostic framework that separates two possible failure modes: *item-access failure*, where an entity or attribute value can no longer be recovered, and *binding failure*, where both remain accessible but the model can no longer determine which value belongs to which entity. BINDPROBE places controlled entity–value associations at different positions in long contexts and evaluates them with separate entity, value, and binding probes. Across several evaluated compression settings, we observe regimes in which relational binding degrades substantially while access to the individual items remains comparatively stable. This produces *item-preserving misbinding*: the model retains the relevant entities and values but associates them incorrectly. The effect is strongly position dependent, with associations in the middle of the context becoming vulnerable at lower compression rates. Mechanistic analyses further show that KV states important for recovering these associations are systematically assigned lower importance by compression criteria, making them more likely to be removed or distorted as the cache budget shrinks. Our results show that compression-induced contextual hallucinations can arise not only because information is forgotten, but also because the relationships among still-accessible information are disrupted.

1 INTRODUCTION

Long-context inference is increasingly constrained by the key–value (KV) cache, whose memory footprint grows linearly with sequence length. This has motivated a broad class of compression methods based on eviction, quantization, and merging (Yuan et al., 2024; Yang et al., 2024). However, compression is not lossless: recent studies show that it can degrade context-dependent reasoning and increase contextual hallucinations and inconsistencies (Liu et al., 2026; 2025a). These methods can substantially reduce inference cost, but the resulting information loss is not benign: recent studies report degraded context-dependent reasoning and increased contextual hallucinations under KV-cache compression (Liu et al., 2026; 2025a). Yet such degradation is typically characterized by benchmark-level evaluations, leaving unclear what information failures cause KV-cache compression to increase contextual hallucinations.

To close this gap, we propose BINDPROBE, a diagnostic framework that decomposes compression-induced contextual hallucinations into two information-level failure modes: *item-access failure*, where a relevant entity or attribute value becomes inaccessible, and *binding failure*, where the constituent items remain accessible but their association is no longer correctly recovered. To distinguish the two, BINDPROBE places entity–value pairs at controlled positions in long contexts and evaluates each pair with three complementary probes: entity and value probes test access to the individual constituents, while binding probes test their correct association. By evaluating these probes on the same underlying associations, BINDPROBE isolates binding failures from errors caused by simple information loss.

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

Across the evaluated compression methods and budgets, we observe a consistent divergence: binding performance degrades substantially faster than item accessibility as compression becomes more aggressive. In particular, models increasingly produce incorrect entity–value associations even when the corresponding entities and attribute values remain accessible. We term this failure mode *item-preserving misbinding*. The effect is strongly position dependent, with middle-context bindings becoming vulnerable at lower compression rates. Our mechanistic analyses further indicate that compression disproportionately disrupts KV states that are causally important for relation recovery: these binding-critical states are systematically underestimated by the compression criteria, making them more vulnerable to removal or distortion as the cache budget tightens. Together, these findings show that compression-induced contextual hallucinations can arise not only from the loss of contextual items, but also from the breakdown of associations among items that remain accessible.

Our contributions are threefold: (1) We introduce BINDPROBE to diagnose compression-induced contextual hallucinations by distinguishing item-access failures from binding failures. (2) We show that binding can degrade before item access as compression intensifies, producing *item-preserving misbinding* before out-of-context predictions become prevalent. (3) We identify KV states that causally support relation recovery and provide evidence that compression importance scores can underestimate their functional contributions.

2 RELATED WORK

KV-cache compression.

A rich line of work has explored KV-cache compression along three major directions: token eviction, which selectively retains important tokens while discarding the rest; quantization, which stores KV states at reduced precision; and merging, which combines similar cache entries to reduce redundancy. Early eviction methods typically rely on recency (Xiao et al., 2024) or attention-based importance (Li et al., 2024), while subsequent works have introduced more refined importance indicators (Devoto et al., 2025) and budget allocation strategies (Cai et al., 2025) to better preserve critical information under aggressive compression. Beyond token eviction, mixed-precision methods retain less important tokens at lower precision rather than discarding them entirely (Yang et al., 2024), while staged compression schemes further distinguish the prefilling and decoding stages to account for their different cache requirements (Zirui Liu et al., 2023). Together, these approaches have substantially expanded the design space of KV-cache compression, enabling increasingly aggressive reduction of cache footprints while attempting to preserve downstream model capabilities.

KV-cache compression and hallucination.

Recent work has systematically examined how KV-cache compression affects different model capabilities across tasks, models, and compression schemes. Liu et al. (2026) evaluated six categories of model abilities under KV-cache compression and showed that arithmetic reasoning, in-context learning (ICL), and chain-of-thought (CoT) reasoning are particularly susceptible to compression. Li et al. (2025) further extended the evaluation to multi-task, multi-turn, multi-request, and retrieval settings, revealing distinct performance patterns associated with different attention structures under compression. At a broader scale, Yuan et al. (2024) conducted comprehensive evaluations across a wide range of benchmarks and model settings, systematically comparing the effects of different compression approaches, while Yang et al. (2024) investigated the impact of KV-cache quantization across diverse models and experimental configurations. Together, these studies establish that the impact of KV-cache compression is highly capability- and setting-dependent. However, existing evaluations predominantly characterize compression-induced degradation through task-level performance, leaving its relationship with hallucination comparatively underexplored. In particular, it remains unclear whether compression-induced failures primarily reflect a general loss of model capability or the disruption of specific mechanisms responsible for faithful contextual generation.

Mechanisms of contextual retrieval and binding.

Prior work has begun to characterize the internal mechanisms that support contextual retrieval and binding. Wu et al. (2024) identified *retrieval heads* that contribute disproportionately to retrieving information from context, while Feng & Steinhardt (2024) studied how language models bind entities to their attributes. More recently, Gur-Arieh et al. (2026) proposed that in-context binding relies

jointly on positional, lexical, and reflexive information. These studies reveal structured mechanisms for recovering contextual associations, but do not examine how they are affected by KV-cache compression. Our work connects these mechanistic analyses with cache compression by separating item access from relational binding and showing that compression can disrupt the latter even when the constituent items remain accessible.

3 DIAGNOSTIC FRAMEWORK

BINDPROBE is designed to distinguish two information-level failure modes under KV-cache compression. An *item-access failure* occurs when a relevant entity or attribute value can no longer be recognized from the compressed context, whereas a *binding failure* occurs when the constituent items remain accessible but their association is no longer correctly recovered. The central design of BINDPROBE is therefore to evaluate these abilities on the same underlying fact: for each entity–attribute–value triple (e, k, v), we separately probe access to the entity e, access to the value v, and recovery of the relation $(e, k) \mapsto v$. This paired construction allows us to determine whether compression-induced binding errors can be explained by loss of their constituent items, or instead arise from corruption of the association between items that remain accessible.

3.1 CONTROLLED CONTEXTS

We construct T independently sampled contexts $\{\mathcal{C}^{(t)}\}_{t=1}^{T}$ with controlled entity–attribute associations. Each context contains N entities

$$\mathcal{E}^{(t)} = \{e_1, \ldots, e_N\}$$

the attribute set K_{e_i} for each entity e_i is randomly sampled from a global candidate pool $\mathcal{U}_K$

$$\mathcal{K}_{e_i} = \{k_1, \ldots, k_K\}.$$

Every entity is assigned exactly one value for every attribute type, giving the ground-truth fact set

$$\mathcal{B}^{(t)} = \{(e_i, k, v_{ik}) : e_i \in \mathcal{E}^{(t)}, k \in \mathcal{K}_{e_i}\}. \tag{1}$$

Entities are sampled without replacement from a global candidate pool $\mathcal{U}_E$. For each attribute type k, we likewise sample N values without replacement from its candidate pool $\mathcal{U}_{V,k}$ and randomly assign them to the N entities. Consequently, entity identities are unique within a context, and values are unique within each attribute type:

$$i \neq j \quad \Longrightarrow \quad e_i \neq e_j \ \text{ and } \ v_{ik} \neq v_{jk}, \qquad \forall k \in \mathcal{K}. \tag{2}$$

Thus, each entity and each instantiated attribute value participates in a single unambiguous binding for a given attribute type. Each entity is represented by a single compact record containing the same number and types of attributes, for example,

```
Record 001:  name Alice; occupation doctor; pet cat; city Paris; hobby
                                chess.
 Record 002:  name Bob; occupation teacher; pet dog; city Rome; hobby
                                tennis.
Record 003:  name Carol; occupation nurse; pet bird; city Tokyo; hobby
                                piano.
```

Each entity name and attribute value appears only once in the context. We randomize both record order and within-record attribute order, while independently resampling entity–value assignments across contexts to avoid fixed positional or co-occurrence shortcuts.

3.2 PAIRED ITEM-ACCESS AND BINDING PROBES

For each fact $(e, k, v) \in \mathcal{B}^{(t)}$, we construct complementary probes for its constituent items and their association.

Item-access probes. Entity and value probes ask whether an individual item occurs in the context. For example, an entity probe asks *"Is Alice mentioned in the records?"*, while a value probe asks

"Is cat mentioned in the records?". Both require a binary `yes`/`no` response. For an item x, the target label is

$$y_x^{(t)} = \mathbb{I}\left[x \text{ occurs in } \mathcal{C}^{(t)}\right]. \tag{3}$$

Probe items are drawn from the corresponding candidate pools and include items absent from the current context, providing negative controls against an unconditional affirmative response. These probes measure recognition of the individual constituents without requiring the model to recover their relation.

Binding probes. We probe each association in both directions. The forward probe queries the value associated with an entity under a specified attribute type, e.g., *"The pet of Alice is:"*, with ground-truth answer v for fact (e, k, v). The reverse probe queries the entity associated with a given attribute value, e.g., *"The person that has cat as pet is:"*. Because values are unique within each attribute type, each reverse probe also has a single correct entity. Together, the paired probes distinguish access to the individual constituents from recovery of their association.

3.3 Probe Scoring

Because all candidate answers are single tokens, we score each probe using the model logits at the first answer position, allowing direct comparison among candidate answers without requiring free-form generation. For a probe q, let $z_w(q)$ denote the logit assigned to candidate token w, and let $\mathcal{V}_q$ denote its probe-specific candidate set. We normalize over this set as

$$p_q(w) = \frac{\exp z_w(q)}{\sum_{u \in \mathcal{V}_q} \exp z_u(q)}. \tag{4}$$

For item-access probes, $\mathcal{V}_q$ contains the `yes` and `no` tokens. For forward binding, it contains accepted realizations of values in $\mathcal{U}_{V,k}$; for reverse binding, it contains accepted realizations of entities in $\mathcal{U}_E$. These candidate pools are not restricted to items appearing in the current context.

Let $\mathcal{S}(x)$ denote the accepted token realizations of item x. We aggregate their probability mass as

$$P_q(x) = \sum_{w \in \mathcal{S}(x)} p_q(w). \tag{5}$$

For an item-access probe with target y_x, we use

$$\ell_{\text{access}}(q) = \big(p_q(\text{yes}) - y_x\big)^2 + \big(p_q(\text{no}) - (1 - y_x)\big)^2. \tag{6}$$

For a forward binding probe with correct value v, we use

$$\ell_{\rightarrow}(q) = \big(1 - P_q(v)\big)^2 + \sum_{v' \in \mathcal{U}_{V,k} \setminus \{v\}} P_q(v')^2. \tag{7}$$

For a reverse binding probe with correct entity e, we analogously use

$$\ell_{\leftarrow}(q) = \big(1 - P_q(e)\big)^2 + \left(\sum_{e' \in \mathcal{U}_E \setminus \{e\}} P_q(e').\right)^2 \tag{8}$$

We average these losses over probes and report entity access, value access, forward binding, and reverse binding separately. Because the probe families have different candidate spaces, we use these losses to characterize degradation within each probe family as compression increases, rather than comparing their absolute magnitudes directly.

3.4 Compression Diagnosis

Compression protocol. For each context and compression setting, we first prefill the context and apply the specified KV-cache compression method before supplying any probe question. Each probe

Table 1: Evaluated KV-cache compression methods, models, and context configurations. $N \times K$ denotes a context with N entities and K attribute types.

Category	Method	Compression principle	Model(s)	Contexts
Chunk-based	ChunkKV	Chunk selection	Qwen3-8B	$30 \times \{5, 10\}$
Score-based	Knorm	Token norm	Qwen3-8B	$30 \times \{5, 10\}$
	SnapKV	Attention score	Qwen3-8B	$30 \times \{5, 10\}$
	PyramidKV	Layer-wise allocation	Qwen3-8B	$30 \times \{5, 10\}$
Predictive	Expected Attention	Expected attention	Qwen3-8B	$30 \times \{5, 10\}$
	KVZip	Question-aware scoring	Qwen3-8B, Llama-3.1-8B	$30 \times \{5, 10, 11\}$
Other	ThinK	Dimension pruning	Qwen3-8B	$30 \times \{5, 10\}$
	KIVI	Quantization	Qwen3-8B	$30 \times \{5, 10\}$
	Merging Knorm	Token merging	Qwen3-8B	$30 \times \{5, 10\}$

starts from the same compressed context cache, without retaining changes introduced by preceding probes. The same sampled contexts and ground-truth facts are evaluated across compression settings, enabling paired comparisons as the cache budget is reduced. We additionally condition the probe losses on the recorded positions of the corresponding factual items. This allows us to determine not only whether item access and binding degrade at different rates, but also where in the context these failures emerge.

Binding-error diagnostics. Aggregate probe losses reveal whether binding and item access deteriorate differently, but do not characterize the form of individual binding errors. We therefore classify binding predictions according to whether the predicted item is supported by the current context.

For attribute type k, let

$$\mathcal{V}_k^{(t)} = \{v' : (e', k, v') \in \mathcal{B}^{(t)}\} \tag{9}$$

denote the values of type k appearing in context $\mathcal{C}^{(t)}$. For a forward probe with ground-truth value v, a prediction $\hat{v}$ is classified as

$$\textsc{Correct}: \quad \hat{v} = v, \tag{10}$$

$$\textsc{Contextual Misbinding}: \quad \hat{v} \in \mathcal{V}_k^{(t)} \setminus \{v\}, \tag{11}$$

$$\textsc{Out-of-Context}: \quad \hat{v} \notin \mathcal{V}_k^{(t)}. \tag{12}$$

This decomposition distinguishes incorrect reassociation among contextual items from predictions unsupported by the context.

4 Diagnostic Results

We apply BindProbe to examine how KV-cache compression affects item recognition and relation recovery. We assess their degradation across compression settings, its dependence on contextual position, and the types of incorrect associations produced by the model.

4.1 Experimental Setup

Models and compression methods. We evaluate Qwen3-8B across nine compression methods: ChunkKV(Liu et al., 2025b), Knorm (Devoto et al., 2024), SnapKV (Li et al., 2024), PyramidKV (Li et al., 2024), Expected Attention (Devoto et al., 2025), KVZip (Kim et al., 2025), ThinK (Xu et al., 2025), KIVI (Zirui Liu et al., 2023), and Merging Knorm (Bolya et al., 2023; Wan et al., 2025). These methods cover token eviction, dimension pruning, quantization, and merging. We additionally evaluate KVZip on Llama-3.1-8B-Instruct to examine variation across models. Table 1 summarizes the evaluated combinations.

Contexts and evaluation. Our comparisons focus on changes within each probe family relative to its uncompressed performance. We additionally examine losses by factual position and classify binding predictions using the categories defined in subsection 3.4. Our comparisons focus on changes

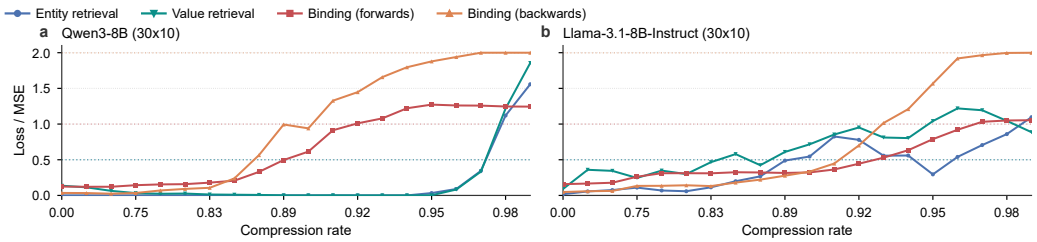


Figure 1: Item-access and binding losses across compression rates for two models.

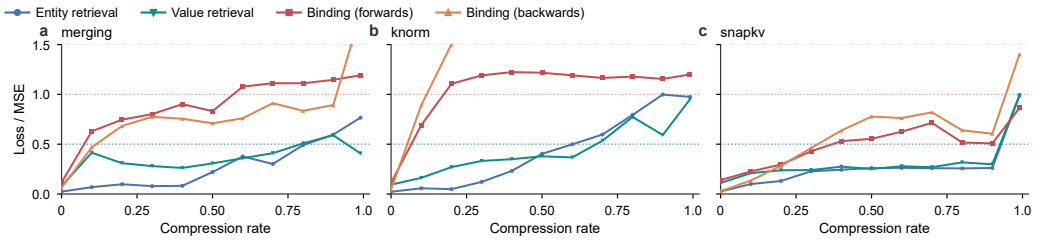


Figure 2: Item-access and binding losses across compression rates for representative KV-cache compression methods on Qwen3-8B.

within each probe family relative to its uncompressed performance. We additionally examine losses by factual position and classify binding predictions using the categories defined in subsection 3.4.

4.2 Results

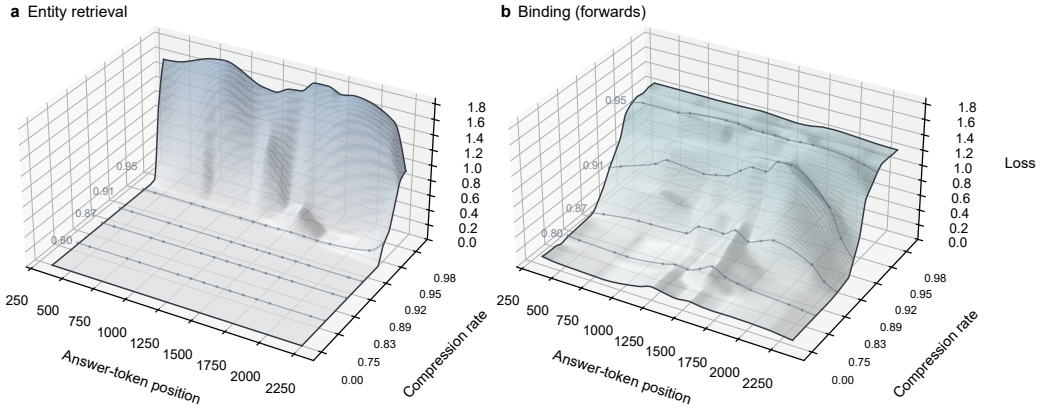


Figure 3: 3D surface plots of entity-access and forward-binding losses across compression rates and target-item positions for KVZip on Qwen3-8B.

Binding can deteriorate while item recognition remains stable. Figure 1 compares the four probe families under KVZip compression. For Qwen3-8B, both binding losses increase over a range of compression settings in which entity and value recognition remain close to their uncompressed levels. At a compression rate of approximately 0.92, for example, both recognition losses remain near zero despite pronounced increases in binding loss. Recognition deteriorates sharply only under more aggressive compression. This separation shows that successful recognition of contextual items does not guarantee recovery of their associations. The separation depends on the model. Llama-3.1-8B-Instruct exhibits greater variation in recognition loss, particularly for attribute values, and a less distinct ordering between recognition and binding degradation. The Qwen result therefore

demonstrates a clear regime of binding vulnerability, while the Llama result limits a universal ordering across models. Figure 2 extends the comparison to Merging Knorm, Knorm, and SnapKV. The curves again distinguish binding degradation from changes in item recognition, although their trajectories differ. Knorm exhibits a sharp early increase in binding loss, whereas SnapKV shows a smaller and non-monotonic separation. These results indicate that the observed vulnerability extends beyond KVZip, while its onset and severity depend on the compression method.

Binding degradation depends on contextual position. Figure 3 examines loss as a function of compression strength and the position of the target item. Forward binding exhibits pronounced increases at interior positions under settings where entity recognition remains largely stable across the context. At intermediate compression levels, the binding surface shows greater vulnerability in the middle region than near the context boundaries. Under stronger compression, elevated binding loss extends across a broader range of positions. Thus, aggregate loss curves conceal substantial variation in which contextual associations are disrupted.

Contextual misbinding precedes widespread out-of-context predictions. Figure 4 examines the highest-probability answers to forward binding probes. As compression increases, answer accuracy falls while the fraction of predictions corresponding to contextual values remains comparatively high. The gap between these quantities identifies predictions that select an existing value but assign it to the wrong entity. The initial deterioration therefore includes incorrect associations among items supplied by the context. With more aggressive compression, the in-context prediction rate also falls, and values absent from the context account for an increasing fraction of responses. This pattern distinguishes an earlier regime characterized by contextual misbinding from a later regime with widespread out-of-context predictions. Together with the recognition results, it provides behavioral evidence that compression can impair relation recovery before broad failure to recognize contextual items.

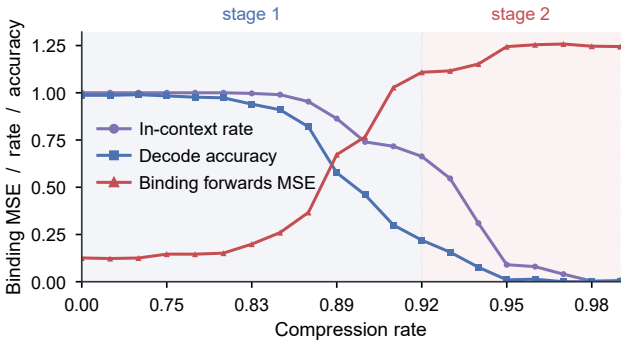


Figure 4: Decode accuracy, in-context prediction rate, and forward-binding loss across compression levels for KVZip on Qwen3-8B, with Stage 1 showing contextual misbinding and Stage 2 increasing out-of-context predictions.

5 Understanding Binding Fragility

The preceding results reveal a regime in which relational binding degrades while access to the constituent entities and values remains largely intact. To investigate the source of this asymmetry, we conduct a mechanistic case study on Qwen3-8B with KVZip using approximately 2K-token contexts containing 30 entities and 10 attributes. We first identify attention heads that are particularly important for binding and validate their functional role through targeted ablation. We then ablate KV states at individual template-aligned positions to identify cache positions that support relation recovery. Finally, we examine how KVZip scores these functionally important positions, testing whether the compression criterion systematically under-preserves information required for binding.

5.1 Results

Identifying heads that support binding. We first ask whether relation recovery depends disproportionately on a subset of attention heads. For each probe, we rank context positions by a head's attention from the answer position and compute a rank-weighted score,

$$s_q^h = \sum_i p_{q,i}^{\text{ans}} \beta^{r_{q,i}^h}, \qquad \beta = 0.95,$$

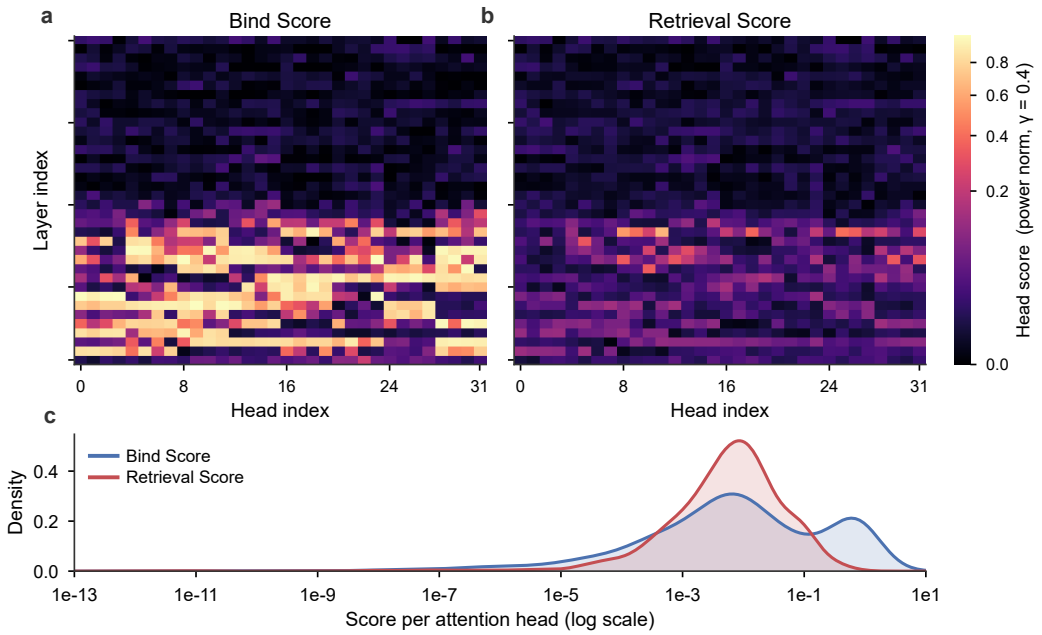


Figure 5: Binding and item-access score heatmaps and distributions on Qwen3-8B. Binding scores are concentrated in a small subset of heads.

where $r^h_{q,i}$ is the rank of position i under head h's attention and $p^{\text{ans}}_{q,i}$ measures its contribution to the correct answer. Averaging s^h_q separately over binding and item-access probes yields a binding score S^h_{bind} and an access score S^h_{access} for each head. Figure 5 visualizes these scores across the model.

The resulting scores are highly heterogeneous, exhibiting a pronounced long-tailed distribution in which a small subset of heads accounts for a large fraction of the total binding-score mass. This concentration suggests that relation recovery may depend disproportionately on a subset of attention heads, consistent with prior observations of functional specialization in contextual retrieval.

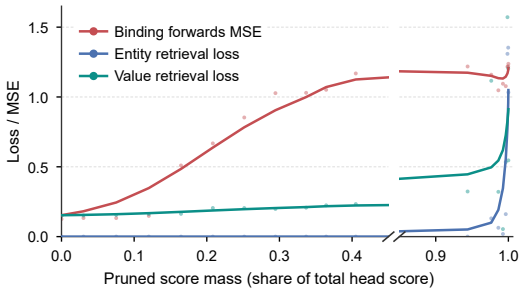


Figure 6: Ablation of binding-sensitive heads in Qwen3-8B.

We next test whether these high-scoring heads are causally important for binding. We progressively remove attention heads in descending order of binding score and measure the resulting binding and item-access losses. As shown in Figure 6, removing the highest-scoring heads sharply increases binding loss, while entity and value access remain comparatively stable. This selective degradation provides evidence that the identified heads contribute disproportionately to relation recovery rather than item access in general.

Localizing contributions across token positions. Having identified binding-sensitive heads, we next ask which cached token positions are functionally important for relation recovery. We ablate the KV states corresponding to each template-aligned position and measure the resulting changes in binding and item-access losses (Figure 7). Removing entity-name and attribute-value positions produces the largest degradation, confirming that these content positions are important for both accessing the constituent items and recovering their association. Importantly, several structural positions, including `binding[0,2]` and `intro[5]`, also produce noticeable degradation in binding. This indicates that relation recovery depends not only on the entity and value representations themselves, but also on contextual positions that help organize their association.

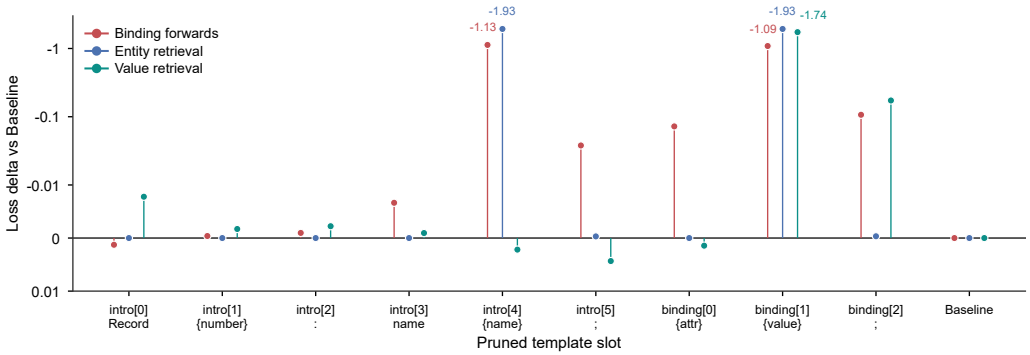


Figure 7: Effect of ablating template-aligned KV positions on binding and item-access losses.

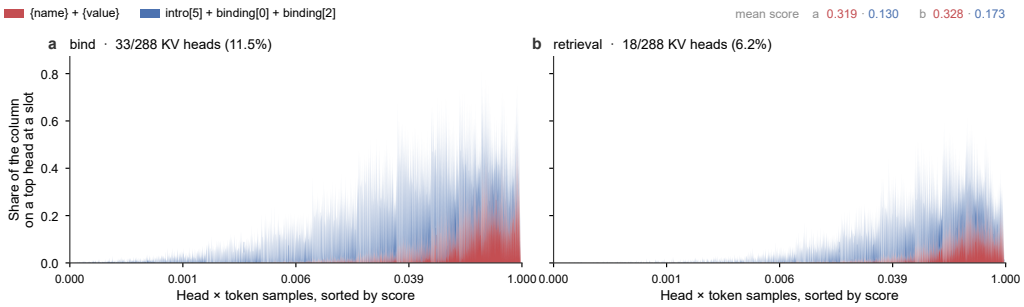


Figure 8: KVZip score distributions on binding- and item-access-sensitive KV heads.

Comparing compression scores with functional importance. We next ask whether KVZip assigns sufficient importance to the cache positions identified above. Figure 8 compares the KVZip scores of entity/value positions with those of the structural positions showing the largest binding effects in Figure 7, evaluated over KV heads accounting for the top 30% of binding- or item-access score mass. The clearest difference is across token positions. Entity and value positions receive relatively high KVZip scores, whereas the binding-relevant structural positions receive substantially lower scores, particularly on binding-sensitive heads. These lower scores make the structural positions more susceptible to eviction as the cache budget is reduced, even when the constituent entity and value representations remain comparatively well preserved. This mismatch between functional importance and compression importance provides a plausible mechanism for the binding-first degradation observed in section 4: KVZip can preserve information about the individual items while under-preserving contextual states that support recovery of their association.

6 Conclusion

We introduced BindProbe, a diagnostic framework that separates *item access* from *relational binding* under KV-cache compression. Our results reveal a regime in which binding degrades while the constituent entities and values remain accessible, producing *item-preserving misbinding*, with middle-context associations being particularly vulnerable. This distinction shows that compression-induced contextual errors cannot always be explained by simple loss of the underlying facts, but may instead arise from failures to recover the relations among facts that remain available. A mechanistic case study with Qwen3-8B and KVZip further identifies binding-sensitive heads and cache positions, and shows that several functionally important positions receive relatively low compression scores, making them more susceptible to eviction. Together, these findings link item-preserving misbinding to vulnerabilities in the compressed cache, showing that KV-cache compression can disrupt relational structure even when contextual information remains accessible.

AI USE STATEMENT

This paper was polished for grammar and presentation using ChatGPT. All AI-assisted suggestions were reviewed and verified by the authors, who take full responsibility for the content, experiments, analyses, and conclusions of this work.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or sensitive personal data. We use publicly available models and datasets for our experiments. To the best of our knowledge, this work does not introduce additional privacy, security, discrimination, or other ethical risks beyond those inherent to the underlying models and datasets. We have followed the ICLR Code of Ethics in conducting and reporting this research.

REPRODUCIBILITY STATEMENT

The anonymized source code and experimental configurations are provided in the supplementary materials and will be publicly released upon acceptance.

REFERENCES

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster, 2023. URL https://arxiv.org/abs/2210.09461.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, and Wen Xiao. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling, 2025. URL https://arxiv.org/abs/2406.02069.

Alessio Devoto, Yu Zhao, Simone Scardapane, and Pasquale Minervini. A simple and effective l_2 norm-based strategy for kv cache compression, 2024. URL https://arxiv.org/abs/2406.11430.

Alessio Devoto, Maximilian Jeblick, and Simon Jégou. Expected attention: Kv cache compression by estimating attention from future queries distribution, 2025. URL https://arxiv.org/abs/2510.00636.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context?, 2024. URL https://arxiv.org/abs/2310.17191.

Yoav Gur-Arieh, Mor Geva, and Atticus Geiger. Mixing mechanisms: How language models retrieve bound entities in-context, 2026. URL https://arxiv.org/abs/2510.06182.

Jang-Hyun Kim, Jinuk Kim, Sangwoo Kwon, Jae W. Lee, Sangdoo Yun, and Hyun Oh Song. Kvzip: Query-agnostic kv cache compression with context reconstruction, 2025. URL https://arxiv.org/abs/2505.23416.

Yucheng Li, Huiqiang Jiang, Qianhui Wu, Xufang Luo, Surin Ahn, Chengruidong Zhang, Amir H. Abdi, Dongsheng Li, Jianfeng Gao, Yuqing Yang, and Lili Qiu. Scbench: A kv cache-centric analysis of long-context methods, 2025. URL https://arxiv.org/abs/2412.10319.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation, 2024. URL https://arxiv.org/abs/2404.14469.

Minghui Liu, Aadi Palnitkar, Tahseen Rabbani, Hyunwoo Jae, Kyle Rui Sang, Dixi Yao, Shayan Shabihi, Fuheng Zhao, Tian Li, Ce Zhang, Furong Huang, and Kunpeng Zhang. Hold onto that thought: Assessing kv cache compression on reasoning, 2025a. URL https://arxiv.org/abs/2512.12008.

Xiang Liu, Zhenheng Tang, Peijie Dong, Zeyu Li, Yue Liu, Bo Li, Xuming Hu, and Xiaowen Chu. Chunkkv: Semantic-preserving kv cache compression for efficient long-context llm inference, 2025b. URL https://arxiv.org/abs/2502.00299.

Xiang Liu, Zhenheng Tang, Hong Chen, Peijie Dong, Zeyu Li, Xiuze Zhou, Bo Li, Xuming Hu, and Xiaowen Chu. Semantic integrity matters: Benchmarking and preserving high-density reasoning in kv cache compression, 2026. URL https://arxiv.org/abs/2502.01941.

Zhongwei Wan, Xinjian Wu, Yu Zhang, Yi Xin, Chaofan Tao, Zhihong Zhu, Xin Wang, Siqi Luo, Jing Xiong, Longyue Wang, and Mi Zhang. D2o: Dynamic discriminative operations for efficient long-context inference of large language models, 2025. URL https://arxiv.org/abs/2406.13035.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality, 2024. URL https://arxiv.org/abs/2404.15574.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2024. URL https://arxiv.org/abs/2309.17453.

Yuhui Xu, Zhanming Jie, Hanze Dong, Lei Wang, Xudong Lu, Aojun Zhou, Amrita Saha, Caiming Xiong, and Doyen Sahoo. Think: Thinner key cache by query-driven pruning, 2025. URL https://arxiv.org/abs/2407.21018.

June Yong Yang, Byeongwook Kim, Jeongin Bae, Beomseok Kwon, Gunho Park, Eunho Yang, Se Jung Kwon, and Dongsoo Lee. No token left behind: Reliable kv cache compression via importance-aware mixed precision quantization, 2024. URL https://arxiv.org/abs/2402.18096.

Jiayi Yuan, Hongyi Liu, Shaochen Zhong, Yu-Neng Chuang, Songchen Li, Guanchu Wang, Duy Le, Hongye Jin, Vipin Chaudhary, Zhaozhuo Xu, Zirui Liu, and Xia Hu. Kv cache compression, but what must we give in return? a comprehensive benchmark of long context capable approaches, 2024. URL https://arxiv.org/abs/2407.01527.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi : Plug-and-play 2bit kv cache quantization with streaming asymmetric quantization. 2023. doi: 10.13140/RG.2.2.28167.37282. URL https://www.researchgate.net/doi/10.13140/RG.2.2.28167.37282.