Understanding the Physics of Key-Value Cache Compression for LLMs through Attention Dynamics

Anonymous authors
Paper under double-blind review

Abstract

As context windows in large language models (LLMs) grow to hundreds of thousands of tokens, the key–value (KV) cache becomes the main memory cost during autoregressive decoding. Recent compression methods cut this cost by 80%–90% with little accuracy loss on standard benchmarks, which has encouraged a common assumption that the cache is largely redundant. We question that assumption — attention is not only a store of past tokens; it also decides how information is routed to the decoder, so a key–value pair can survive compression yet still be unreachable when the answer is produced. We therefore study KV compression as a controlled change to this routing, separating whether a token is kept from whether it can still be used. Using small, tightly controlled datasets for multi-entity tracking, disambiguation, coreference, and multi-hop reasoning, we report four major findings - (1) Moderate compression damages internal representations while leaving answers largely correct, confirming that much of the cache is redundant, (2) Near 90% compression, every model we test shows a sharp jump in hallucination, a "safety cliff", that coincides with spikes in a simple quantity we call the Global Eviction Ratio (GER): the fraction of answer tokens dropped from every head at once, (3) The layers where routing is most fragile sit at different depths in different model families, so depth-based pruning rules do not transfer reliably, (4) Finally, even when answer tokens survive, too much agreement among attention heads can leave the model unable to re-route to them, a second failure mode we call rigidity. What limits compression, then, is not how many tokens are kept but whether a usable route to the answer survives, an inference-time counterpart to lottery-ticket sparsity in attention.

1 Introduction

The recent evolution of large language models (LLMs) has been defined as much by context length as by parameter scale (Grattafiori et al., 2024; Yang et al., 2025; Team et al., 2024). Context windows have expanded from a few thousand tokens to hundreds of thousands and beyond, enabling document-level reasoning, repository-scale code analysis, and persistent conversational memory (Liu et al., 2025b). This mirrors earlier scaling laws for parameters and data (Kaplan et al., 2020; Hoffmann et al., 2022; Muennighoff et al., 2025), with longer context treated as one more axis of capability. Yet unlike parameters, which are shared across tokens, the key–value (KV) cache grows linearly with sequence length during autoregressive decoding. For a transformer with L layers, H_{KV} heads of dimension D_h, batch size B, sequence length S, and precision b, the memory footprint scales as

$$\text{Memory} = B \times S \times L \times H_{KV} \times D_h \times b. \tag{1}$$

As S increases, KV storage rapidly dominates inference memory, often exceeding parameter storage itself. In practice, memory, not compute, becomes the principal bottleneck (Meng et al., 2026) for long-context LLMs.

KV cache compression (Li et al., 2024a) is the standard engineering response to this problem. Methods based on quantization (Hooper et al., 2024; Liu et al., 2024; Kang et al., 2024), token eviction (Wan et al., 2025; Cai et al., 2025; Li et al., 2024b), adaptive head pruning (Feng et al., 2025), or clustering (Liu et al., 2025a;

Zhang et al., 2025) routinely remove 80%–90% of the cache with little measured loss on benchmarks such as LongBench and RULER (Bai et al., 2023; Hsieh et al., 2024; Zhang et al., 2024). The natural reading of these results is that the cache holds a lot of redundancy and that careful pruning removes it without changing what the model does. We think this reading misses something: attention is not a passive store, it also decides how the tokens it keeps are routed to the decoder.

Work on attention circuits supports this view. Induction-head and circuit-level analyses show that reasoning runs along specific routes that cross heads and layers, not on tokens stored in isolation (Clark et al., 2019; Voita et al., 2019; Elhage et al., 2021; Olsson et al., 2022). If that is how attention works, then keeping a key–value pair is not enough: a token can stay in the cache but become unreachable (Lee, 2025) once the routes that reach it are cut. The Lottery Ticket Hypothesis points the same way from the other direction, suggesting that a small subset of attention components can preserve behavior as long as the routes that matter stay intact (Frankle & Carbin, 2019; Otsuka et al., 2025a). This motivates treating KV compression as a controlled way of cutting routes and watching for the point at which the answer becomes unreachable.

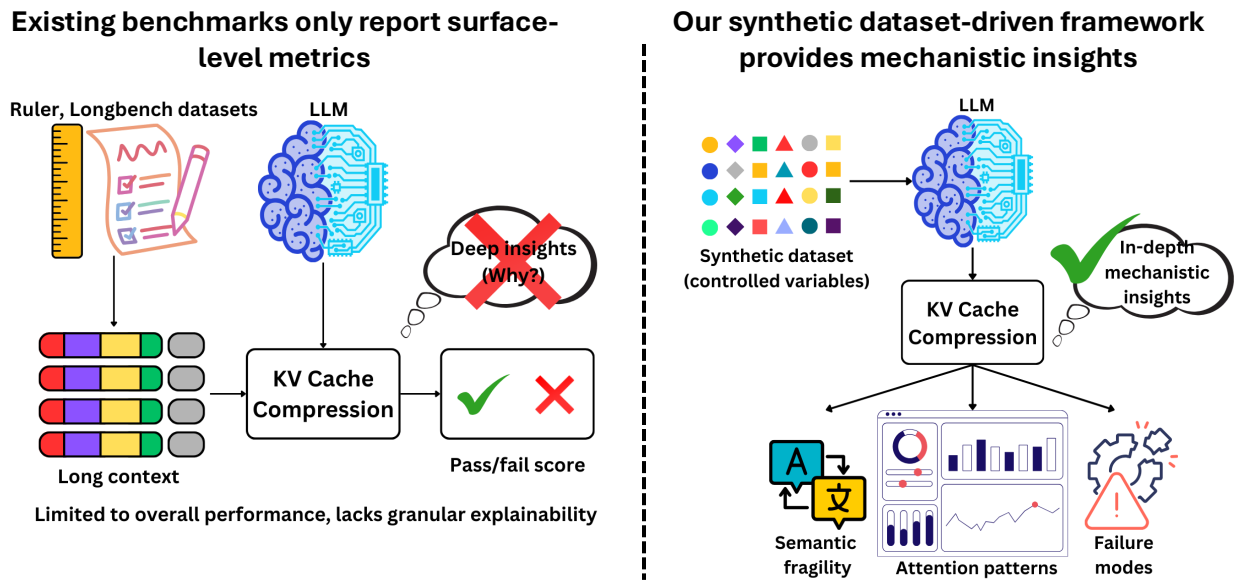


Figure 1: Standard benchmarks versus our controlled setup for studying KV cache compression. Existing benchmarks (left) mostly report coarse task accuracy, while our controlled framework (right) tracks which answer tokens stay reachable, when routing breaks down, and where hallucination begins under compression.

This leads to our main hypothesis. Inside the dense attention layers we expect a few sparse *token-route lottery tickets* (TR-LTs): small sets of cross-head, cross-layer routes that are enough to reach the answer and generate it correctly. Mild compression may simply expose these routes, while heavy compression destroys them. If so, compression should not produce a smooth, gradual decline; performance should hold up while the routes survive and then fail sharply once they are gone (Ma et al., 2026; Ersoy et al., 2025).

Research Objective and Hypothesis

Our objective is to explain *why* accuracy under KV compression stays high until it collapses abruptly at high compression. We test the hypothesis that correct generation requires at least one surviving attention route from the query to answer-relevant tokens, and that the abrupt failure near high compression happens when these routes are globally deleted, not when stored memory is gradually reduced. Under this hypothesis, the accuracy retained at moderate compression reflects route redundancy rather than the absence of information loss.

To test this, we build a controlled evaluation setup with synthetic datasets and targetted evaluation metrics (Figure 1). Instead of looking only at answer accuracy, we measure two concepts : the Global Eviction Ratio (GER), the fraction of answer tokens that have been dropped from every head, and head-level consensus, how much the heads in a layer agree on what to read. Across several models (the LLaMA and Qwen families) and compression levels up to 90%, the same patterns recur.

The findings are consistent across models. Moderate compression noticeably degrades internal representations but leaves task accuracy almost unchanged, so there is a wide margin of redundancy. Near 90% compression, every model shows a sharp *safety cliff* in hallucination that tracks spikes in GER, marking the point where the answer is no longer reachable. The layers where routing is most fragile fall at different depths in different model families, so a fixed depth-based pruning rule (such as a "pyramid" schedule) does not carry over cleanly from one architecture to another. Finally, probing shows that some answer types hold up better than others: common, concrete categories like 'Person' and 'Location' stay stable, while relational or rarer ones like 'Organization', 'Event', and especially 'Creature' fall off quickly.

Main Finding

Compression does not degrade reasoning smoothly. The model keeps working as long as a sparse token-route lottery ticket survives, and fails sharply when global eviction removes the route or excessive head agreement leaves it unusable.

This points to two failure modes. In *erasure*, the answer tokens are dropped from every head, so no route to them remains. In *rigidity*, the tokens survive but the heads agree so strongly on what to read that the model cannot re-route to them. Either way, how much compression a model tolerates is set not by the number of tokens kept but by whether usable routes to the answer remain.

By linking these measurements to sparsity theory for attention (Huang et al., 2022; Otsuka et al., 2025b;a), we change the question. Instead of asking how much of the KV cache can be thrown away before accuracy drops, we ask which routes have to survive for the model to keep answering from context. Our answer is that reliability under compression depends on whether a few sparse token-route lottery tickets survive, not on how much memory is kept.

2 Background

2.1 Self-Attention and the KV Cache in Autoregressive Decoding

Transformers (Vaswani et al., 2017) are built upon the self-attention mechanism. Given an input sequence $X \in \mathbb{R}^{S \times d}$, each layer computes queries, keys, and values:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V,$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$. Scaled dot-product attention is then defined as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V.$$

In multi-head attention (MHA), this operation is replicated across H heads and concatenated. In autoregressive decoding, recomputing K and V for all previous tokens at each time step would incur $\mathcal{O}(S^2)$ cost. Instead, models cache previously computed keys and values, forming the *KV cache*. At decoding step t, attention is computed between the current query q_t and all stored keys $K_{1:t}$, while values $V_{1:t}$ are reused. Unlike parameters, which are shared across tokens, the KV cache grows linearly in S and B as seen in Equation 1, making it the dominant memory consumer in long-context inference (Li et al., 2024a).

2.2 Efficient Attention Mechanisms

The quadratic $\mathcal{O}(S^2)$ cost of full self-attention has motivated many efficiency-oriented architectures (Sun et al., 2026). One line of work restricts attention to a fixed or learned subset of token pairs: Sparse

Transformers (Child, 2019), Longformer (Beltagy et al., 2020), and BigBird (Zaheer et al., 2020) use locality, block structure, or random connectivity to keep some global context while cutting cost. A second line approximates the attention matrix: Linformer (Wang et al., 2020) projects it to low rank before the softmax, and Performer (Choromanski et al., 2021) replaces the softmax kernel with a linear feature map, reaching $\mathcal{O}(S)$ cost at the price of approximation error. Other methods limit attention to sliding windows (Beltagy et al., 2020), which favors short-range dependencies. State-space models such as Mamba (Gu & Dao, 2024) drop explicit attention entirely and keep an implicit memory in linear time. Each of these choices trades global expressivity, compute, approximation quality, and hardware fit differently.

2.3 IO-Aware Exact Attention and KV Sharing Mechanisms

FlashAttention (Dao et al., 2022) takes a different route: it speeds up exact softmax attention without changing the math. Instead of writing the full $S \times S$ attention matrix to high-bandwidth memory (HBM), it tiles and fuses the computation and does the softmax and accumulation in on-chip SRAM. Since memory traffic, not arithmetic, is often the real bottleneck, this gives large speedups while staying exact. FlashAttention-2 (Dao, 2024) improves the work partitioning and parallelism for higher throughput across GPUs. The lesson is that attention is often limited more by data movement than by raw compute.

A separate idea reduces memory by sharing key and value projections across heads. Multi-Query Attention (MQA) (Shazeer, 2019) uses a single shared KV head, and Grouped-Query Attention (GQA) (Ainslie et al., 2023) lets H_Q query heads share a smaller number $H_{KV} < H_Q$ of KV heads. This shrinks KV cache memory by a factor of H_{KV}/H_Q with little quality loss, which is why most current large models use GQA or MQA (Li et al., 2024a). The two directions are complementary: exact kernels target data movement, while KV sharing reduces redundancy in the key–value projections.

2.4 KV Cache Compression

A third, fast-growing line compresses the KV cache itself at inference time, treating it as a memory-management problem. The methods fall into a few groups. *Quantization* methods (Hooper et al., 2024; Yang et al., 2024b) lower the bit precision b of stored keys and values, saving memory at a small perplexity cost. *Eviction and heavy-hitter* methods drop tokens judged unlikely to matter, using cues such as recency, cumulative attention mass, token norms, or learned predictors (e.g., H2O (Zhang et al., 2023), StreamingLLM (Xiao et al., 2024), Scissorhands (Liu et al., 2023), CurDKV (Sengupta et al., 2025)); they assume attention is concentrated on a few tokens. *Chunking and summarization* methods (Liu et al., 2025c; Bai et al., 2024) replace groups of related tokens with compact representatives. *Query-agnostic* methods (Kim et al., 2025; Chari & Durme, 2025) compress the cache before seeing the query. *Adaptive* methods (Ge et al., 2024; Zhou et al., 2025; Feng et al., 2025; Qin et al., 2025) try to predict which tokens later decoding will need (e.g., expected-attention (Devoto et al., 2025) or AdaKV (Feng et al., 2025)).

These methods are usually judged by compression ratio, latency, perplexity or accuracy loss, robustness on long or adversarial inputs, and whether they need fine-tuning. The common finding is that much of the cache can be removed, often 70%–90%, with little benchmark cost. But these evaluations almost always look at final accuracy, which quietly treats keeping a token as the same thing as being able to use it. They rarely separate storage, access, and use, and so leave open the question we study here: whether a compressed cache still keeps the routes needed to reach the answer during generation.

3 Methodology

3.1 Datasets and Controlled KV Compression Evaluation

Standard long-context benchmarks such as LongBench (Bai et al., 2023), InfiniteBench (Zhang et al., 2024), and RULER (Hsieh et al., 2024) measure task accuracy on long inputs, but they are not built to isolate how KV compression affects attention routing. They mix together three things: whether information is still stored, whether it can still be reached through attention, and whether it is actually used when generating. A model might answer correctly only because redundancy hid the routing damage, or fail for reasons that

have nothing to do with compression. Controlled synthetic data is a good way to separate these effects, since it exposes internal mechanisms that natural benchmarks blur (Allen-Zhu & Li, 2025; 2024). Following that approach, we build a set of synthetic datasets that let us treat KV compression as a controlled cut to token routes.

We build the datasets in two ways. Some are generated by prompting a larger language model (Gemini 2.5 pro) with tightly structured instructions that fix the entity–attribute links, control how often each entity is mentioned, and ask questions in both directions. Others are filled from templates by slot substitution, which removes linguistic noise and gives us full control. Used together, the two keep the answers route-sensitive while spanning a range of realism.

Design principles. Each dataset is constructed to satisfy four constraints: (1) *route sensitivity*, where correct answers depend on preserving specific cross-layer token pathways; (2) *minimal redundancy*, preventing accidental recoverability after aggressive eviction; (3) *bidirectional querying*, enforcing both subject→attribute and attribute→subject tracing; and (4) *failure interpretability*, enabling classification into representational erasure or rigidity. Under these principles, compression functions as a controlled ablation of token-route structures.

Generation framework. For generative datasets, we use structured prompting to enforce attribute-entity bindings and controlled semantic relationships. The generic procedure is formalized in Algorithm 1.

Algorithm 1 Controlled Synthetic Example Generation

Require: Subject s, structural constraints C, length range L
1: Prompt LLM with instructions enforcing C
2: Replace entity names with synthetic identifiers
3: Generate passage P
4: Generate QA pairs under bidirectional linkage rules
5: Validate explicit answer grounding
6: Store $(P, \{(q_i, a_i)\})$

For template-based datasets, examples are produced via deterministic slot substitution from pre-generated entity pools, optionally followed by controlled perturbations (e.g., pronoun swaps). This removes surface variability while preserving structural dependencies.

Dataset suite. The resulting suite spans short to long contexts (150–1300+ tokens) and targets distinct routing stressors:

- **Base task:** Short passages with one subject and tightly bound attributes; probes precision of eviction under minimal redundancy.
- **Knowledge manipulation:** Slot-filled biographies testing minimally distributed factual structures and systematic eviction biases.
- **Multi presence:** Repeated entity mentions requiring instance disambiguation; evaluates positional robustness under pruning.
- **Multi entity:** Multiple semantically linked entities; tests cross-entity and cross-head routing integrity.
- **Long context:** Extended passages requiring distributed reasoning across distant spans; probes route capacity at scale.
- **Coreference:** Controlled pronoun perturbations with "I don't know" ground truth; detects fine-grained routing failures and hallucination under compression.
- **Hops:** Chain-structured entities requiring multi-hop reasoning; evaluates preservation of sequential semantic routes.

Representative prompt and template structures are summarized below to ensure reproducibility. Table 3 highlights samples of each task while Table 6 shows forward and reverse samples in Appendix A.

Name	Passage Length (words)	# Queries	Total Passages	Total Queries	LLM Generated
Base task	175 ± 25	6	200	1199	Yes
Knowledge manipulation	32	13	4000	52000	No
Multi presence	500 ± 100	10	100	1000	Yes
Multi entity	500 ± 100	10	100	1000	Yes
Long context	1100 ± 200	10	100	1000	Yes
Coreference	44	9	8000	72000	No
Hops	1100 ± 20	16	100	1600	Yes

Table 1: Structural summary of the synthetic dataset suite used for controlled KV compression evaluation.

Base task (generative prompt)

```
Generate a passage about <subject> (150-200 words)
- >=4 directly linked attributes
- Fake entity names
Generate 6 QA pairs linking subject <-> attributes.
```

Knowledge manipulation (Template)

```
<Firstname Lastname> was born on <Date>.
They studied in <University>.
They work as <Occupation>.
They live in <City>.
```

Multi presence / Multi entity / Long context (prompt)

```
Generate passage (400-1300 words)
- Multiple entities or repeated mentions
- Explicit attribute linkage
- Bidirectional QA generation
```

Coreference (template)

```
<Firstname Lastname> was born on <Date> in <Location>.
<He/She> studied at <University>.
...
# Pronoun-swapped queries -> answer = "I don't know"
```

Hops (prompt)

```
Generate subject + 3 linked entities
- Sequential semantic linkage
- 16 QA pairs spanning multi-hop chains
```

Dataset statistics. Table 1 summarizes the suite: average passage length, queries per passage, and totals. The passages range from short, tightly controlled ones (Base task, Knowledge manipulation) to multi-hop narratives over a thousand words (Long context, Hops), which lets us apply the same compression ratios across very different context sizes. The suite deliberately mixes LLM-generated passages, which add controlled variety, with template-built ones, which give precise control, so we can study compression under both.

Why short, controlled contexts are appropriate. The contexts in our suite are shorter (roughly 150–1300 tokens) than those in LongBench (Bai et al., 2023), InfiniteBench (Zhang et al., 2024), or RULER (Hsieh et al., 2024), which typically use several thousand tokens. This is a deliberate design choice rather than a limitation, for three reasons. First, the failure mode we study is governed by the *compression ratio*, not the absolute context length: a token route is deleted when its key–value entry is evicted across all heads, and this can be induced at any sequence length by raising the fraction of entries removed. Sweeping compression from 10% to 90% on a 1000-token context therefore exercises the same eviction dynamics that a fixed-budget method would impose on a much longer one, while keeping the number of answer-critical tokens small enough to track exactly. Second, our reachability metrics (GER) and linear probes require a *ground-truth set of answer-relevant token positions*. In naturalistic multi-thousand-token passages this annotation is ambiguous and the answer is often supported redundantly by paraphrase, which confounds whether a failure stems from routing loss or from incidental recovery. Short, slot-controlled passages remove this confound and let us attribute each failure to a specific structural cause. Third, replacing all entities with synthetic names prevents the model from answering through parametric memory, so observed performance must come from in-context routing — the quantity we wish to perturb. We emphasize that the suite is not limited to new entity names: it also varies entity multiplicity (Multi presence, Multi entity), enforces coreference and abstention (Coreference), and chains relations across hops (Hops), so that disambiguation, relational, and multi-step routing are stressed alongside simple binding. Extending these probes to longer contexts and to relational or temporal edits is a natural next step (Section 6).

Significance of the generated synthetic datasets. Unlike conventional long-context benchmarks that primarily report aggregate accuracy, our framework enables causal attribution of compression failures to specific structural mechanisms within attention. The Base task and Knowledge manipulation datasets isolate eviction precision and entity–attribute binding stability. Multi presence and Multi entity expose positional sensitivity and cross-head routing fragility. Long context stresses route capacity under large token budgets, while Hops directly probes preservation of multi-step semantic chains. Coreference evaluates fine-grained representational consistency and hallucination sensitivity. This layered structure permits systematic classification of failures into representational erasure, where all viable token routes to critical information are removed, and representational rigidity, where tokens survive but routing flexibility collapses. Compression is thus analyzed as a structural phase transition in semantic reachability rather than as a monotonic decline in accuracy.

3.2 Tagging Framework

To move beyond aggregate performance metrics, each question–answer pair is annotated using a structured tagging framework that enables fine-grained behavioral analysis under compression.

Tagging operates along two orthogonal axes: answer type and question difficulty. The answer-type axis categorizes responses according to semantic domain, allowing us to examine whether certain conceptual representations are more resilient to compression. The answer categories include *Person*, *Thing*, *Organization*, *Creature*, *Location*, *Numerals*, *Date/Time*, and *Event* (Table 4 in Appendix A). By decomposing performance along these semantic dimensions, we can detect systematic degradation patterns, identify hierarchies of representational stability, and evaluate whether compression disproportionately affects particular classes of concepts.

The second axis captures reasoning demand through question-difficulty tags. Questions are classified as *Standard* (direct retrieval of explicitly stated information), *Manipulated* (requiring implicit contextual interpretation or transformation), or *Part* (requiring aggregation of information distributed across multiple textual regions). This dimension separates failures caused by token loss from those arising due to reasoning complexity (Table 5 in Appendix A). A model may retain the necessary tokens yet fail on high-difficulty queries that require multi-step inference; conversely, it may fail simple retrieval tasks due to representational erasure.

3.3 Experimental Setup

All compression experiments were conducted using NVIDIA's KVPress library[1], which provides a unified interface for scoring-based and pruning-based KV cache compression during autoregressive decoding. For all models, KV compression was applied at inference time without additional fine-tuning. We systematically vary compression ratios from mild pruning (10%) to aggressive eviction (up to 90% removal of KV entries) and report performance as a function of the retained KV budget.

Scoring and Pruning Strategies. We adopt Expected Attention (Devoto et al., 2025) as the scoring function, which estimates token importance by aggregating expected attention weights across decoding steps and layers. This per-layer scoring mechanism allows high-salience key–value pairs, particularly those residing in structurally significant heads, to be preferentially preserved. Two complementary pruning strategies are employed:

- **FINCH (Chunk) Press** (Corallo & Papotti, 2024) performs chunk-level pruning by partitioning the context into contiguous segments and removing the lowest-scoring tokens within each segment. This approach is computationally efficient and preserves coarse-grained structural coverage across the document.
- **AdaKV Press** (Feng et al., 2025) performs head-wise global pruning by ranking KV entries across all heads simultaneously and removing the lowest-scoring entries irrespective of positional grouping. This provides finer-grained adaptive compression at the cost of increased overhead.

The combination of FINCH and AdaKV enables evaluation under both coarse structural pruning and globally adaptive head-aware compression, allowing us to analyze whether routing collapse depends on local chunk structure or global head-level importance.

Inference Regimes. Compression is evaluated under two inference settings:

- **Question-agnostic:** The model first receives the full context and performs KV pruning without access to the downstream question. The query is provided only after compression. This setting ensures that eviction decisions are independent of retrieval demands and reflects realistic deployment scenarios.
- **Question-aware:** The list of candidate questions is provided prior to pruning, allowing compression decisions to condition on anticipated query structure. This setting measures the upper bound of compression tolerance when routing can be optimized for known retrieval targets.

Comparing these regimes isolates whether compression robustness arises from intrinsic route redundancy or from query-conditioned pruning strategies.

Models. We evaluate five instruction-tuned LLMs spanning two architectural families:

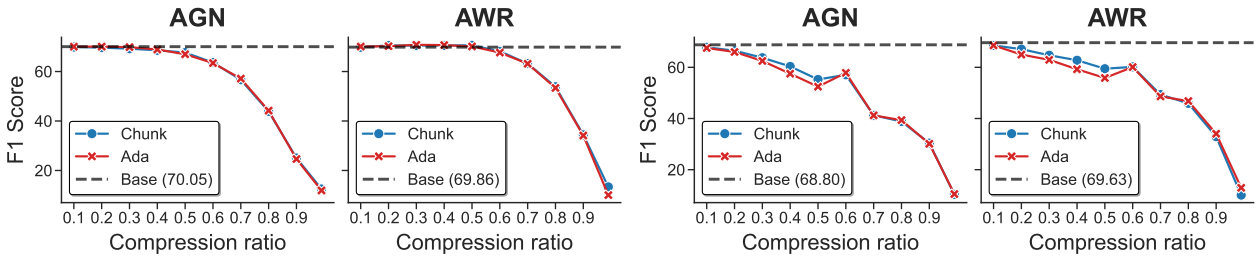
- LLaMA-3.2 3B Instruct and LLaMA-3 8B Instruct (Grattafiori et al., 2024),
- Qwen-2.5 3B Instruct, Qwen-2.5 7B Instruct, and Qwen-2.5 14B Instruct (Yang et al., 2024a).

These models cover a range of parameter scales (3B–14B) and architectural variations in attention design and training data, enabling analysis of model-dependent compression tolerance. All checkpoints were obtained from the Hugging Face model hub[2] and used without modification.

Implementation Details. All experiments were executed on a single NVIDIA RTX A6000 GPU using the KVPress text generation pipeline. The default configuration enforces question-agnostic pruning unless explicitly overridden. Generation is performed using greedy decoding to avoid variability introduced by sampling. Performance is evaluated at the token level, with answers matched using the LLaMA-3 8B tokenizer to ensure consistent segmentation across models and compression settings.

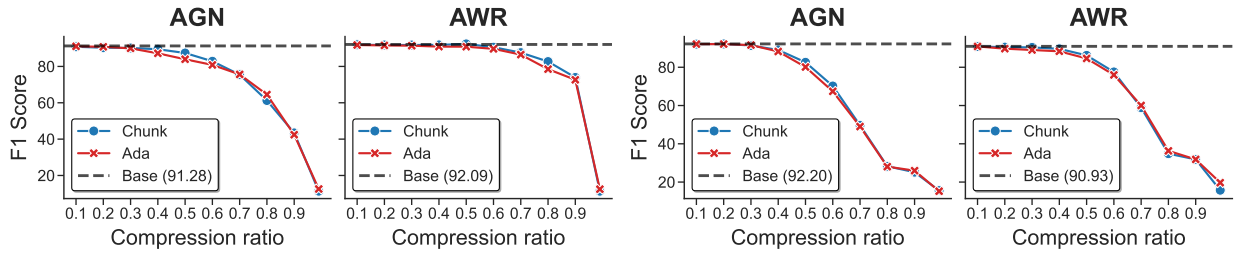
With greedy decoding and a fixed tokenizer, any difference we observe comes from the compression itself rather than from sampling noise or tokenization.

[1] `https://github.com/NVIDIA/kvpress/`
[2] `https://huggingface.co/models`



(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

Figure 2: Base task performance across compression levels. Localized spikes under moderate compression suggest sparse substructure effects. Corresponding results for LLaMA-3.2 3B, Qwen-2.5 3B, and Qwen-2.5 14B are shown in Figure 27 of Appendix B.



(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

Figure 3: Knowledge manipulation results. Qwen exhibits a more gradual rate of degradation under compression, particularly in the question-aware setting.

4 Results

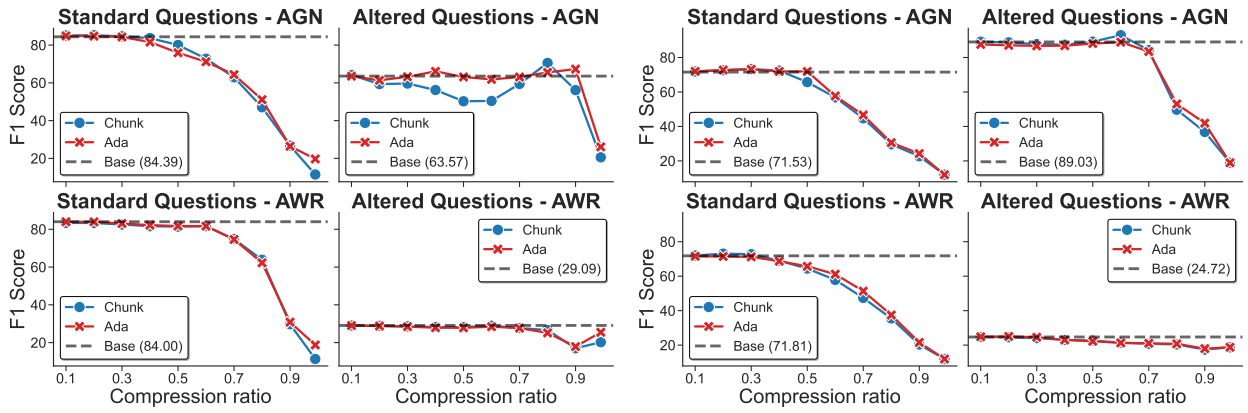
4.1 Aggregate Performance Trends

Table 2 reports the base (0% compression) F1 scores across all tasks under question-agnostic (AGN) and question-aware (AWR) settings.

Model	Setup	Base	Know	Multi-P	Multi-E	Long	Coref	Hops
LLaMA-3 8B	AGN	70.05	91.28	79.64	80.33	45.93	75.63	31.54
	AWR	69.86	92.09	77.66	81.93	45.54	60.90	31.25
Qwen-2.5 7B	AGN	68.80	92.20	73.63	77.46	38.68	78.89	27.13
	AWR	69.63	90.93	71.88	79.38	40.20	52.00	26.35

Table 2: Base (0% compression) F1 scores across datasets.

Base task. On the Base task (Figure 2), both models operate near 70% F1 (LLaMA: 70.05 AGN, 69.86 AWR; Qwen: 68.80 AGN, 69.63 AWR). Under increasing compression, performance generally declines; however, the question-aware setup exhibits localized non-monotonic spikes, particularly for Qwen around intermediate compression levels. This suggests that moderate pruning can remove interfering or redundant KV routes, temporarily improving alignment. Such behavior is consistent with the existence of sparse token-route (Zhu et al., 2025) substructures within attention that remain functionally intact under partial eviction.



(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

Figure 4: Coreference performance across setups. Question-aware pruning significantly increases overconfident errors.

Knowledge manipulation presents a contrasting pattern (Figure 3). Qwen slightly surpasses LLaMA in the question-agnostic setting (92.20 vs. 91.28), and while LLaMA is marginally stronger in question-aware (92.09 vs. 90.93), Qwen degrades more gradually under higher compression. This task emphasizes structured transformation rather than pure retrieval, indicating that Qwen's robustness derives from instruction-conditioned reasoning, whereas LLaMA's strength lies in stable factual retention. Notably, both models remain above 90% F1 until approximately 40% compression.

Coreference (Figure 4) exposes a pronounced divergence between AGN and AWR. LLaMA drops from an aggregate performance of 75.63 (AGN) to 60.90 (AWR), while Qwen drops from aggregate performance of 78.89 (AGN) to 52.00 (AWR), a 26.89-point reduction. Conditioning pruning on anticipated queries appears to bias models toward premature commitment, reducing their ability to correctly abstain when inconsistencies are introduced, supplementing the results obtained by Jin et al. (2025).

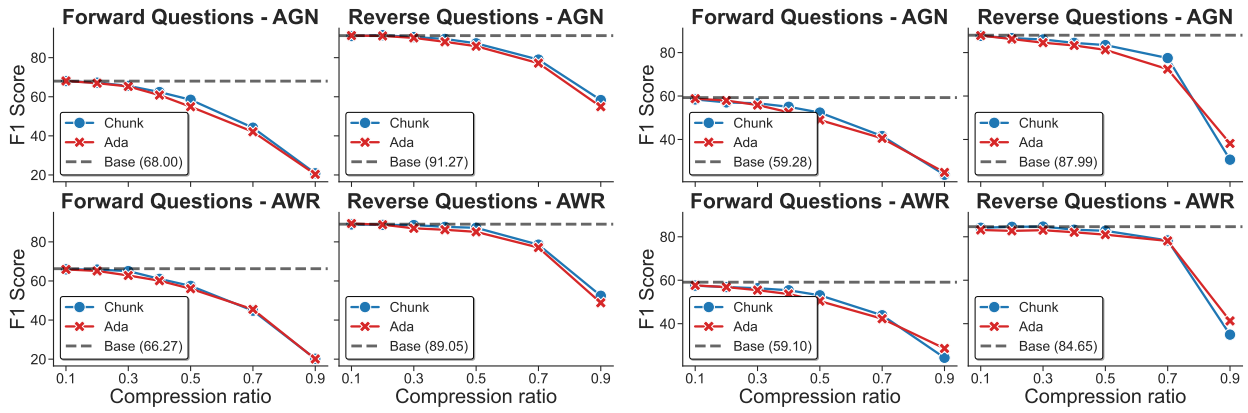
Multi presence (Figure 5): LLaMA outperforms Qwen overall (79.64 vs. 73.63 aggregate AGN). Even at full cache, forward queries are much weaker than reverse ones (68.00 forward vs. 91.27 reverse AGN; 66.27 forward vs. 89.05 reverse AWR), and this gap widens under compression, indicating asymmetric routing fragility when an entity is mentioned repeatedly.

Multi entity, in contrast, shows more balanced behavior (LLaMA 80.33 aggregate AGN; Qwen 77.46 aggregate AGN), suggesting that distributing attributes across distinct entities mitigates interference (79.94 forward and 80.72 reverse AGN vs. 80.88 forward and 82.98 reverse AWR).

Long context and Hops reveal (Figures 7, 8 and 9) deeper structural limits. Even without compression, Long context F1 remains modest (LLaMA 45.93 AGN; Qwen 38.68 AGN), and Hops is lower still (LLaMA 31.54 AGN; Qwen 27.13 AGN). Multi-hop reasoning degrades more rapidly than direct retrieval (Li et al., 2025) as compression increases, suggesting that relational pathways are more fragile than token presence.

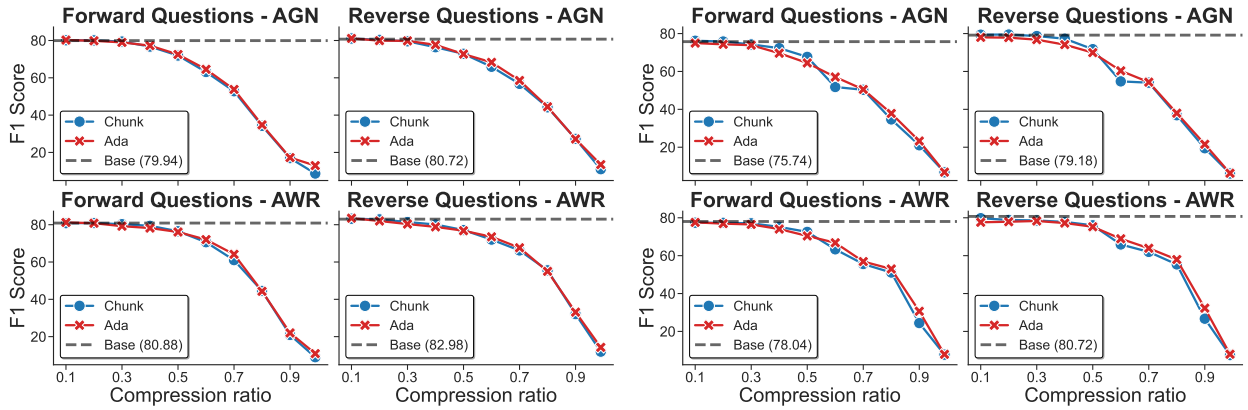
Structural Summary

(1) Mild compression can briefly improve accuracy, apparently by clearing away interfering routes and leaving a cleaner sparse one. (2) Robustness does not track retained token count alone: at comparable eviction rates, accuracy depends on whether redundant cross-head routes to the answer survive. (3) Multi-hop and long-range reasoning degrade at lower compression ratios than single-hop retrieval, and begin to fail while answer tokens are still largely retained (low GER); this indicates that relational routing fails before the underlying memory is removed, rather than the truism that connectivity is gone once tokens are.



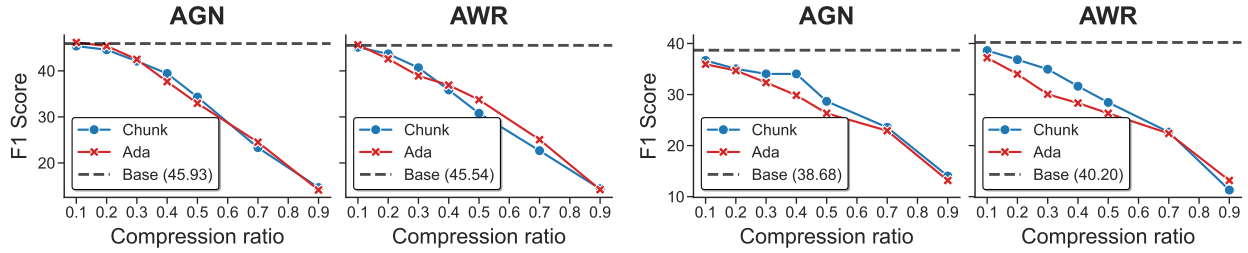
(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

Figure 5: Multi presence forward vs. reverse asymmetry.



(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

Figure 6: Multi entity results showing reduced directional asymmetry. Corresponding results for LLaMA-3.2 3B, Qwen-2.5 3B, and Qwen-2.5 14B are shown in Figure 28 of Appendix B.

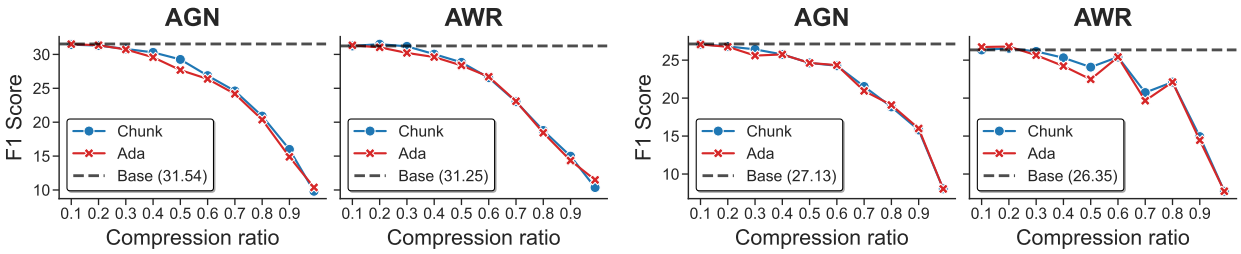


(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

Figure 7: Long context degradation across compression levels.

4.2 Tag-Level Analysis

Aggregate F1 shows that compression degrades performance, but the tags show *where*. Splitting accuracy by answer type and question type separates plain token loss from harder-to-spot weaknesses, and it turns



(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

Figure 8: Hops task performance. Multi-step semantic chaining collapses rapidly under compression.

out compression does not hit all categories equally: it tends to erode relational and morphologically varied answers first.

Answer-Type Tags. Figure 10 shows answer-type performance for the Base task. Although the overall F1 score remains near 70% at 0% compression, tag-level trends reveal systematic variation. Categories such as *Person* and *Location* are relatively stable across compression levels, complementing the similar results obtained by Liu et al. (2025a). These entities are typically realized as concrete nouns with consistent surface forms, enabling direct lexical anchoring in the KV cache. Even when pruning removes a subset of tokens, surviving mentions are often sufficient for correct retrieval.

Event answers drop off much faster. An event answer usually has to be matched across different verb forms and paraphrases, so it is spread over many tokens rather than sitting in a single noun. When compression fragments that spread-out encoding, the answer becomes hard to reassemble, and the Event curve turns down earlier than the noun-based categories.

The *Organization* tag reveals a distinct but related weakness. Organizational entities frequently encode nested relational structure, including sub-entities, roles, or hierarchical relationships. In the question-aware setup, both models show sharper declines for Organization-tagged answers compared to the question-agnostic condition. This suggests that query-conditioned pruning may collapse hierarchical distinctions into coarse labels. Under compression these fine-grained relational links are among the first to go. The pattern looks less like losing tokens and more like losing the structure that ties them together.

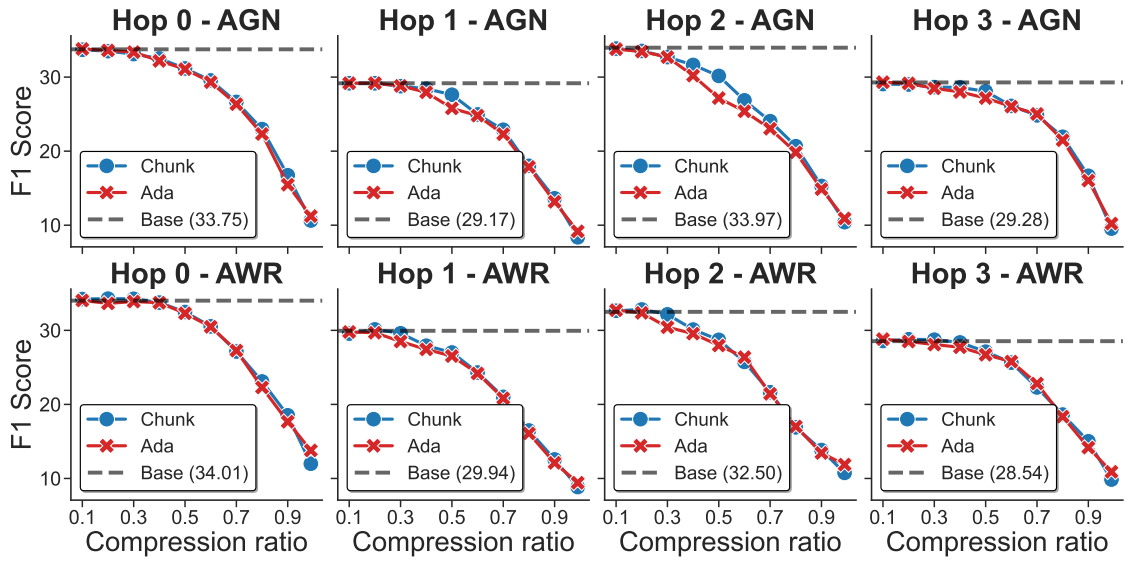
Multi-entity tag results reinforce this interpretation, as shown in Figure 11. The *Thing* and *Organization* tags display more irregular degradation curves, especially under moderate-to-high compression, while *Person* and *Location* remain comparatively stable. This asymmetry suggests that compression preferentially harms categories requiring cross-token relational binding rather than those represented by isolated lexical anchors.

Knowledge manipulation further clarifies architectural divergence. As shown in Figure 12, Qwen exhibits differentiated degradation across tags, whereas LLaMA shows more synchronized decline. This suggests that Qwen allocates representational capacity unevenly across semantic domains, preserving certain categories longer under compression, while LLaMA maintains more uniform retention but collapses more abruptly when compression becomes extreme.

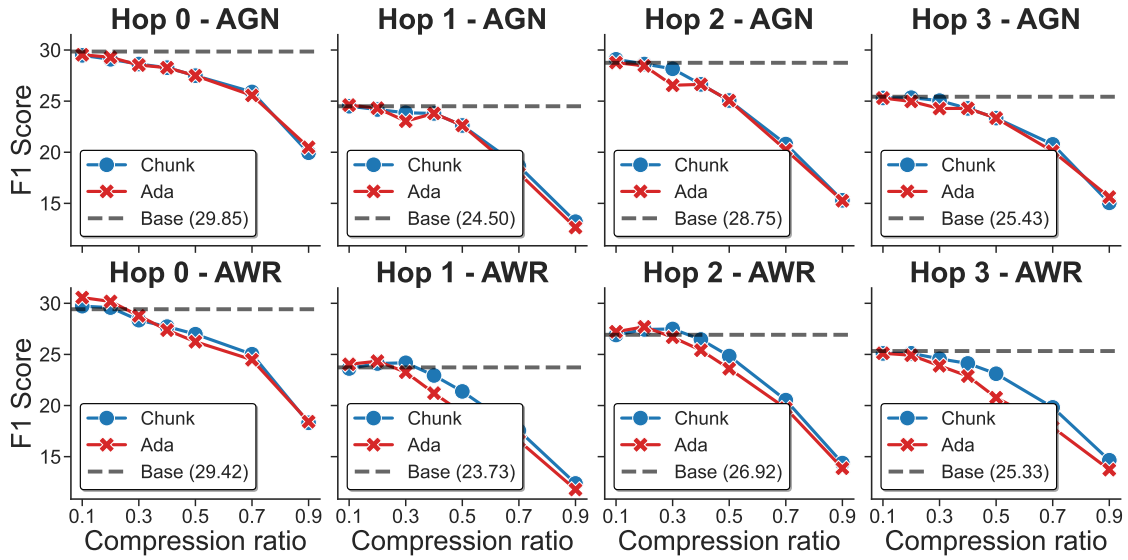
Question-Type Tags. Question-type tags reveal how compression interacts with cognitive demands. In the Base task (Figure 13), *Standard* retrieval questions are relatively stable until moderate compression. These tasks require direct extraction of explicitly stated facts and thus depend primarily on token survival.

Manipulated questions, which require transformation or reinterpretation of context, expose architectural differences. Qwen maintains closer alignment between AGN and AWR settings, indicating robust instruction-conditioning. LLaMA, while strong in raw retrieval, exhibits sharper divergence under AWR in manipulation-heavy tasks.

Part retrieval questions degrade earliest and most irregularly. These queries require integrating scattered phrases across different regions of the context. Even when individual tokens survive eviction, the relational



(a) LLaMA-3 8B Instruct



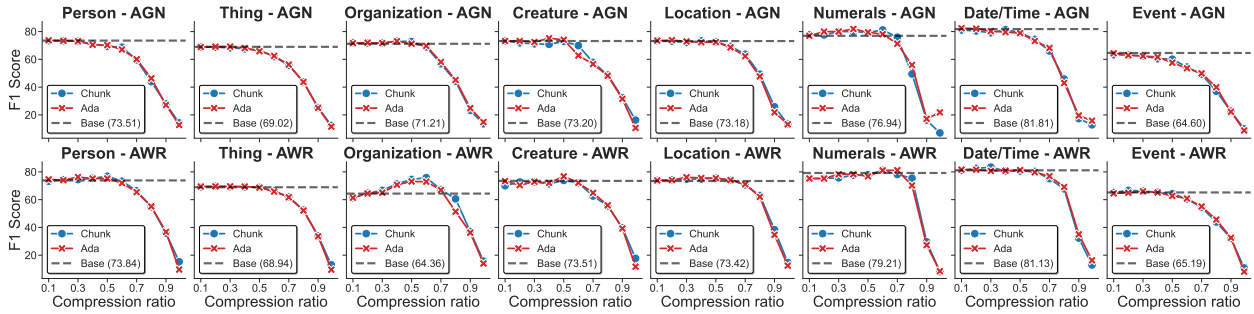
(b) Qwen-2.5 7B Instruct

Figure 9: Individual hop breakdown showing instability in intermediate semantic links.

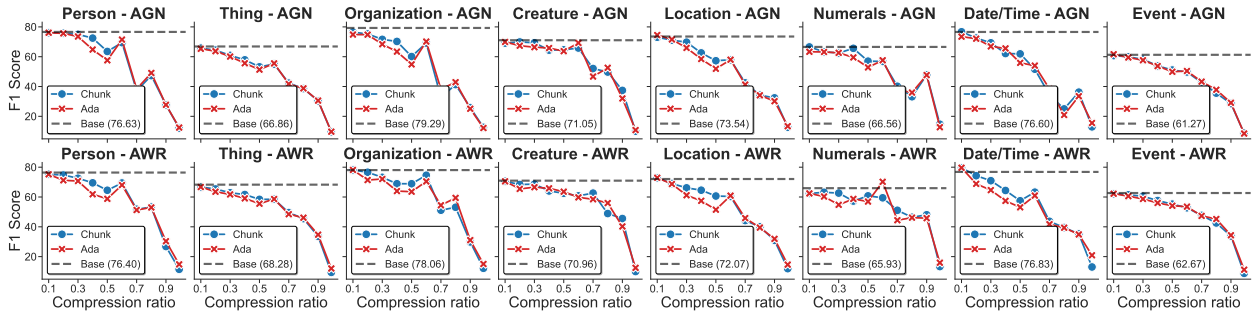
pathways connecting them become fragile. Performance curves for Part questions exhibit jagged transitions rather than smooth decline, indicating intermittent route failure rather than uniform capacity reduction.

Knowledge manipulation question tags (Figure 14) highlight Qwen's strength in *Manipulated* queries, where it outperforms LLaMA more consistently under compression. However, in pure *Standard* retrieval, LLaMA retains an advantage. This reinforces the interpretation that the two architectures allocate representational resources differently: one optimized for instruction-conditioned flexibility, the other for dense storage and direct recall.

In Multi entity (Figure 15), *Part* retrieval again exhibits the most instability. Performance oscillates across compression levels, indicating a threshold phenomenon where distributed semantic cues become intermittently unreachable. This instability aligns with the notion of representational rigidity: tokens remain present, but routing flexibility diminishes, preventing consistent multi-span integration.

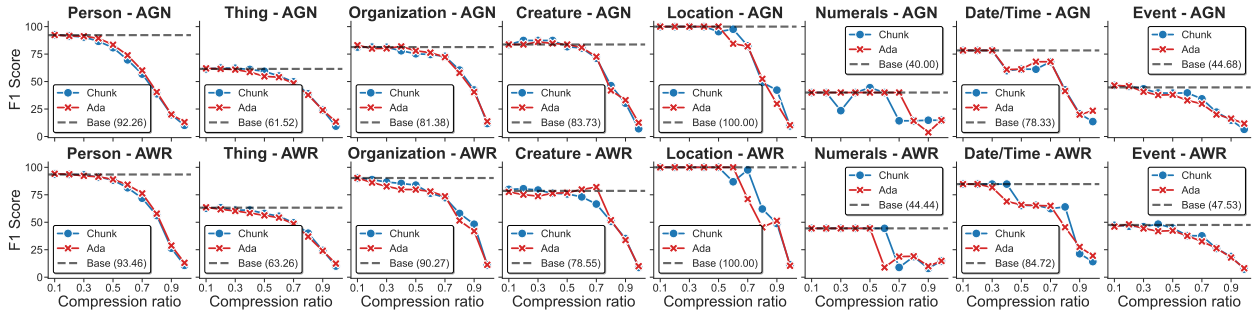


(a) LLaMA-3 8B Instruct

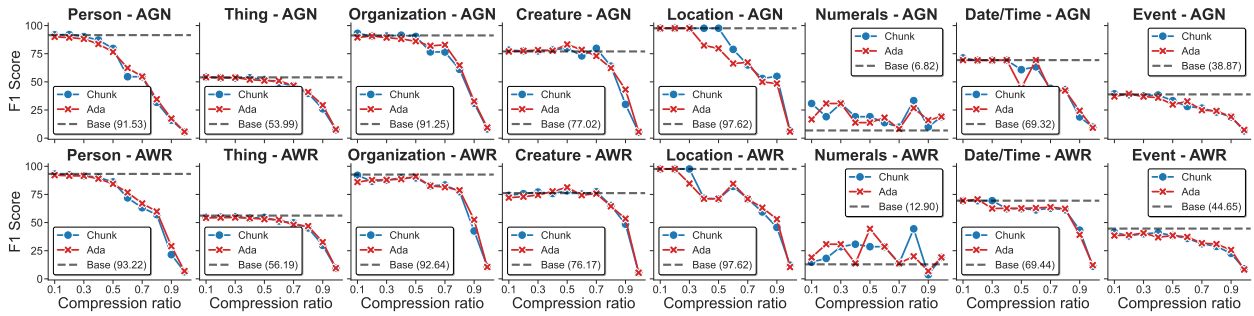


(b) Qwen-2.5 7B Instruct

Figure 10: Answer-type tag performance in Base task. Corresponding results for LLaMA-3.2 3B, Qwen-2.5 3B, and Qwen-2.5 14B are shown in Figure 29 of Appendix B.

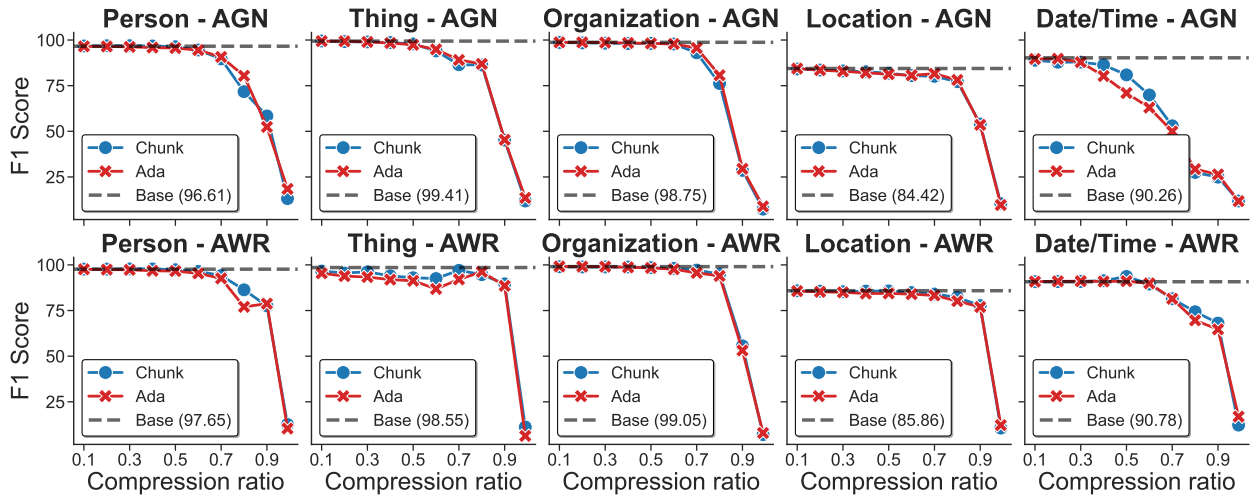


(a) LLaMA-3 8B Instruct

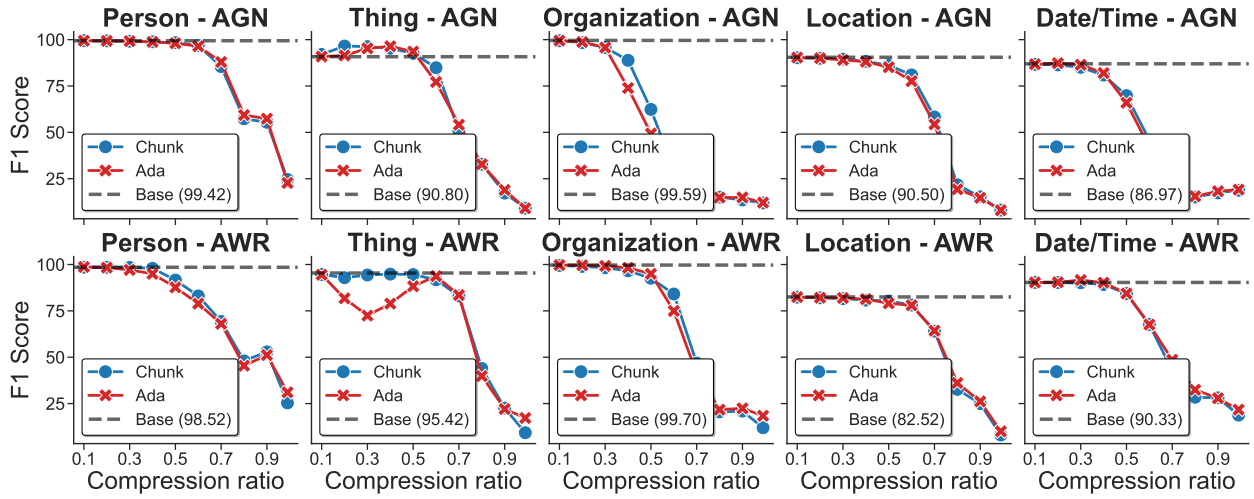


(b) Qwen-2.5 7B Instruct

Figure 11: Answer-type tag behavior in Multi entity. Corresponding results for LLaMA-3.2 3B, Qwen-2.5 3B, and Qwen-2.5 14B are shown in Figure 30 of Appendix B.

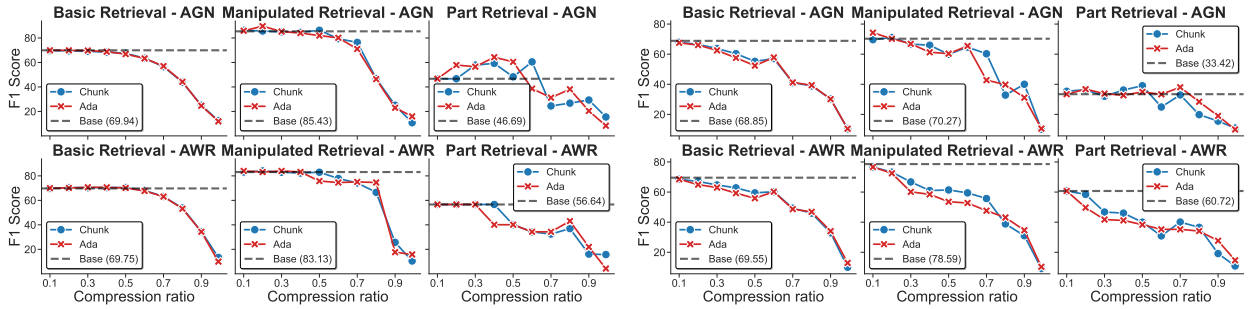


(a) LLaMA-3 8B Instruct



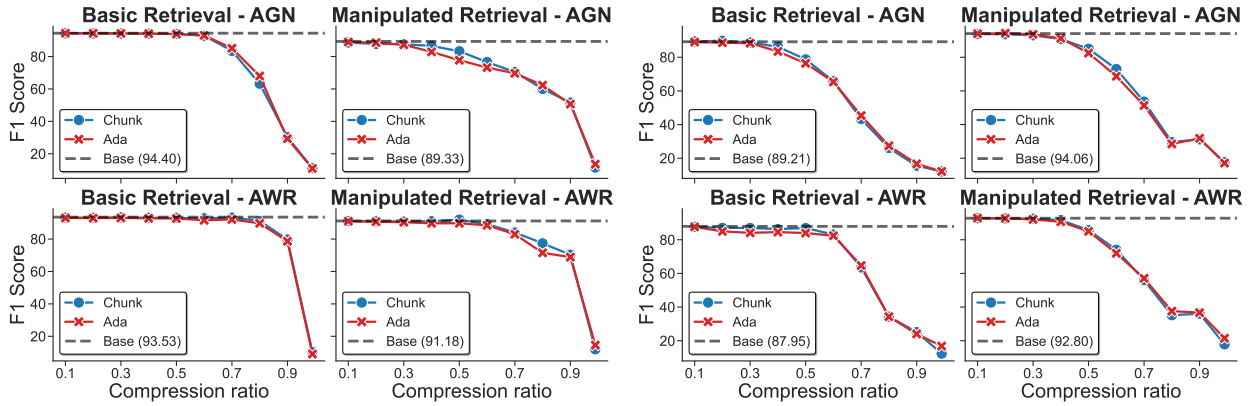
(b) Qwen-2.5 7B Instruct

Figure 12: Knowledge manipulation answer-type tag-level behavior.



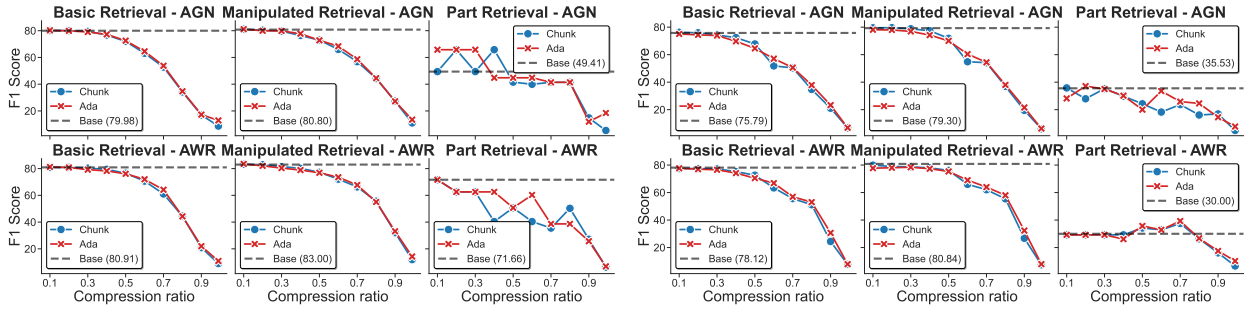
(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

Figure 13: Question-type tags for Base task. Corresponding results for LLaMA-3.2 3B, Qwen-2.5 3B, and Qwen-2.5 14B are shown in Figure 31 of Appendix B.



(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

Figure 14: Knowledge manipulation question-type analysis.



(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

Figure 15: Multi entity question-type behavior shows a similar trend to Base task but with much smoother curves in question agnostic. The question aware setup has a far more jagged curve with distinct points of significant drop off. Corresponding results for LLaMA-3.2 3B, Qwen-2.5 3B, and Qwen-2.5 14B are shown in Figure 32 of Appendix B.

Synthesis of Tag-Level Patterns. Two patterns hold up across both kinds of tags. Relational categories (*Event*, *Organization*, *Part*) degrade earlier than simple entity categories, and the degradation often shows up as jagged, unstable curves rather than a smooth slope, which fits a picture of routes dropping out one at a time rather than capacity shrinking evenly. Which categories a given model holds onto longest varies from model to model, but the ordering, relational before atomic, is consistent.

Overall, the tag results tell the same story as the aggregate curves: compression breaks the connections between tokens before it removes the tokens themselves, and relational categories are where this shows up first.

5 Discussion

We read the KV cache as more than storage: compression changes not just how much is kept, but which token routes still carry evidence to the decoder. Two things matter throughout. The first is *reachability*: whether an answer token can still be reached through some surviving head. The second is *adaptivity*: whether enough heads stay diverse for the model to re-route when the context is thinned out. The rest of this section is organized around these two, and connects them to how models use depth, to the safety cliff, and to the gap between what the model represents and what it actually does.

5.1 Structural Metrics for KV Compression

Compression can cause failure either by *erasing* necessary evidence or by leaving evidence present but *poorly used*. To separate these effects, we track *eviction* (survival) and *consensus* (coordination). Let T be the context token set and H the head set. For token t, compression level α, head h, and survived-token set $t^{(\alpha,h)}$, define

$$S_h(t,\alpha) = \begin{cases} 1 & \text{if } t \in t^{(\alpha,h)}, \\ 0 & \text{otherwise.} \end{cases} \qquad \text{GlobalEvicted}(t,\alpha) = \mathbb{1}\Big[\sum_{h\in H} S_h(t,\alpha) = 0\Big].$$

The Eviction Rate is

$$\text{EvictionRate}(\alpha) = \frac{1}{|T|}\sum_{t\in T} \text{GlobalEvicted}(t,\alpha),$$

and the task-aware Global Eviction Ratio (GER) restricts to answer-relevant tokens T_{ans}:

$$\text{GER}(\alpha) = \frac{1}{|T_{\text{ans}}|}\sum_{t\in T_{\text{ans}}} \text{GlobalEvicted}(t,\alpha).$$

Here T_{ans} is the set of context token positions spanning the gold answer, taken directly from the answer annotation produced during dataset construction (Section 3.1); for template tasks these positions are known exactly, and for generative tasks they are the span used for evaluation and probing. The distinction from a plain per-head eviction count lies in the GlobalEvicted operator: a token contributes to GER only when it is pruned in *every* head ($\sum_h S_h = 0$). Because a single surviving head suffices to route a token's value into the query-conditioned state, a token that any head can still read remains reachable. GER is therefore not simply "the fraction of answer tokens removed," but the fraction for which *all* attention pathways to the evidence have been severed simultaneously—a necessary condition for the model to lose contextual grounding, which is why it correlates with hallucination (Figure 24). Survival alone is insufficient; however, the model must also *coordinate* its remaining capacity. Let $A_{\ell,h}(t)$ be normalized attention weight from head h at layer ℓ to token t, with top-attended token

$$t^*_{\ell,h} = \arg\max_t A_{\ell,h}(t).$$

We measure layer-wise head consensus as

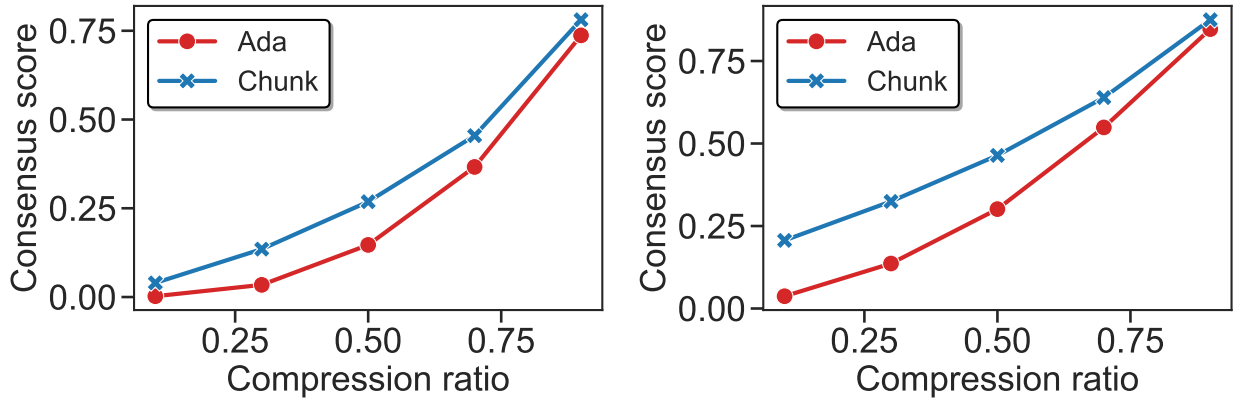
$$\text{Consensus}(\ell) = \frac{\big|\{t^*_{\ell,h}\}_{h\in H}\big|}{|H|},$$

where lower values indicate stronger agreement (many heads focus on the same token) and higher values indicate diversity. We stress that consensus measures *positional* routing diversity, not semantic equivalence across heads: two heads attending the same position need not assign it the same role, and we make no such claim. What the metric captures is how many *distinct* token positions a layer actively reads, and hence how many independent positional routes remain available if some of those positions are pruned. Low consensus thus signals route redundancy (many independent positions read), whereas high consensus signals that the layer's read is concentrated on a few positions, leaving little fallback when they are evicted. Together, eviction and consensus let us distinguish *representational erasure* (routes destroyed) from *representational rigidity* (routes present but insufficiently adaptable).

5.2 What does KV compression reveal about how different architectures allocate computation across depth?

That two model families behave differently is unsurprising on its own; the result we wish to isolate is *where* along depth the compression-sensitive routing sits, and whether that location is fixed. Layer-wise consensus trends show that it is not (Figure 16): the depth at which routing diversifies versus consolidates differs across the models we study—agreement concentrates earlier in LLaMA and later in Qwen—and shifts under increasing compression (Figures 17 and 18). Read mechanistically, these profiles correspond to different "decision depths": early-layer agreement can act like normalization that supports later specialization, while

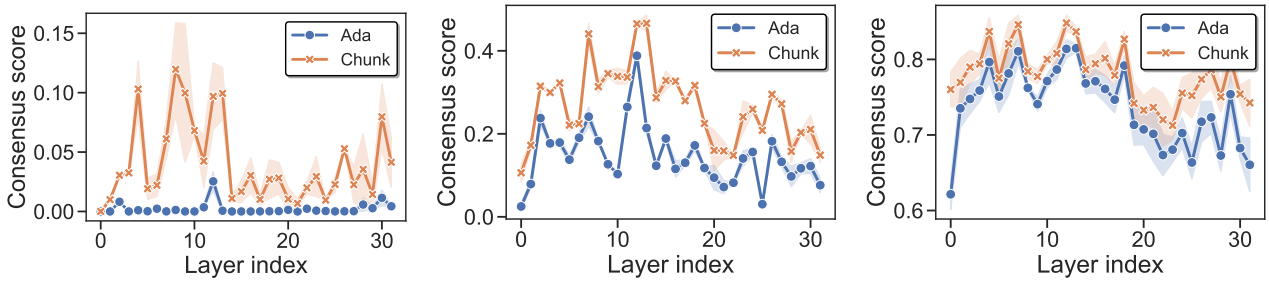
deep-layer consolidation can act like a late-stage decision cascade that concentrates routing onto a narrow token set. The actionable consequence, rather than the difference itself, is what matters for compression design.



(a) LLaMA-3 8B Instruct

(b) Qwen-2.5 7B Instruct

Figure 16: Consensus across layers. LLaMA exhibits early stabilization and late specialization; Qwen exhibits early exploration and late consolidation. Corresponding results for LLaMA-3.2 3B, Qwen-2.5 3B, and Qwen-2.5 14B are shown in Figure 33 of Appendix B.



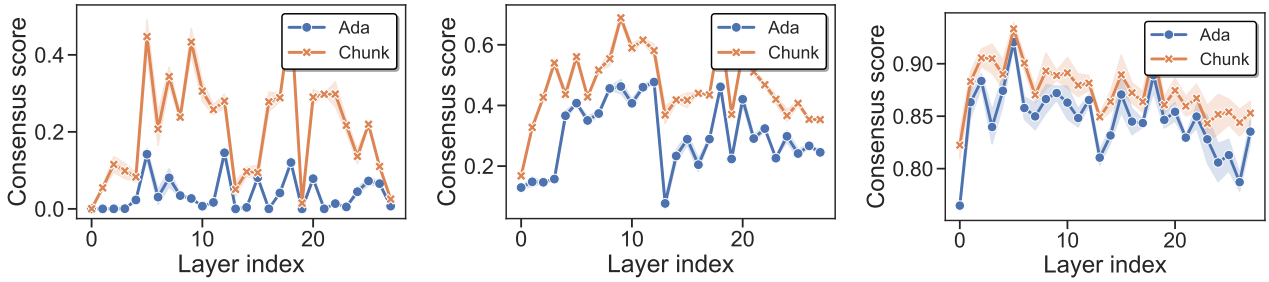
(a) LLaMA-3 8B Instruct 10% Compression

(b) LLaMA-3 8B Instruct 50% Compression

(c) LLaMA-3 8B Instruct 90% Compression

Figure 17: LLaMA layerwise consensus under compression. Early agreement persists, with depth-wise diversification supporting parallel feature channels. Corresponding results for LLaMA-3.2 3B are shown in Figure 34 of Appendix B.

This depth allocation has direct implications for compression design: policies that assume a universal "pyramid" (monotone entropy decrease with depth) may transfer poorly across families, since the fragile computation can occur at different depths. We treat the attention heatmaps (Figures 19 and 20) as qualitative and illustrative rather than as primary evidence for routing: they show that the two presses perturb token positions and disrupt the attention matrices differently at moderate compression (chunk-style contiguous vs. Ada-style non-contiguous removal) and that both converge to widespread removal at extreme compression, but a single heatmap cannot establish that a deleted position was on a *used* route. The quantitative support for the routing account therefore comes from the GER–hallucination correlation (Section 5.3, Figure 24) and the reachability argument of Section 6, with the heatmaps serving to visualize the pruning geometry that those metrics summarize.

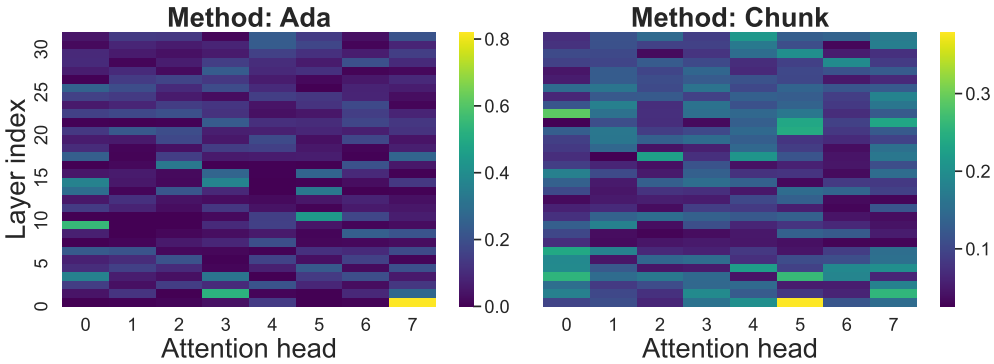


(a) Qwen-2.5 7B Instruct 10% Compression (b) Qwen-2.5 7B Instruct 50% Compression (c) Qwen-2.5 7B Instruct 90% Compression

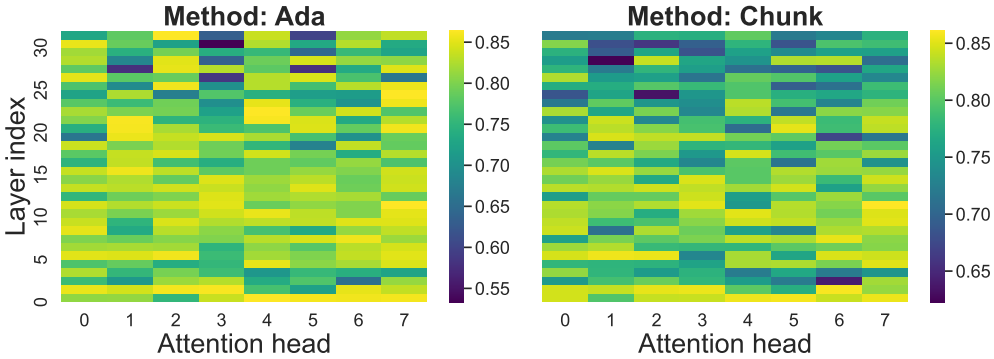
Figure 18: Qwen layerwise consensus under compression. Diversity persists deeper and collapses late, consistent with a consolidation-heavy tail. Corresponding results for Qwen-2.5 3B and Qwen-2.5 14B are shown in Figure 35 of Appendix B.

Observation: Compression must be architecture-aware

Figures 16–18 indicate that the layers most responsible for stabilization vs. consolidation differ by family, so the same "depth prior" for pruning can be safe for one model and brittle for another.



(a) 10% Compression



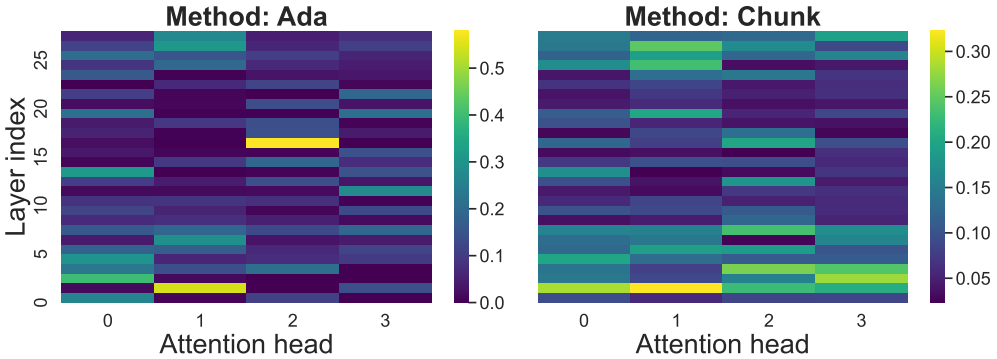
(b) 90% Compression

Figure 19: LLaMA attention heatmaps under compression. Chunk pruning removes contiguous segments; Ada pruning is more irregular. At 0.9, global route destruction dominates. Corresponding results for LLaMA-3.2 3B are shown in Figure 37 of Appendix B.

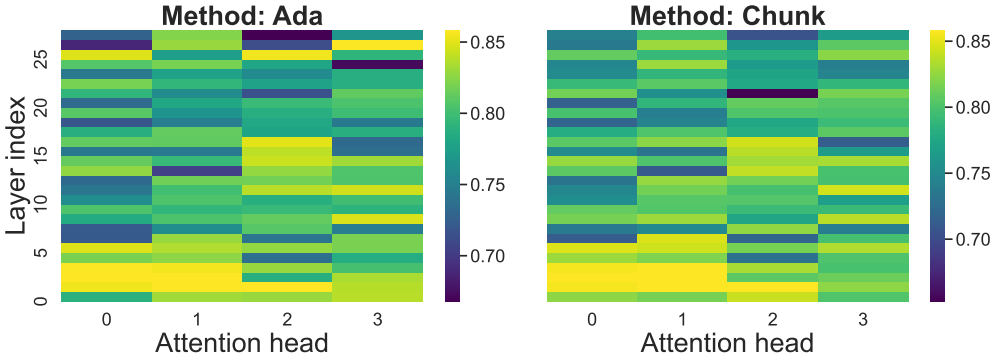
Algorithm 2 Linear Probing Under KV Compression

Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$; model f; compression α; press config Π; probed layers $\mathcal{L}$; train/val indices $\mathcal{I}_{tr}, \mathcal{I}_{va}$

1: **for** $i = 1$ to N **do**
2: Run f on x_i with KVPress (Π, α); record $\{h_{i,\ell}\}_{\ell \in \mathcal{L}}$
3: $z_i \leftarrow \mathrm{Pool}(\{h_{i,\ell}\}_{\ell \in \mathcal{L}})$
4: **end for**
5: Train linear $g(z) = \mathrm{softmax}(Wz + b)$ on $\{(z_i, y_i)\}_{i \in \mathcal{I}_{tr}}$ with f frozen
6: Evaluate macro-F1 on $\{(z_i, y_i)\}_{i \in \mathcal{I}_{va}}$



(a) 10% Compression



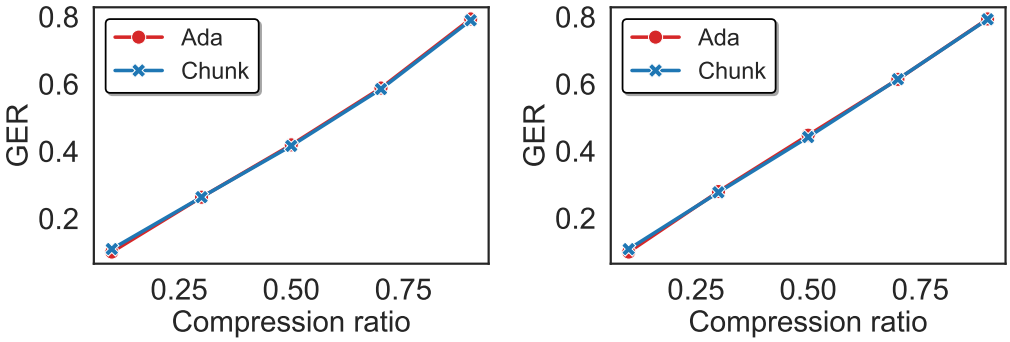
(b) 90% Compression

Figure 20: Qwen attention heatmaps under compression. Policy differences are visible at low compression but are overwhelmed at 0.9 by large-scale pruning. Corresponding results for Qwen-2.5 3B, and Qwen-2.5 14B are shown in Figure 38 and Figure 39 of Appendix B.

5.3 Why does "safe" KV compression suddenly fail at extreme eviction?

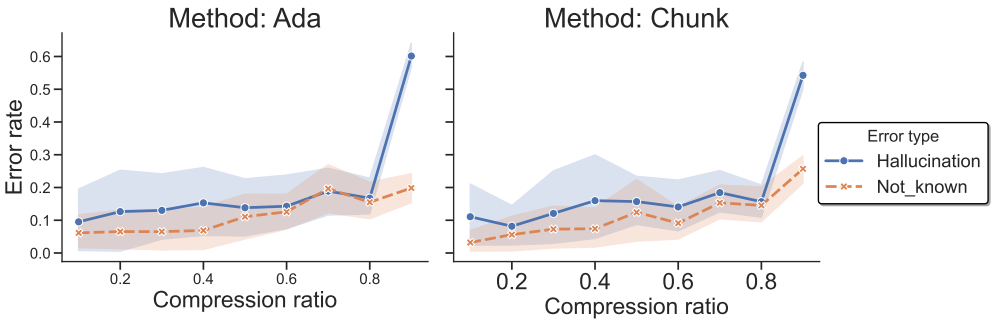
Every model and press we test shows the same safety cliff near $\alpha \approx 0.9$: hallucination jumps once about 90% of the KV entries are gone (Figure 22). What changes at the cliff is not a gradual loss of representation quality but a sudden rise in *global route deletion*, where answer tokens are evicted from every head at once and no path to the evidence is left. This is why the eviction rate can stay fairly flat across α (Figure 21) while the error rate turns sharply (Figure 22), and why GER tracks hallucination so closely (Figure 24).

At the same time, some failures occur even when GER is low, pointing to *rigidity*: if attention consolidates too strongly onto a narrow focal set, the model may not re-route effectively when that set is pruned or when the remaining evidence is weakly aligned. This is most plausible in architectures with strong late consolidation (Figures 16 and 18), where token survival can coexist with under-utilization.

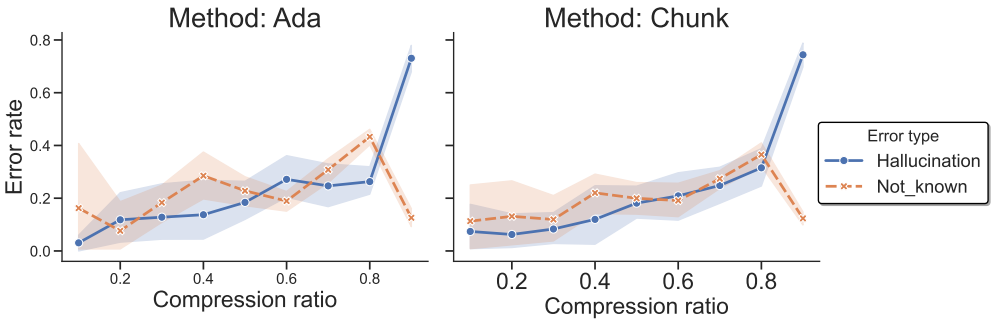


(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

Figure 21: Eviction rates vs. compression. Pruning is broadly question-agnostic unless explicitly conditioned. Corresponding results for LLaMA-3.2 3B, Qwen-2.5 3B, and Qwen-2.5 14B are shown in Figure 42 of Appendix B.



(a) LLaMA-3 8B Instruct

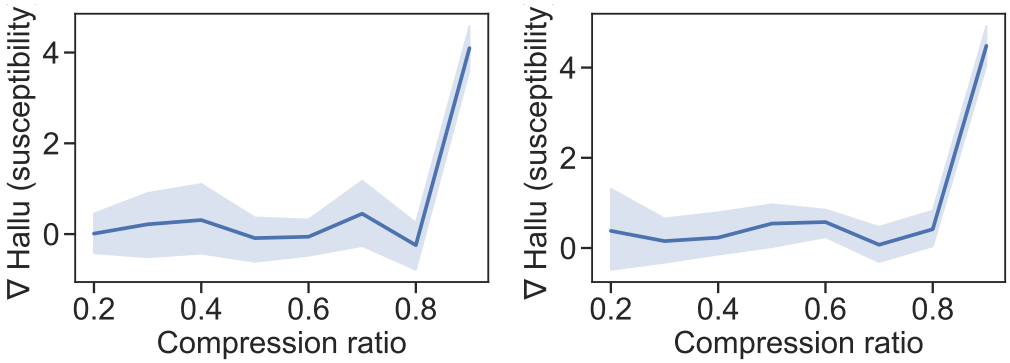


(b) Qwen-2.5 7B Instruct

Figure 22: Error rates vs. compression. Hallucination exhibits a cliff near $\alpha \approx 0.9$, with "unknown" behavior more erratic under AdaKV. Corresponding results for LLaMA-3.2 3B, Qwen-2.5 3B, and Qwen-2.5 14B are shown in Figure 43 of Appendix B.

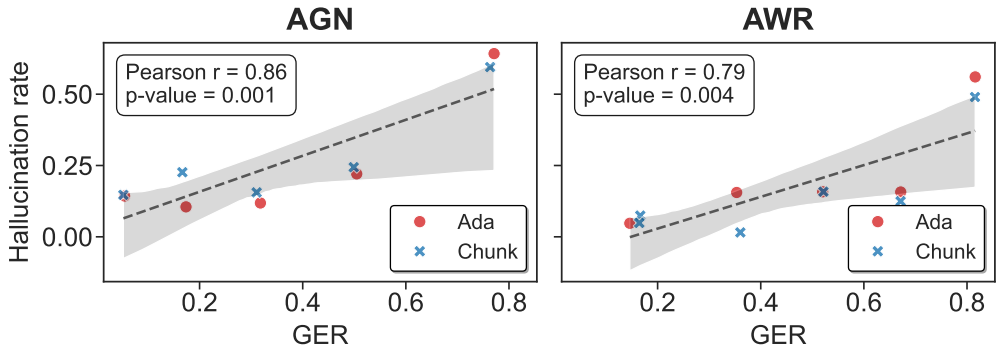
Observation: Two mechanisms explain the safety cliff

Figures 24 and 16 together support two failure modes: (i) **erasure** (high GER removes all evidence routes), and (ii) **rigidity** (routes survive but consolidation limits re-routing), both of which manifest sharply near $\alpha \approx 0.9$ (Figure 22).

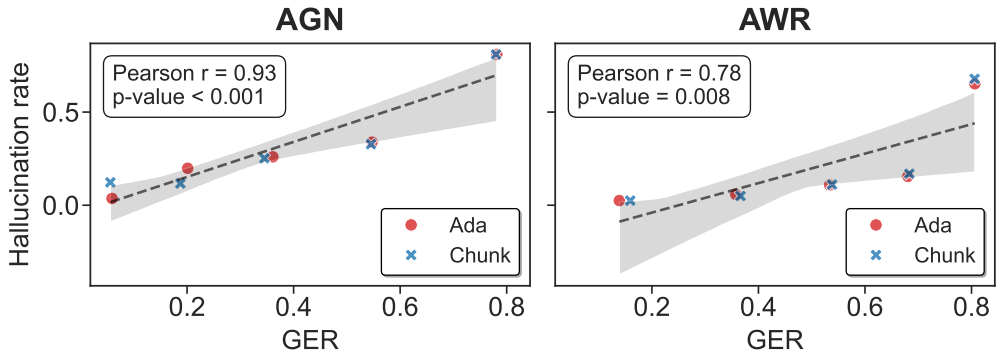


(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

Figure 23: Compression susceptibility $\chi = \frac{\partial H}{\partial \alpha}$ peaks sharply near $\alpha \approx 0.9$, consistent with a critical transition in the error landscape.



(a) LLaMA-3 8B Instruct



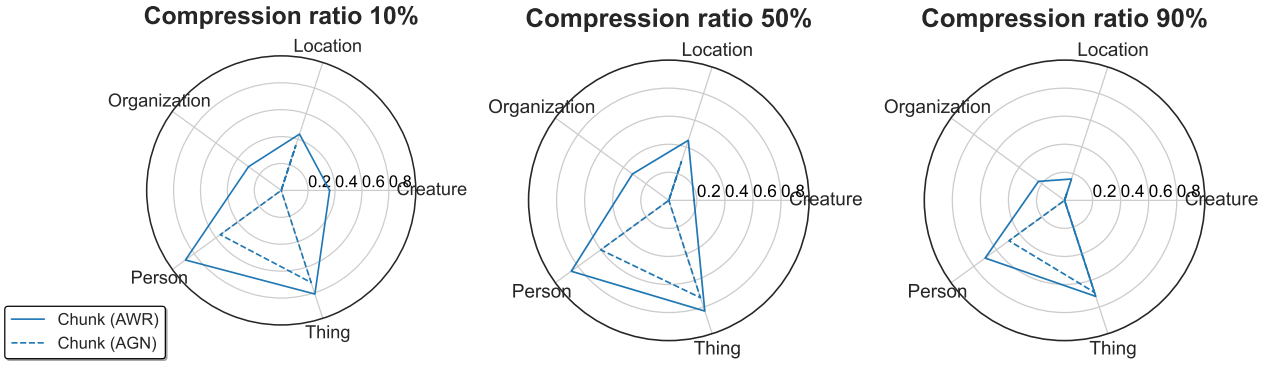
(b) Qwen-2.5 7B Instruct

Figure 24: GER is strongly associated with the hallucination rate, indicating that global route deletion is a primary driver of failure. Corresponding results for LLaMA-3.2 3B, Qwen-2.5 3B, and Qwen-2.5 14B are shown in Figure 44 of Appendix B.

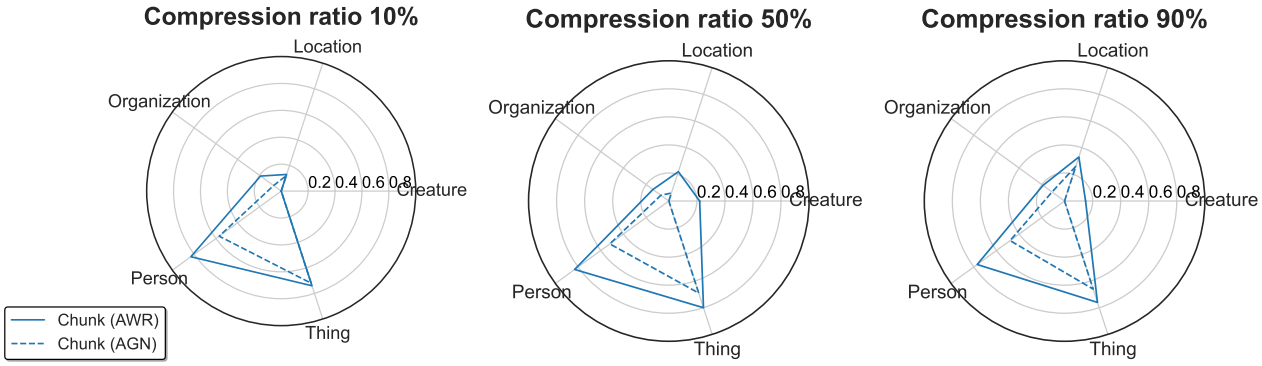
5.4 What survives compression inside the model, and why does survival not imply use?

Probing mechanism. To see what information is still readable after compression, we train linear probes on the model's frozen hidden states, using the same press settings as during generation. For each model, dataset, and ratio α, we cache the hidden states and fit a small linear classifier to predict the answer's concept tag (e.g., *Person*, *Location*, *Organization*), and report macro-F1 (Algorithm 2). This measures whether the

information is *present and readable*, not whether the model actually *uses* it: generation can still fail if routing collapses (high GER) or becomes too rigid, which is why probe scores can stay high while hallucination rises near the cliff (Figures 25 and 22).



(a) LLaMA-3 8B Instruct



(b) Qwen-2.5 7B Instruct

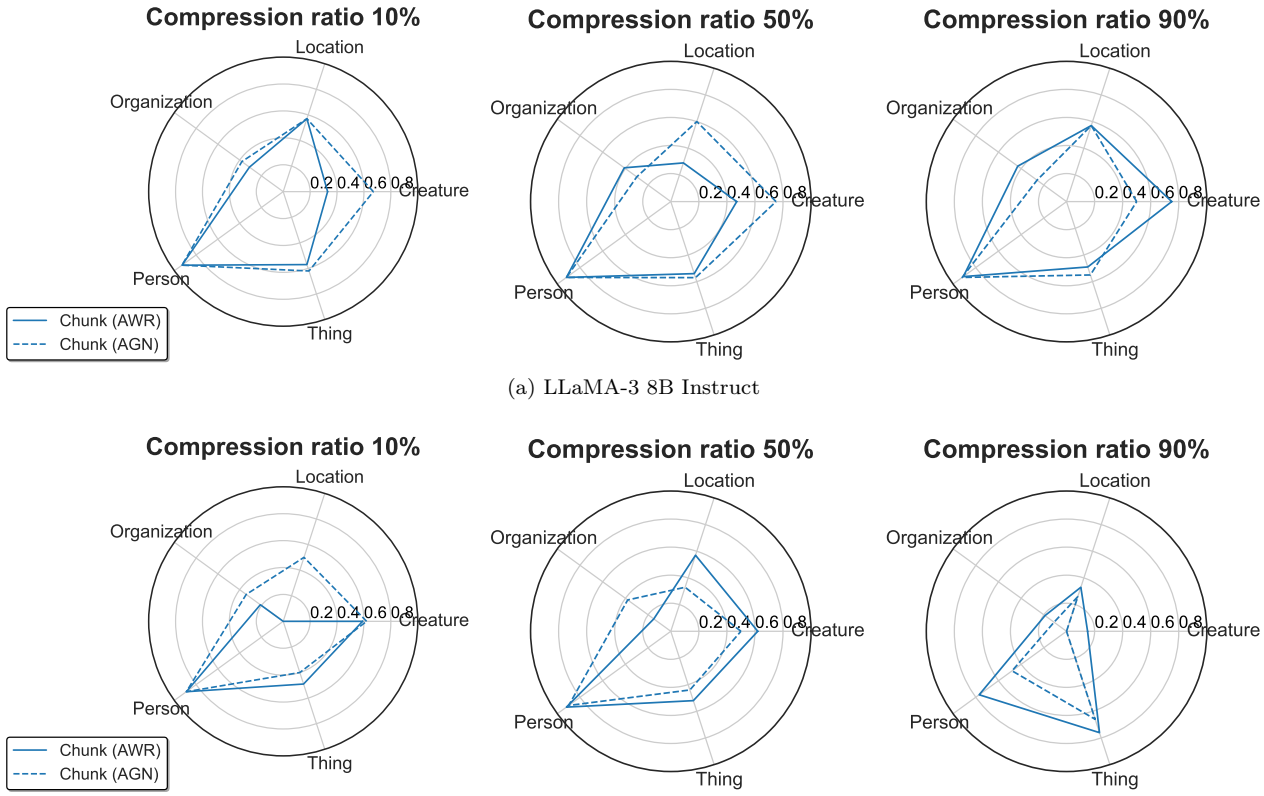
Figure 25: Base task probing across compression. Some tags remain decodable; others collapse early and may partially recover under question-aware setups. Corresponding results for LLaMA-3.2 3B, Qwen-2.5 3B, and Qwen-2.5 14B are shown in Figure 40 of Appendix B.

Observation: Retention ≠ utilization

Probing can remain high while generation fails: concept vectors may stay decodable (Figures 25–26) even as hallucination spikes near the cliff (Figure 22), indicating that routing availability and flexibility, not just representational presence, limit reliability.

5.5 When does question-aware compression help, and why does it help some models more than others?

Question-aware compression amounts to giving pruning an early hint about the answer: it keeps the tokens that best separate the right answer from the alternatives, which matches the cleaner probe separation we see in several cases (Figures 25 and 26). It helps most when deeper layers build on that early selection, as in the late-consolidating models (Figures 16 and 18). It helps least on tasks that depend on bridge tokens, such as multi-hop or discrepancy detection, where the useful token may not look like the query, so good question-aware pruning has to keep *routes*, not just query-similar tokens.



(a) LLaMA-3 8B Instruct

(b) Qwen-2.5 7B Instruct

Figure 26: Multi-Entity probing across compression. Question awareness helps some tags and not others, rather than uniformly. Corresponding results for LLaMA-3.2 3B, Qwen-2.5 3B, and Qwen-2.5 14B are shown in Figure 41 of Appendix B.

Observation: Query signals improve discriminative preservation

Question-aware pruning tends to preserve tokens that maintain decision boundaries for the correct answer, improving probe separability (Figures 25–26) and reducing failures near $\alpha \approx 0.9$ (Figure 22), especially in late-consolidating models (Figure 16).

5.6 Is KV compression best understood as gradual representational decay or as a phase transition?

Finally, probing and generation move differently, which argues against treating compression as smooth decay: probe scores can fall early (Figure 25) while generation holds steady until the cliff (Figure 22). A reachability-driven transition fits better: the model stays reliable as long as one usable route to the answer survives, and breaks when routes are globally deleted (high GER) or become too rigid to use (high consolidation), in line with the flat eviction trend (Figure 21) and the sharp change in error (Figure 23).

Observation

Figures 21–22 indicate that reliability depends on whether the token-route graph still contains viable paths to answer evidence (Figure 24) and whether routing remains sufficiently flexible (Figure 16), supporting a transition-like view of extreme KV compression.

6 Theoretical Implications of Token-Route Sparsity

6.1 From Parameter Lottery Tickets to Token-Route Subgraphs

Recent theory gives strong *parameter-level* sparsity guarantees for multi-head self-attention. In particular, the Strong Lottery Ticket Hypothesis (SLTH) for attention (Otsuka et al., 2025a) shows that, when an attention module is sufficiently wide, there exist sparse *parameter masks* whose restricted attention operator can approximate a target attention behavior without retraining. This is a statement about *weights*: a subset of parameters can implement (approximately) the same mapping as the dense module. KV cache compression is qualitatively different. It leaves the parameters untouched but removes *rows* of the key–value memory at inference time (i.e., removes tokens from the set that attention can read), thereby changing which token-to-token interactions are even possible. This motivates a complementary notion of sparsity defined on the *token connectivity* induced by attention during a forward pass.

Definition 1 (Token-Attention Graph). Fix an input sequence of length S and a transformer with layers $\ell \in \{0, \dots, L\}$ and (KV) heads $h \in \{1, \dots, H\}$. Let $V = \{(\ell, i) : 0 \le \ell \le L,\ 1 \le i \le S\}$ denote token-nodes indexed by layer and position, and let $A_{\ell,h} \in \mathbb{R}^{S\times S}$ be the (row-stochastic) attention matrix of head h at layer ℓ, with $A_{\ell,h}(i, j)$ the weight that destination position i places on source position j. Information propagates *forward* through two kinds of directed edges: attention edges, which carry the value of a source j at layer ℓ into a destination i at layer $\ell{+}1$ whenever some head attends to it above a threshold $\varepsilon > 0$, and residual edges, which carry each position forward to the next layer:

$$E \;=\; \big\{(\ell, j) \to (\ell{+}1, i) : \exists h,\ A_{\ell,h}(i, j) > \varepsilon\big\} \;\cup\; \big\{(\ell, i) \to (\ell{+}1, i)\big\}.$$

The directed layered graph $G = (V, E)$ records how token information can flow *across* layers, so that a multi-hop route corresponds to a path traversing several layers. The threshold $\varepsilon > 0$ is essential: at $\varepsilon = 0$ every softmax weight would induce an edge and reachability would be trivial, so G is read as the graph of non-negligible routes (we take ε to retain the bulk of the attention mass).

Definition 2 (Compressed Token-Attention Graph). At compression level α, a press selects, for each head h and layer ℓ, a subset of *surviving* key–value positions $U^{(\alpha)}_{\ell,h} \subseteq \{1, \dots, S\}$. The compressed graph $G_\alpha = (V, E_\alpha)$ retains every residual edge but keeps an attention edge only when some head that routes along it still retains the source:

$$E_\alpha \;=\; \big\{(\ell, j) \to (\ell{+}1, i) \in E :\ \exists h,\ j \in U^{(\alpha)}_{\ell,h} \wedge A_{\ell,h}(i, j) > \varepsilon\big\} \;\cup\; \big\{(\ell, i) \to (\ell{+}1, i)\big\}.$$

Equivalently, G_α is induced by zeroing all pruned columns $j \notin U^{(\alpha)}_{\ell,h}$ (and renormalizing) in each $A_{\ell,h}$.

Definition 3 (Token-Route Lottery Ticket, TR-LT). Let q denote the output (generation) position and let $T_{\text{ans}} \subseteq \{1, \dots, S\}$ be the answer-relevant context positions.[3] A TR-LT at compression α is a subgraph $H_\alpha \subseteq G_\alpha$ such that (i) some $t \in T_{\text{ans}}$ has a directed path to the output node (L, q) within H_α, and (ii) routing restricted to the edges of H_α already preserves the model's correct output on that instance. Intuitively, TR-LTs are *token-level* sparse routing backbones that remain functional under compression, complementing SLTH's *parameter-level* sparse backbones.

6.2 Reachability and the Erasure Failure Mode

The central structural question is whether answer evidence can still flow to the output through the compressed graph. Let $R_\alpha(q)$ be the set of context positions t from which the output node (L, q) is reachable in G_α, and define the reachability event

$$\mathcal{C}_\alpha(q) \;:=\; \big(T_{\text{ans}} \cap R_\alpha(q) \neq \emptyset\big).$$

If $\mathcal{C}_\alpha(q)$ fails, no answer-relevant token can route into the output, so the model cannot condition its generation on answer evidence through self-attention. When evidence is *redundantly* accessible—several head-disjoint

[3] In our experiments, T_{ans} is obtained from the gold answer span / annotation used for evaluation and probing.

routes carry the same answer token—moderate compression rarely severs all of them simultaneously, which is the structural reason accuracy is largely preserved well before the cliff; the sharp failure appears only once this redundancy is exhausted.

Proposition (Erasure implies loss of contextual grounding). Assume the model has not already copied the answer content into surviving positions *before* the KV entries for T_{ans} are removed.[4] If there exists $\alpha^\star$ with $T_{\text{ans}} \cap R_{\alpha^\star}(q) = \emptyset$, then no self-attention routing can condition the model's output on answer evidence at $\alpha^\star$, so correct generation cannot be grounded in context and the model must fall back on parametric priors, making hallucination structurally likely.

Proof sketch. Every head output is a weighted combination of *surviving* value vectors. If no answer token can reach (L, q) in $G_{\alpha^\star}$, then no sequence of attention aggregations introduces a value vector originating from T_{ans} into the output state; under the no-leakage assumption the decoder has no contextual evidence and defaults to priors. □

This is precisely what GER operationalizes: it estimates the prevalence of the erasure event $T_{\text{ans}} \cap R_\alpha(q) = \emptyset$ at the evidence level, rather than counting how many tokens remain overall. That is why GER, and not the raw eviction rate, tracks the hallucination cliff (Figures 21, 22, 24).

6.3 The Rigidity Failure Mode

Reachability is necessary but not sufficient. We also observe failures where answer tokens survive (low GER) yet accuracy still drops, which we attribute to *rigidity*. Let $\rho_\ell = \max_t \frac{1}{H}\big|\{h : t^*_{\ell,h} = t\}\big|$ measure how strongly heads at layer ℓ agree on a single top-attended position $t^*_{\ell,h} = \arg\max_t A_{\ell,h}(t)$. When ρ_ℓ is high, a large fraction of heads read the same few positions; pruning those positions forces all of them to re-route at once, and if the surviving attention mass does not cover T_{ans} the layer cannot recover the evidence even though it remains in memory. Head-level consensus (Figure 16) is our measurable proxy for this effect and is most pronounced in consolidation-heavy regimes. We treat this as a qualitative mechanism rather than a quantitative bound: together, erasure (high GER) and rigidity (low routing diversity) describe the two ways a compressed graph can stop supporting correct generation.

6.4 Sparse Token-Routes as the Inference-Time Counterpart of SLTH

SLTH (Otsuka et al., 2025a) implies that attention contains *sparse parameter* subnetworks capable of approximating dense attention behavior. Our results point to a complementary inference-time phenomenon: robustness under KV compression is governed by the survival of sparse *token-route* subnetworks (TR-LTs) within the compressed token-attention graph. In this view, KV compression is a controlled perturbation that reveals the minimal routing backbone required for correct generation, and it induces failure when either cross-head reachability collapses (the erasure mode, high GER) or routing becomes too rigid to exploit surviving evidence (the rigidity mode, low routing diversity). It is therefore not the *amount* of KV memory retained that governs reliability, but whether the compressed routing graph preserves at least one viable, usable token-route to answer evidence.

7 Limitations and Future Work

Our framework gives a mechanistic view of KV compression, but it has clear limits and openings. The synthetic datasets are built to isolate routing-sensitive behavior such as multi-hop chaining, entity disambiguation, and coreference, which is what makes the failure analysis clean, but it also means they do not capture the full messiness of natural text. An obvious next step is to check whether the same reachability–rigidity split holds on long-document QA, retrieval-augmented generation, code, and multimodal inputs. Systems with external memory or tools may also show forms of routing resilience that internal KV pruning alone does not.

[4] This rules out degenerate cases where answer information is perfectly "copied" into other surviving tokens prior to eviction; empirically, the GER–hallucination correlation (Figure 24) suggests such leakage is limited at the safety cliff.

Our theory leans on two simple metrics, the Global Eviction Ratio and head-level consensus. They line up well with where performance collapses, but a closer analysis of attention under structured token removal could give tighter results, for instance through spectral properties of the routing graph or bounds on cross-head redundancy. A more ambitious direction is to design for route capacity rather than prune after the fact: training objectives could encourage redundant cross-head routes or a controlled level of consensus, so that the routes needed to reach the answer survive compression by construction.

8 Conclusion

We treated KV cache compression as a change to how self-attention routes information, not just as a way to save memory. Using controlled synthetic data, layer-wise routing analysis, and two simple metrics for reachability and rigidity, we showed that compression fails in two ways: the answer tokens are dropped from every head (erasure), or they survive but the heads agree too strongly to re-route to them (rigidity). The sharp cliff at high compression is the point where reachable routes run out, and it shows that a small token-route backbone, not the bulk of the cache, decides whether the model still answers from context. This carries the lottery-ticket idea from network weights over to the token routes used at inference time. The practical message is that long-context and efficient-attention systems should try to preserve a minimal set of routes across heads and layers, not just keep token representations around.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, 2023.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction, 2024. URL https://arxiv.org/abs/2309.14316.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 1, learning hierarchical language structures. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=mPQKyzkA1K.

Yu Bai, Xiyuan Zou, Heyan Huang, Sanxing Chen, Marc-Antoine Rondeau, Yang Gao, and Jackie CK Cheung. CItruS: Chunked instruction-aware state eviction for long sequence modeling. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5908–5930, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.338. URL https://aclanthology.org/2024.emnlp-main.338/.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer, December 2020. URL http://arxiv.org/abs/2004.05150. Number: arXiv:2004.05150 arXiv:2004.05150 [cs].

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. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=ayi7qezU87.

Vivek Chari and Benjamin Van Durme. Compactor: Calibrated query-agnostic kv cache compression with approximate leverage scores, 2025. URL https://arxiv.org/abs/2507.08143.

Rewon Child. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, 2019.

Giulio Corallo and Paolo Papotti. Finch: Prompt-guided key-value cache compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1517–1532, 11 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00716. URL https://doi.org/10.1162/tacl_a_00716.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mZn2Xyh9Ec.

Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Re. Flashattention: Fast and memory-efficient exact attention with IO-awareness. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=H4DqfPSibmx.

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.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Ibrahim Talha Ersoy, Andrés Fernando Cardozo Licha, and Karoline Wiesner. Phase transitions reveal hierarchical structure in deep neural networks. *arXiv preprint arXiv:2512.11866*, 2025.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S. Kevin Zhou. Ada-KV: Optimizing KV Cache Eviction by Adaptive Budget Allocation for Efficient LLM Inference, January 2025. URL http://arxiv.org/abs/2407.11550. arXiv:2407.11550 [cs].

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uNrFpDPMyo.

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.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

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. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 30016–30030, 2022.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W Mahoney, Yakun S Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *Advances in Neural Information Processing Systems*, 37:1270–1303, 2024.

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

Zhongzhan Huang, Senwei Liang, Mingfu Liang, Wei He, Haizhao Yang, and Liang Lin. The lottery ticket hypothesis for self-attention in convolutional neural network, 2022. URL https://arxiv.org/abs/2207.07858.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan Arık. Long-context llms meet rag: Overcoming challenges for long inputs in rag. In *13th International Conference on Learning Representations, ICLR 2025*, pp. 65721–65759. International Conference on Learning Representations, ICLR, 2025.

Hao Kang, Qingru Zhang, Souvik Kundu, Geonhwa Jeong, Zaoxing Liu, Tushar Krishna, and Tuo Zhao. Gear: An efficient kv cache compression recipe for near-lossless generative inference of llm. *arXiv preprint arXiv:2403.05527*, 2024.

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.

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.

Hunjae Lee. Understanding the failure modes of transformers through the lens of graph neural networks. *arXiv preprint arXiv:2512.09182*, 2025.

Haoyang Li, Yiming Li, Anxin Tian, Tianhao Tang, Zhanchao Xu, Xuejia Chen, Nicole Hu, Wei Dong, Qing Li, and Lei Chen. A survey on large language model acceleration based on kv cache management. *arXiv preprint arXiv:2412.19442*, 2024a.

Yuankai Li, Jia-Chen Gu, Di Wu, Kai-Wei Chang, and Nanyun Peng. Brief: Bridging retrieval and inference for multi-hop reasoning via compression. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5449–5470, 2025.

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, June 2024b. URL http://arxiv.org/abs/2404.14469. arXiv:2404.14469 [cs].

Guangda Liu, Chengwei Li, Jieru Zhao, Chenqi Zhang, and Minyi Guo. Clusterkv: Manipulating llm kv cache in semantic space for recallable compression. In *2025 62nd ACM/IEEE Design Automation Conference (DAC)*, pp. 1–7. IEEE, 2025a.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, et al. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*, 2025b.

Xiang Liu, Zhenheng Tang, Peijie Dong, Zeyu Li, Bo Li, Xuming Hu, and Xiaowen Chu. Chunkkv: Semantic-preserving kv cache compression for efficient long-context llm inference. *arXiv preprint arXiv:2502.00299*, 2025c.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=JZfg6wGi6g.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. In *International Conference on Machine Learning*, pp. 32332–32344. PMLR, 2024.

Ziyang Ma, Zuchao Li, Lefei Zhang, Gui-Song Xia, Bo Du, Liangpei Zhang, and Dacheng Tao. Phase transitions in large language model compression. *npj Artificial Intelligence*, 2(1):21, 2026.

William Meng, Benjamin Lee, and Hong Wang. Understanding bottlenecks for efficiently serving llm inference with kv offloading. *arXiv preprint arXiv:2601.19910*, 2026.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. *Journal of Machine Learning Research*, 26(53):1–66, 2025. URL http://jmlr.org/papers/v26/24-1000.html.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Hikari Otsuka, Daiki Chijiwa, Yasuyuki Okoshi, Daichi Fujiki, Susumu Takeuchi, and Masato Motomura. The strong lottery ticket hypothesis for multi-head attention mechanisms. *arXiv preprint arXiv:2511.04217*, 2025a.

Hikari Otsuka, Yasuyuki Okoshi, Daichi Fujiki, Susumu Takeuchi, Masato Motomura, and Daiki Chijiwa. On the existence of hidden subnetworks within a randomly weighted multi-head attention mechanism. In *High-dimensional Learning Dynamics 2025*, 2025b. URL https://openreview.net/forum?id=gB1LZhA8Oy.

Ziran Qin, Yuchen Cao, Mingbao Lin, Wen Hu, Shixuan Fan, Ke Cheng, Weiyao Lin, and Jianguo Li. CAKE: Cascading and adaptive KV cache eviction with layer preferences. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EQgEMAD4kv.

Ayan Sengupta, Siddhant Chaudhary, and Tanmoy Chakraborty. Value-guided KV compression for LLMs via approximated CUR decomposition. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=klmc4fwPLd.

Noam Shazeer. Fast transformer decoding: One write-head is all you need, 2019. URL https://arxiv.org/abs/1911.02150.

Yutao Sun, Zhenyu Li, Yike Zhang, Tengyu Pan, Bowen Dong, Yuyi Guo, and Jianyong Wang. Efficient attention mechanisms for large language models: A survey, 2026. URL https://arxiv.org/abs/2507.19595.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, December 2017. URL http://arxiv.org/abs/1706.03762. Number: arXiv:1706.03762 arXiv:1706.03762 [cs].

Elena Voita, Rico Sennrich, and Ivan Titov. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. *arXiv preprint arXiv:1909.01380*, 2019.

Zhongwei Wan, Xinjian Wu, Yu Zhang, Yi Xin, Chaofan Tao, Zhihong Zhu, Xin Wang, Siqi Luo, Jing Xiong, Longyue Wang, and Mi Zhang. D_{2}O: Dynamic discriminative operations for efficient long-context inference of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=HzBfoUdjHt.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient Streaming Language Models with Attention Sinks, April 2024. URL http://arxiv.org/abs/2309.17453. arXiv:2309.17453 [cs].

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. Qwen2.5-1m technical report, 2025. URL https://arxiv.org/abs/2501.15383.

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, 2024b. URL https://arxiv.org/abs/2402.18096.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.

Minwei Zhang, Haifeng Sun, Jingyu Wang, Shaolong Li, Wanyi Ning, Qi Qi, Zirui Zhuang, and Jianxin Liao. ClusterAttn: KV cache compression under intrinsic attention clustering. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14451–14473, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.703. URL https://aclanthology.org/2025.acl-long.703/.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞Bench: Extending long context evaluation beyond 100K tokens. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.814.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H$_2$O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models, December 2023. URL http://arxiv.org/abs/2306.14048. arXiv:2306.14048 [cs].

Xiabin Zhou, Wenbin Wang, Minyan Zeng, Jiaxian Guo, Xuebo Liu, Li Shen, Min Zhang, and Liang Ding. DynamicKV: Task-aware adaptive KV cache compression for long context LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 8042–8057, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.426. URL https://aclanthology.org/2025.findings-emnlp.426/.

Jiaying Zhu, Dong Li, Xueyang Fu, Gege Shi, Jie Xiao, Aiping Liu, and Zheng-Jun Zha. A lottery ticket hypothesis approach with sparse fine-tuning and mae for image forgery detection and localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10968–10976, 2025.

A Dataset Descriptions

Task	Context	Question	Response
Base task	Gonzalo Batistuta, an Argentine forward, is widely regarded as one of the greatest football players of all time...	What is Gonzalo Batistuta's nationality?	Argentine
Knowledge man.	Cora Delaine was born on December 11, 1954. They studied in Daegu Global Science University...	What is Cora Delaine's first name?	Cora
Multi presence	The Astralis Spire stands tall as one of the most remarkable achievements of modern architecture and engineering. Rising high above the skyline of its bustling city...	What was the main reason for constructing Astralis Spire?	To improve television and radio broadcasting disrupted by urban development
Multi entity	Harland Kane is one of the most influential authors of contemporary horror, suspense, and supernatural fiction. Over a career spanning decades, Kane has...	Which novel by Kane is set in a haunted hotel?	The Shadow of the Pines
Long context	During the mid-Cretaceous period, the Riverback Sailbacks were formidable semi-aquatic predators, inhabiting river systems, floodplains, and coastal wetlands. Known for their elongated jaws, conical teeth, and large...	Where was Kaelin Vireo from?	The riverine fossil station of Thaloris
Coreference	Julian Foster was born on September 7, 1962 in Aleppo, Aleppo Governorate, Syria. He is male. He studied in Tianjin Harbor University in Podgorica, Podgorica, Montenegro...	What does he work as?	Tour Guide
Hops	Long before colonial records traced the legends of the high Mexican plateau, the City of Teotilac rose from volcanic plains as a geometric marvel of light, shadow, and faith.Built by the enigmatic Itzaca people, Teotilac's avenues and pyramids were laid out to capture the solar zenith, with precise alignments connecting celestial motion to civic life...	What was the City of Teotilac, and what made it distinctive? (Hop 0)	A sacred city built by the Itzaca, aligned with celestial cycles (Hop 0)
		How did Teotilac's design integrate cosmic and civic order? (Hop 1)	Its pyramids and avenues encoded solar and temporal order (Hop 1)
		What was the Prism of Yulnah project? (Hop 2)	A modern optical archaeology project at Teotilac (Hop 2)
		Who led the Prism of Yulnah and what modern tools were employed? (Hop 3)	Led by Dr. Lira Montoya using spectral drones and mirrors (Hop 3)

Table 3: Sample contexts and inputs for each task along with the ground truth answer.

Tag	Context	Question	Response
Person	Joao Fernandes, widely considered one of the greatest footballers of all time...	What is the name of the person described in the passage?	Joao Fernandes
Thing	Gonzalo Batistuta, an Argentine forward, is widely regarded as one of the greatest...	What is Gonzalo Batistuta's nationality?	Argentine
Organization	Mookie Blaylock, a rock band from Northport formed in Riverbend in 1990...	What is the subject of the given passage?	Mookie Blaylock
Creature	The Void Stalker is a fictional endoparasitoid extraterrestrial species that serves as the main...	What kind of species is the Void Stalker?	An endoparasitoid extraterrestrial species
Location	The Stones of Pharaohs are a complex of monumental structures located on the Giza Plateau...	In which city are The Stones of Pharaohs located?	Cairo
Numerals	The Thistlewood Vale Witch Trials represent a chilling episode in early American history...	How many individuals were convicted and executed during the trials?	20
Dates/Times	Gonzalo Batistuta, an Argentine forward, is widely regarded as one of the greatest...	When did Gonzalo Batistuta win the world championship?	2022
Events	During a time of great peril for his people, a child named Jabez was born. To save him from a decree of...	How did Jabez's mother save him from the decree of death?	His mother placed him in a woven basket and set him adrift on the river

Table 4: Examples of answer-type tags.

Tag	Context	Question	Response
Standard	The Basalt Bloom is an immense igneous province, covering a vast expanse of central India. It formed through a series of massive volcanic eruptions...	Where is the Basalt Bloom located?	Central India
Manipulated	Xena Bell was born on November 20, 1968. They studied in Shanghai Maritime University...	What is Xena Bell's last name?	Bell
Part	Breezeflare is characterized by unusual sounds during breathing caused by airflow obstruction in the airways. The condition is commonly observed...	Nyla Soren is recognized for leading what type of initiatives?	Community workshops, public education

Table 5: Examples of question-type tags

Task	Direction	Context	Question	Response
Mutli presence	FWD	The series known as Chrono Nexus is widely regarded as one of the most compelling science fiction narratives in Japanese animation. Created by Itsuki Aoyama...	Who created Chrono Nexus?	Itsuki Aoyama
	REV	Few creatures in Earth's history capture the imagination quite like the Thunderfang Predator, one of the largest and most fearsome carnivorous dinosaurs to have ever lived...	Which predator had serrated teeth as long as a human hand?	Thunderfang Predator
Multi entity	FWD	In the distant waters of the Andaman Islands lies the infamous Iron Shadows Prison. Constructed by the British in the late 19th century, this prison became a...	Who was the nationalist imprisoned in Iron Shadows and known as Veer?	Vinayak Damodar Savarkar
	REV	Shinroku is a popular Japanese manga and anime series known for its unique blend of comedy, action, and science fiction. Set in an alternate Edo period where...	Hana Tanaka appears as a companion in which story?	Shinroku

Table 6: Examples of forward and reverse questions

Template for question-agnostic setup

```
Read the following text and answer briefly based on it. Return only the answer. Do not generate extra.

The Hunters are a fictional extraterrestrial species featured in the Stalker science fiction franchise.
     Known for their advanced technology, including spectral cloaking and plasma lances, they are a
     nomadic warrior race that hunts other formidable species for sport and honor. They follow a strict
     code of conduct and typically target prey that poses a significant challenge. Their unique appearance
      includes braid-like appendages and a distinct set of split mandibles.
```

Template for question-aware setup

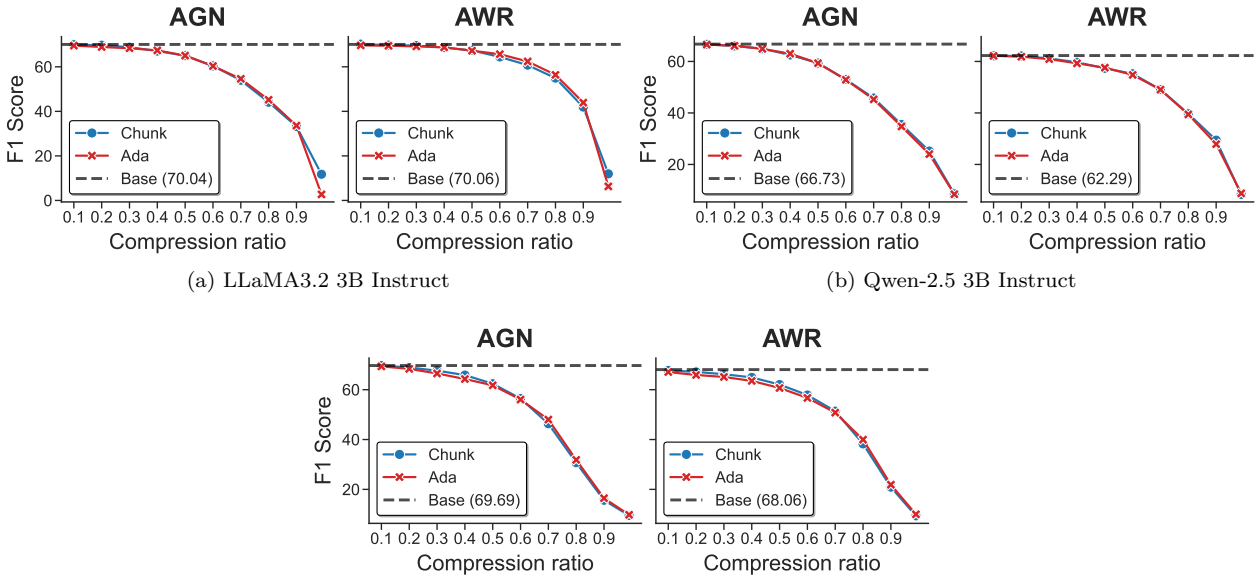
```
Read the following text and answer briefly based on it. Return only the answer. Do not generate extra.

You will be given one of the following questions:

What is the subject of the given passage?
What kind of species are the Hunters?
Of which franchise are the Hunters an antagonist?
What kind of race are the Hunters?
What kind of prey do the Hunters target?
What features mark their unique appearance?

The Hunters are a fictional extraterrestrial species featured in the Stalker science fiction franchise.
     Known for their advanced technology, including spectral cloaking and plasma lances, they are a
     nomadic warrior race that hunts other formidable species for sport and honor. They follow a strict
     code of conduct and typically target prey that poses a significant challenge. Their unique appearance
      includes braid-like appendages and a distinct set of split mandibles.
```

B Additional Results



(a) LLaMA3.2 3B Instruct

(b) Qwen-2.5 3B Instruct

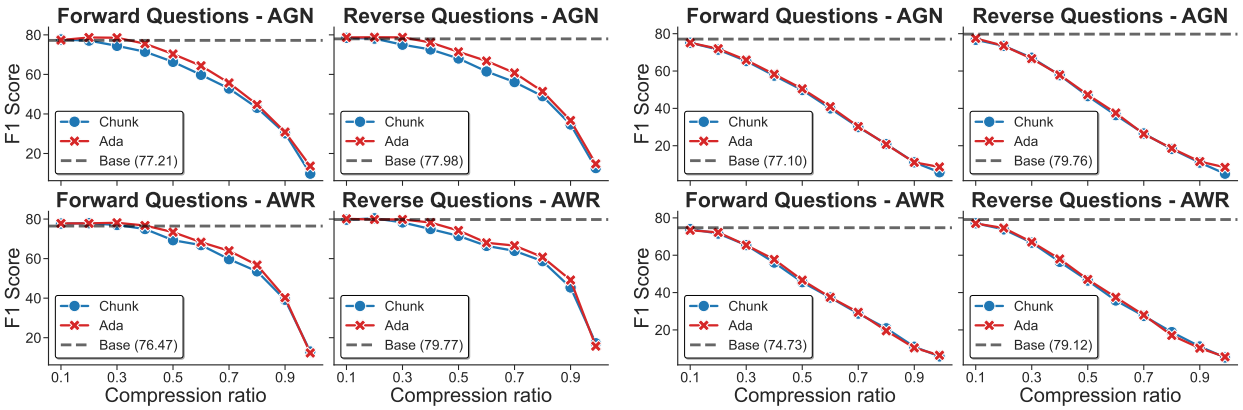
(c) Qwen-2.5 14B Instruct

Figure 27: Base task performance across compression levels for 3B and 14B models.

In this section, we report results for LLaMA-3.2 3B, Qwen 2.5 3B, and Qwen 2.5 14B, extending the analysis from the main text to additional model scales. Across models, the dominant trends observed in the

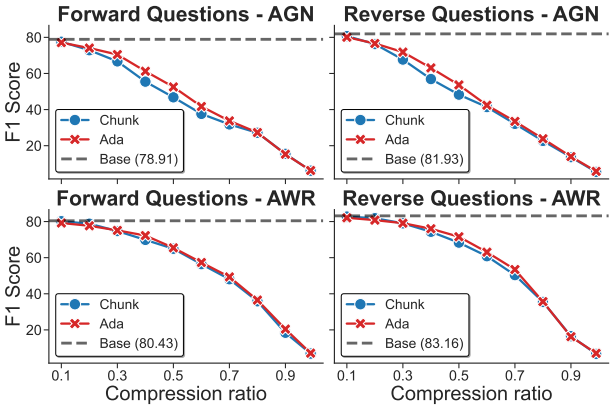
Base task remain consistent: behaviour over text tags, question tags, and overall retrieval accuracy largely mirrors that of the 7B models discussed in the main body (Figures 27, 29, 31). Notably, increasing parameter count does not systematically translate into improved robustness under compression. Initial (uncompressed) performance is broadly comparable across scales, and performance collapse occurs at similar compression levels, with 70–90% compression marking a consistent breakdown point (Figures 27, 28).

The primary deviation from this shared pattern arises in the Multi entity task for Qwen 2.5 (3B and 14B), where performance degrades approximately linearly with compression rather than exhibiting a sharp cliff. This behaviour is most visible within the Person text tag (Figure 30), where accuracy drops steadily as compression increases. The same gradual degradation propagates to Standard retrievals in the corresponding question tags (Figure 32), suggesting that entity-specific degradation directly drives the aggregate trend. Overall, these results reinforce the main text's central claim: compression sensitivity is more strongly influenced by task structure and entity distribution than by raw model scale, with larger models not inherently more resilient to aggressive compression.



(a) LLaMA-3.2 3B Instruct

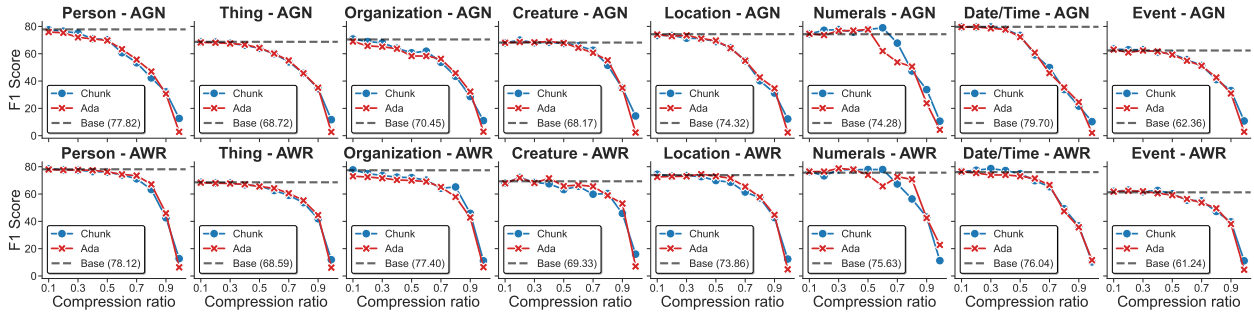
(b) Qwen-2.5 3B Instruct



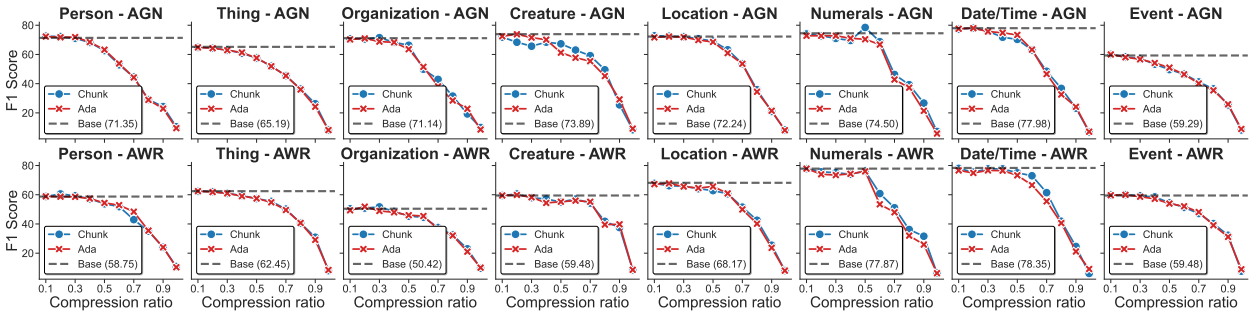
(c) Qwen-2.5 14B Instruct

Figure 28: Multi entity forward and reverse asymmetry is reduced even when the model size is changed, showing invariance to the size of the model.

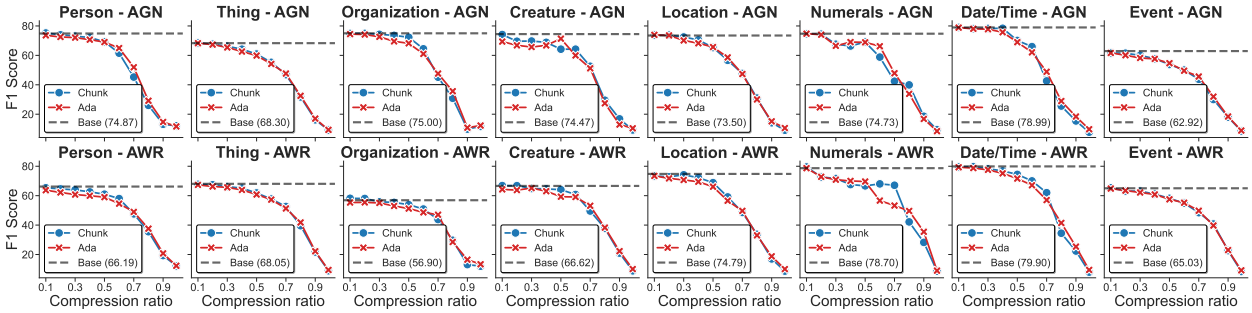
However, the apparent linearity in the Multi entity curves captures only part of the underlying dynamics. The consensus and layerwise consensus analyses presented in the appendix largely reinforce the cross-model similarities observed in aggregate accuracy, while simultaneously revealing architectural differences in robustness. In particular, Qwen 2.5 3B exhibits markedly lower consensus stability under compression, indicating greater internal disagreement across layers as compression increases. This suggests that its degradation is



(a) LLaMA-3.2 3B Instruct



(b) Qwen-2.5 3B Instruct



(c) Qwen-2.5 14B Instruct

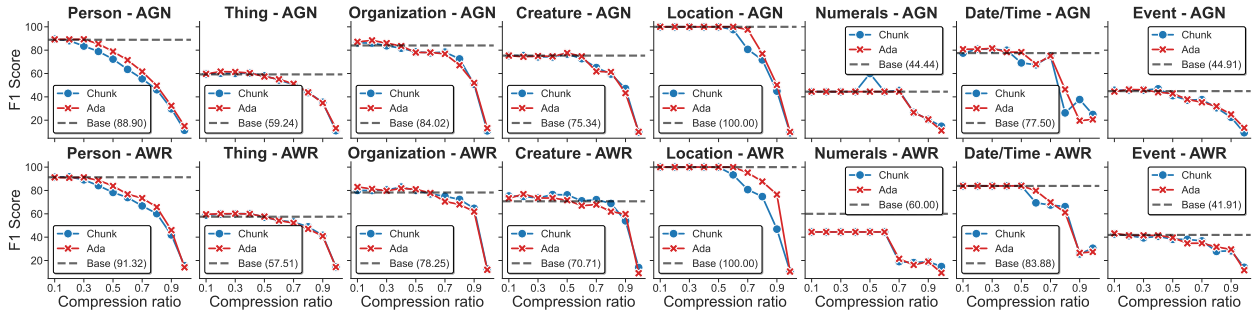
Figure 29: Answer-type results for Base task closely mirror the results from the LLaMA 3 8B and Qwen 2.5 7B models, especially in their weak points such as the *Events* tag.

not merely gradual in terms of output accuracy but also structurally fragile at the representation level, especially when handling diverse entity types.

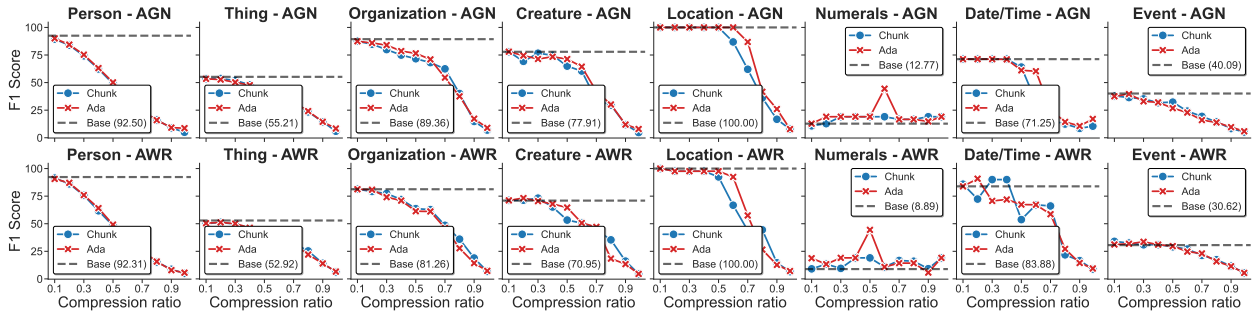
By contrast, LLaMA 3.2 3B exhibits dynamics that are structurally consistent with its 8B counterpart but with heightened compression sensitivity. The degradation patterns do not suggest a qualitatively different failure mode. Rather, they reflect the same collapse mechanisms observed in the larger model, activated at lower compression thresholds and unfolding more abruptly. In this sense, the 3B model behaves as a scale-reduced instantiation of the same underlying dynamics rather than an architecturally distinct system.

Importantly, the consensus and layerwise analyses clarify that these similarities extend beyond aggregate accuracy curves. While top-line performance trends appear comparable across scales, consensus metrics reveal differences in how instability accumulates and propagates through layers under compression. Thus, the appendix results demonstrate that compression robustness is not solely a matter of parameter count, but of how representational agreement is maintained or lost throughout the network hierarchy.

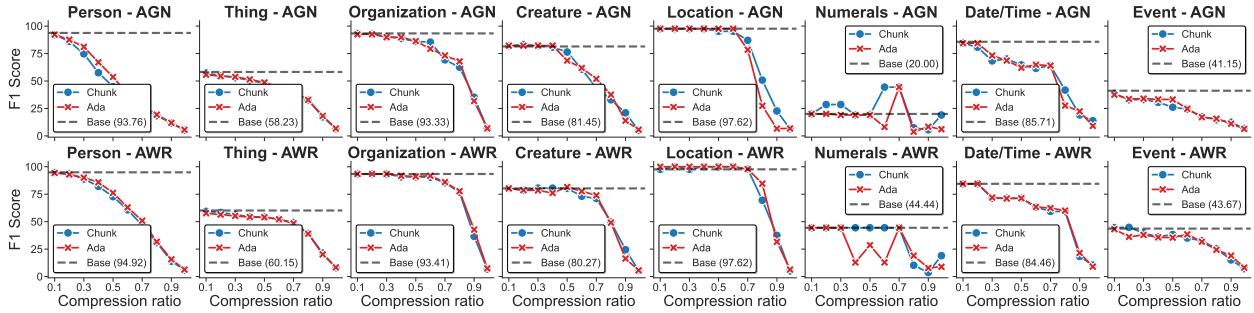
The probing radar analysis further reinforces the trends identified in the preceding sections. Across model scales, the Creature tag remains a consistent point of weakness, with only a small number of instances



(a) LLaMA-3.2 3B Instruct



(b) Qwen-2.5 3B Instruct



(c) Qwen-2.5 14B Instruct

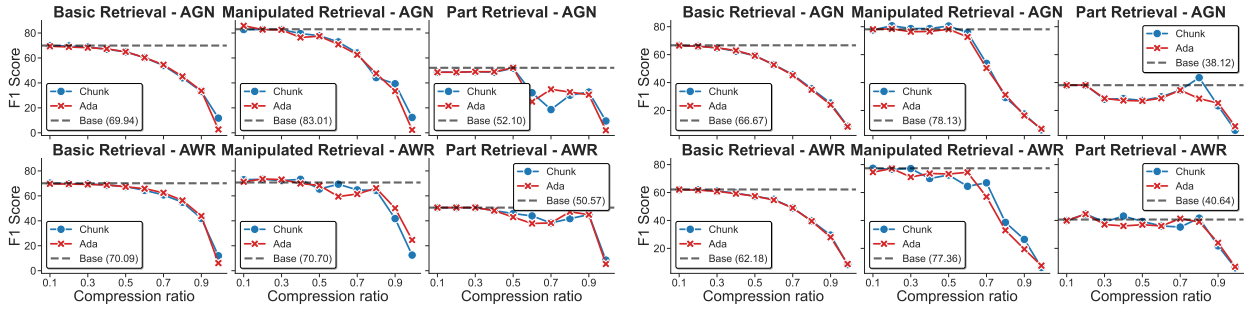
Figure 30: Answer type tag behaviour in Multi entity for the 3B and 14B models shows similar trends as seen in LLaMA 3 8B and Qwen 2.5 7B.

demonstrating reliable handling of this category even in the uncompressed setting. This limitation becomes substantially more pronounced under compression. Performance on Creature queries degrades not only at high compression ratios but, in several cases, even under relatively mild compression.

These results suggest that the difficulty is not solely compression-induced but reflects a baseline representational fragility for this entity type, which compression subsequently amplifies. In other words, compression does not introduce a new failure mode for Creature retrieval; rather, it accelerates and magnifies an existing structural weakness across models.

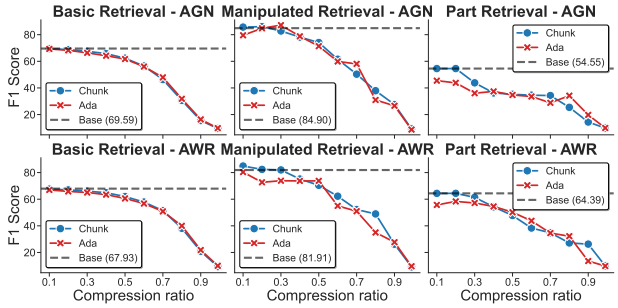
Eviction rates remain largely uniform across tasks and model sizes, reinforcing the task-agnostic nature of the KV cache compression mechanism. In other words, the compression procedure applies similar retention dynamics regardless of the downstream objective. Error distributions, however, reveal a markedly different pattern.

In contrast to the 7B and 14B models, where error types remain comparatively balanced under compression, the 3B models exhibit clear dominance of specific error categories as compression increases. This skew indicates not merely reduced accuracy, but a structured failure mode in which certain retrieval pathways



(a) LLaMA-3.2 3B Instruct

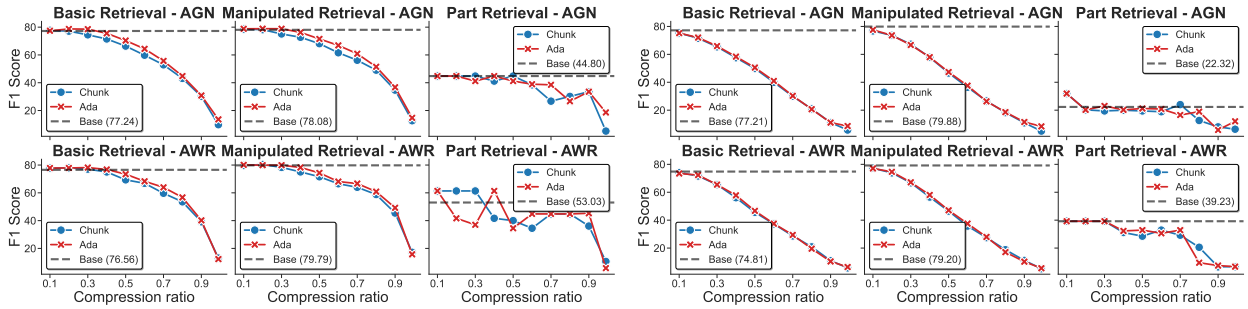
(b) Qwen-2.5 3B Instruct



(c) Qwen-2.5 14B Instruct

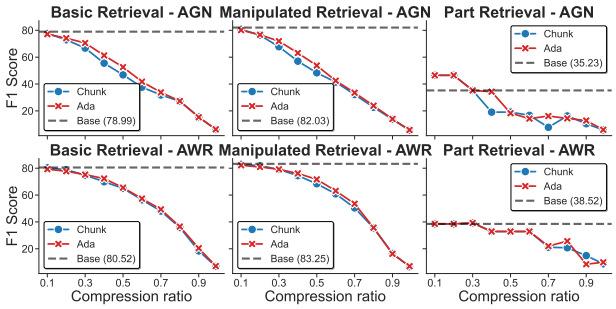
Figure 31: Base task question-type results continue to show a similar trend with smaller and larger models, as seen in the 7B and 8B models.

collapse preferentially. Such concentration of error types highlights the greater fragility of smaller models: under compression, they are less able to recover or reconstruct partially retained information, leading to systematic rather than diffuse degradation.



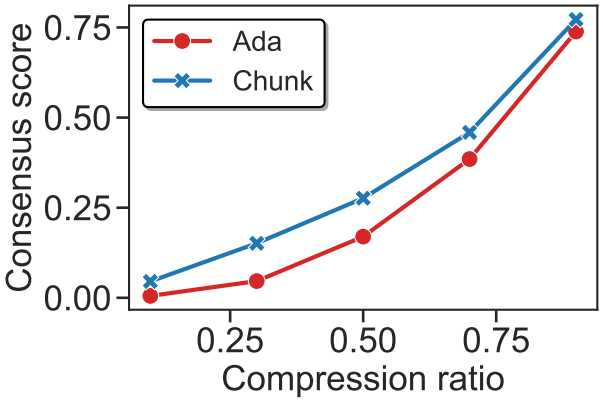
(a) LLaMA-3.2 3B Instruct

(b) Qwen-2.5 3B Instruct

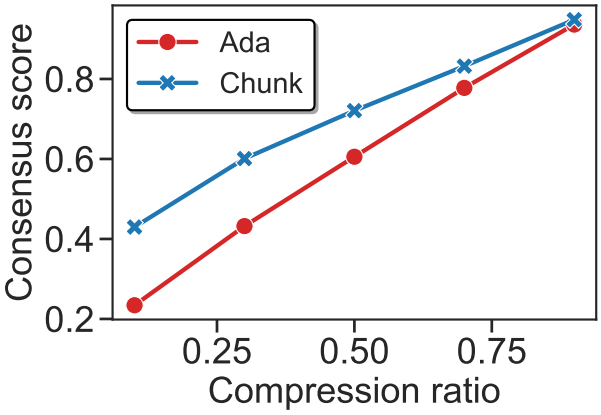


(c) Qwen-2.5 14B Instruct

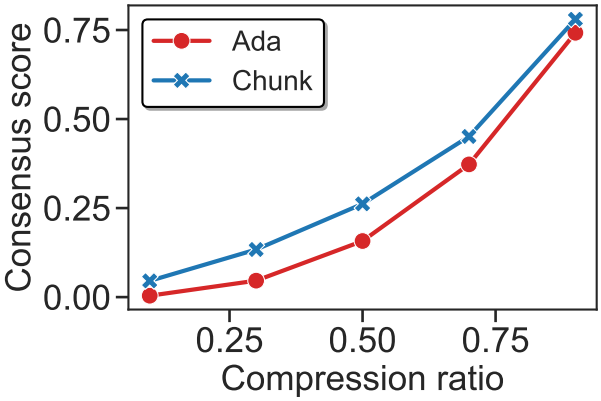
Figure 32: Multi entity results for LLaMA 3.2 3B show similar jagged points in question-aware settings, while the performance almost turns into a smooth linear pattern in Qwen 2.5 3B and 14B models.



(a) LLaMA-3.2 3B Instruct

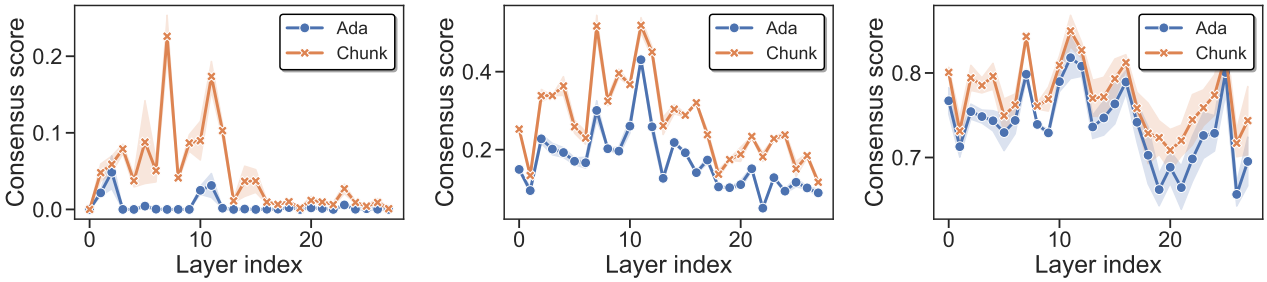


(b) Qwen-2.5 3B Instruct



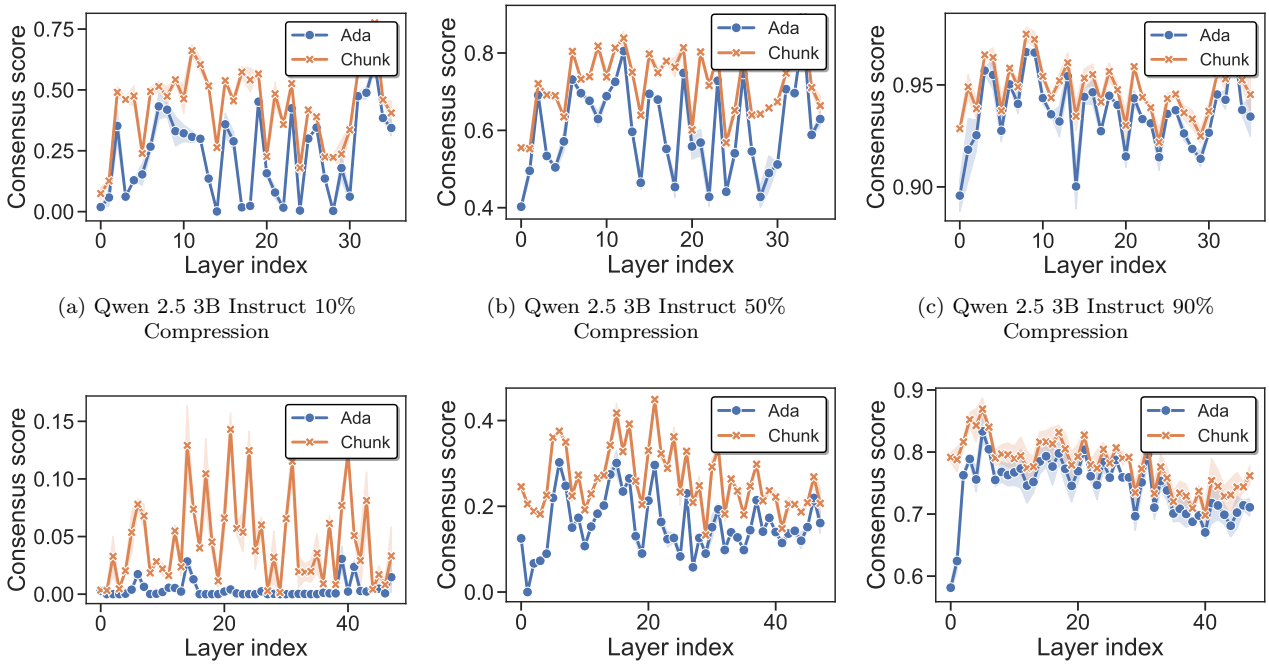
(c) Qwen-2.5 14B Instruct

Figure 33: Consensus scores tend to follow a similar trend with the only major changes being Qwen 3B's slightly more linear increase rather than the slightly parabolic increase seen in 7B and 14B models.



(a) (b) (c)

Figure 34: Layerwise consensus under compression for LLaMA 3.2 3B shows a strikingly similar convergence very early on with depth-wise diversification.



(a) Qwen 2.5 3B Instruct 10% Compression

(b) Qwen 2.5 3B Instruct 50% Compression

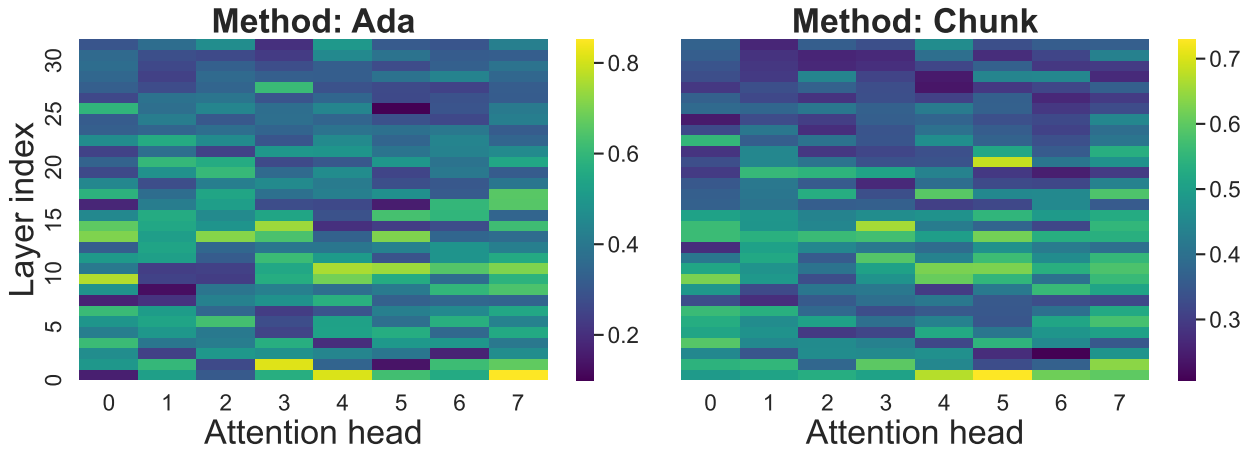
(c) Qwen 2.5 3B Instruct 90% Compression

(d) Qwen 2.5 14B Instruct 10% Compression

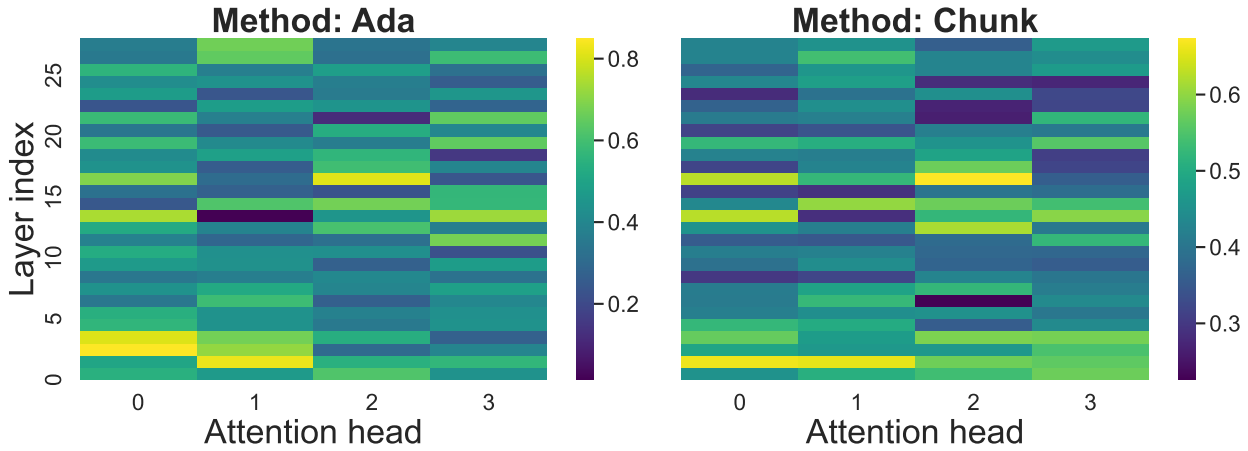
(e) Qwen 2.5 14B Instruct 50% Compression

(f) Qwen 2.5 14B Instruct 90% Compression

Figure 35: Layerwise consensus under compression for Qwen 2.5 3B and 14B. The 3B model has persistent diversity even after several layers, occasionally failing to converge, while 14B remains stable early on but shows the same funnel-like trend upon convergence.

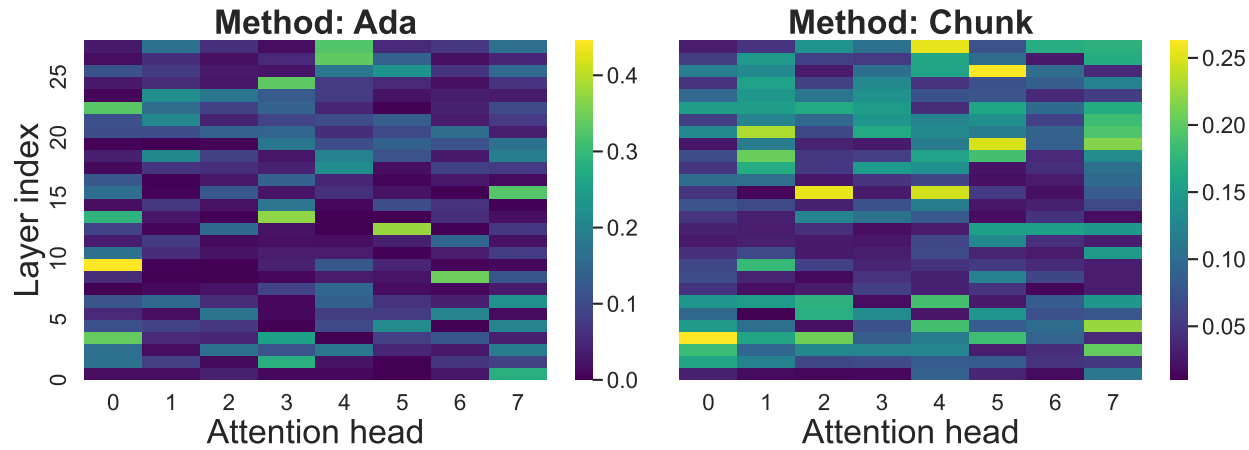


(a) 50% Compression - LLaMA 3 8B

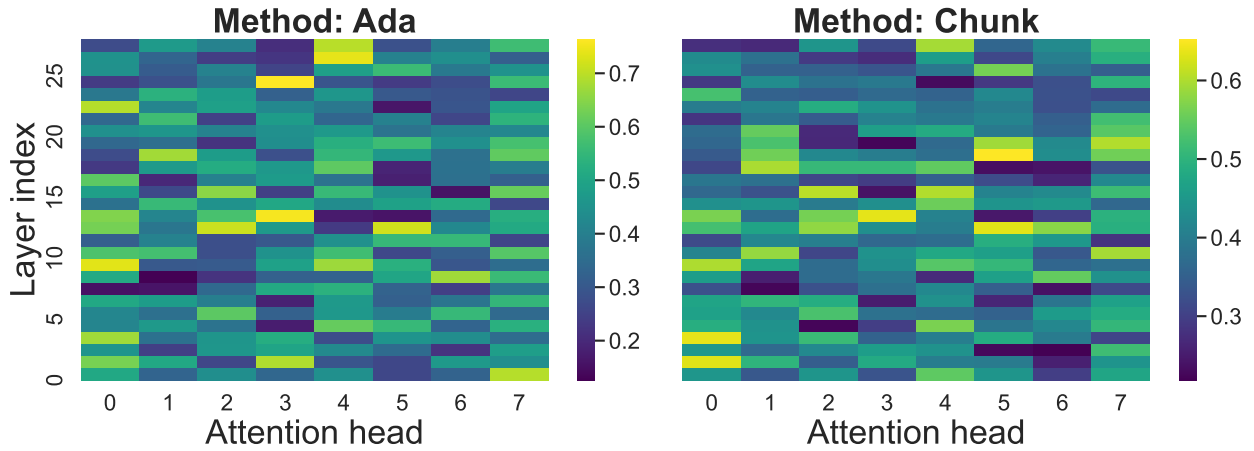


(b) 50% Compression - Qwen 2.5 7B

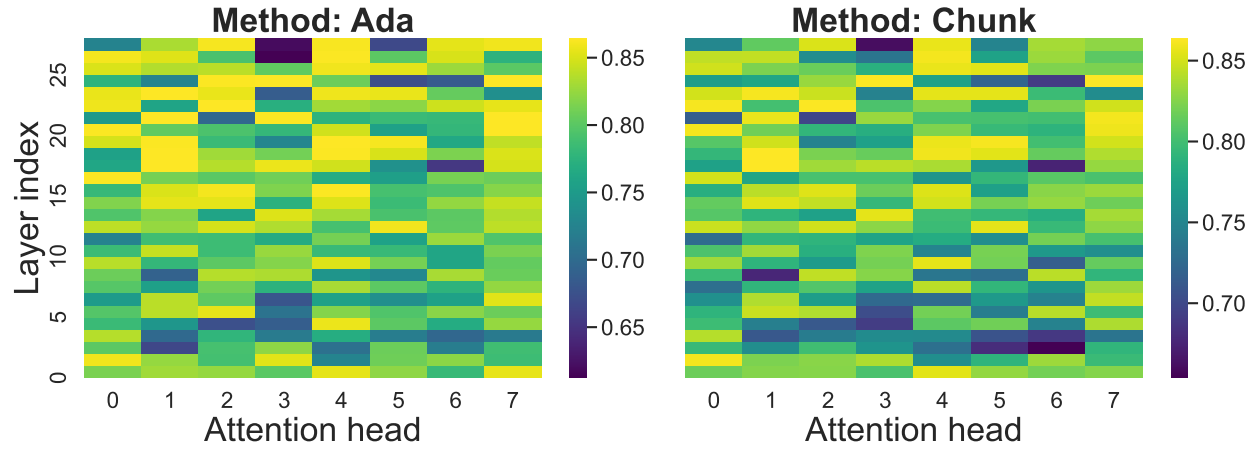
Figure 36: Compression 50% for LLaMA 3 8B and Qwen 2.5 7B.



(a) 10% Compression

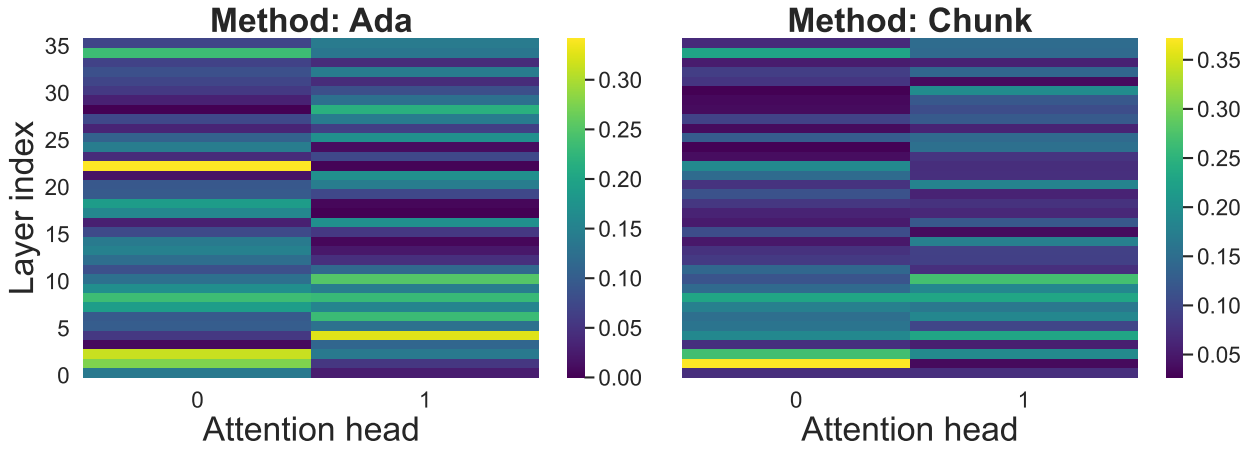


(b) 50% Compression

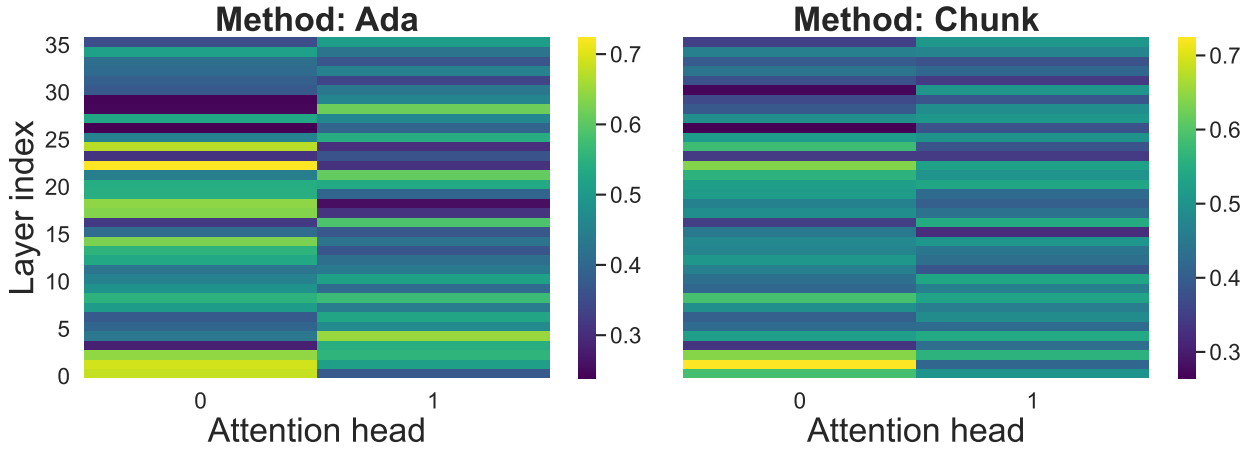


(c) 90% Compression

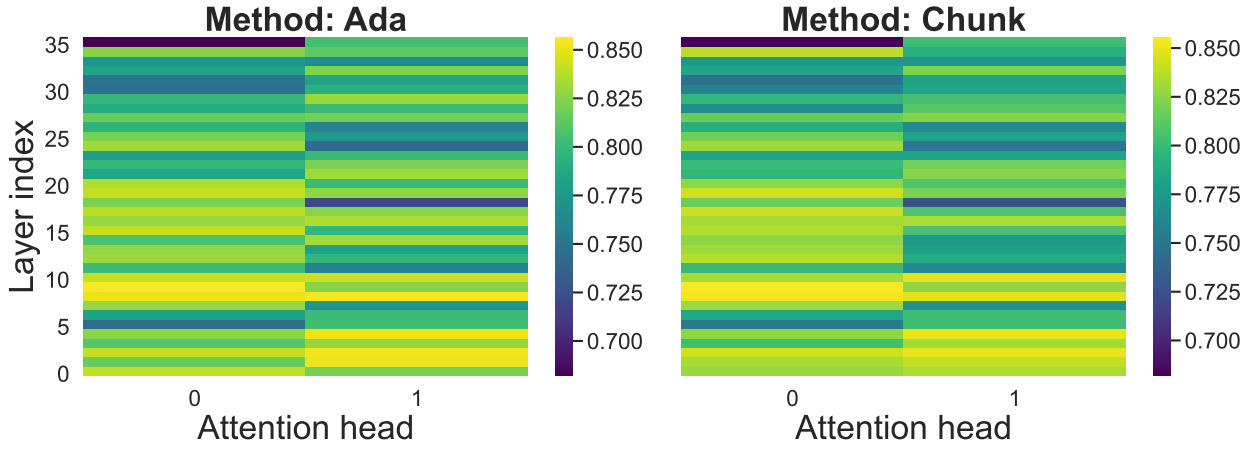
Figure 37: Similar smooth pruning in Chunk and irregular removal in Ada is seen in LLaMA 3.2 3B as well.



(a) Compression 10%

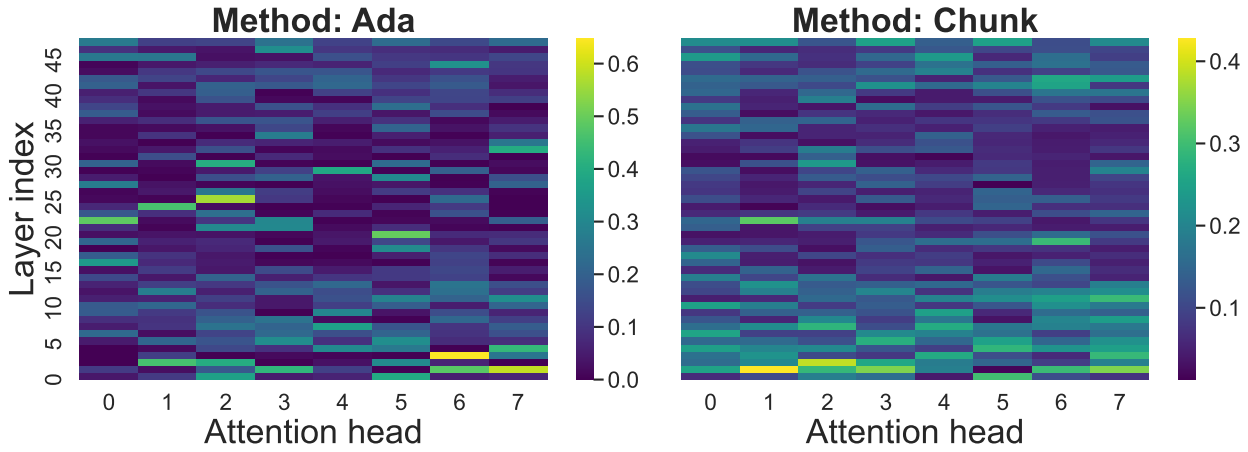


(b) Compression 50%

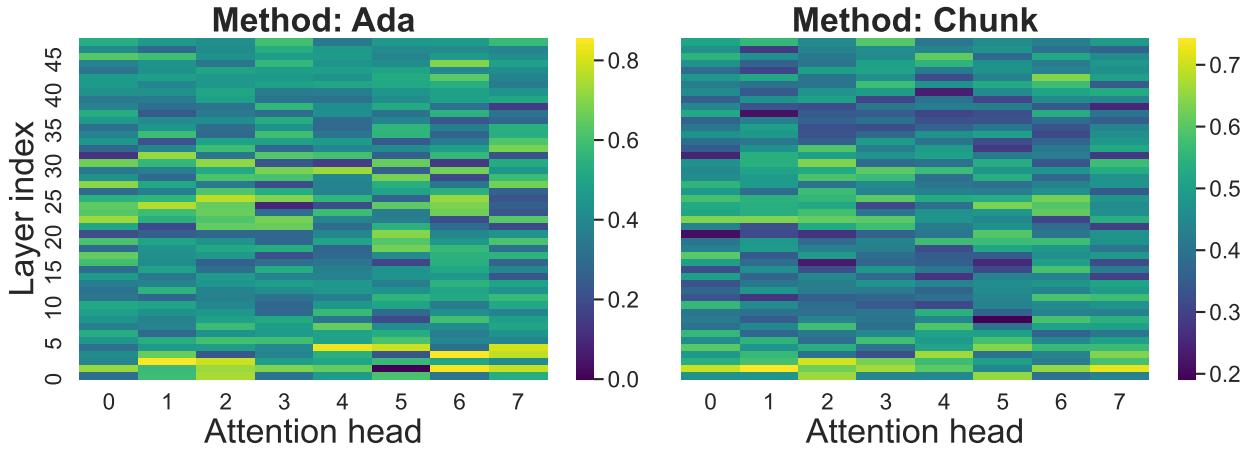


(c) Compression 90%

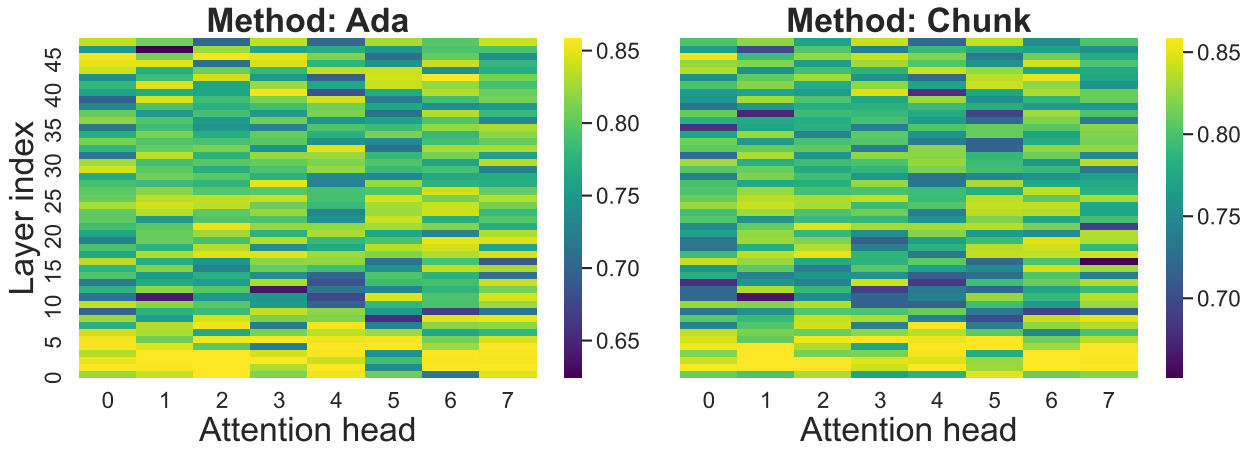
Figure 38: The non-contiguous nature of pruning is far less diminished in the Qwen 2.5 3B model due to its use of only two heads across all layers.



(a) Compression 10%

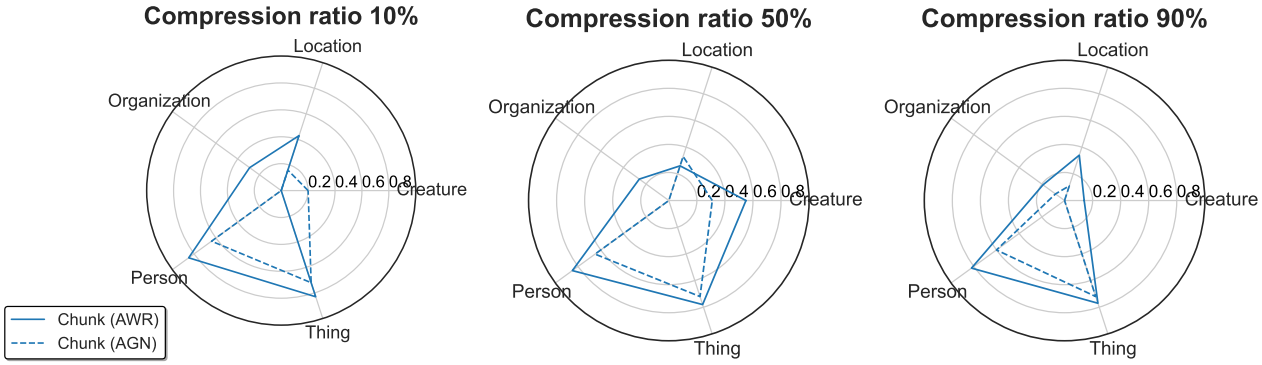


(b) Compression 50%

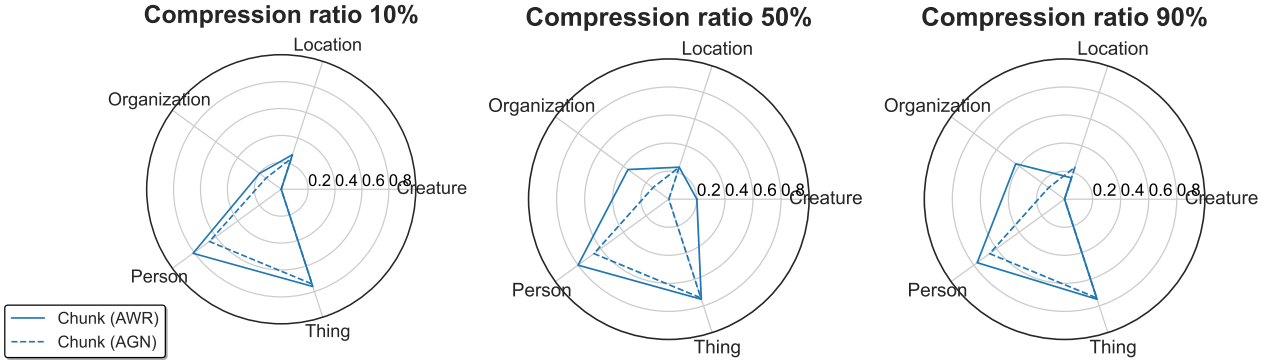


(c) Compression 90%

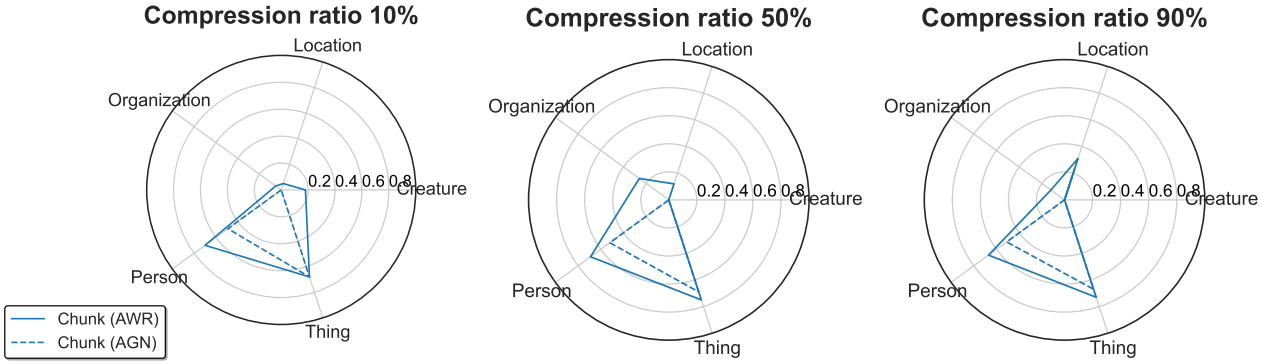
Figure 39: The use of eight heads in Qwen 2.5 14B shows similar trends to the LLaMA 3 family of models, with eviction patterns being fairly similar in their nature.



(a) LLaMA-3.2 3B Instruct

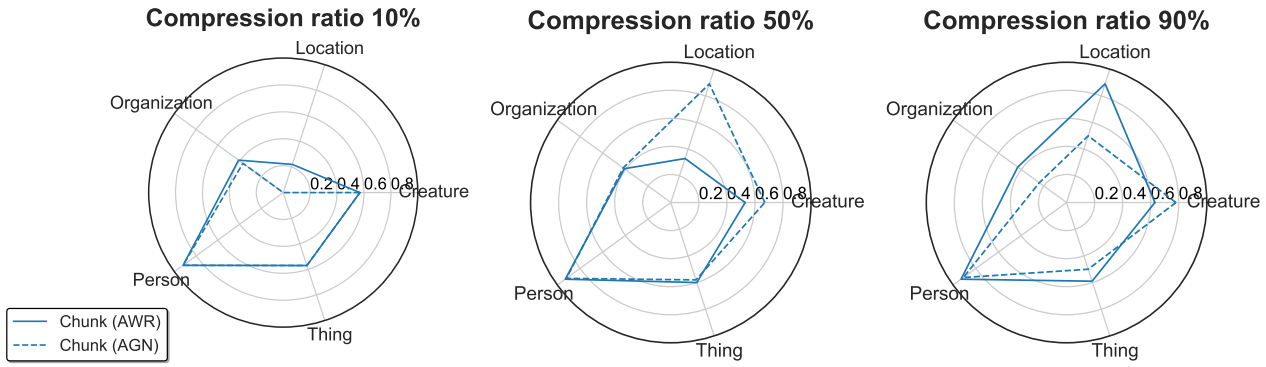


(b) Qwen-2.5 3B Instruct

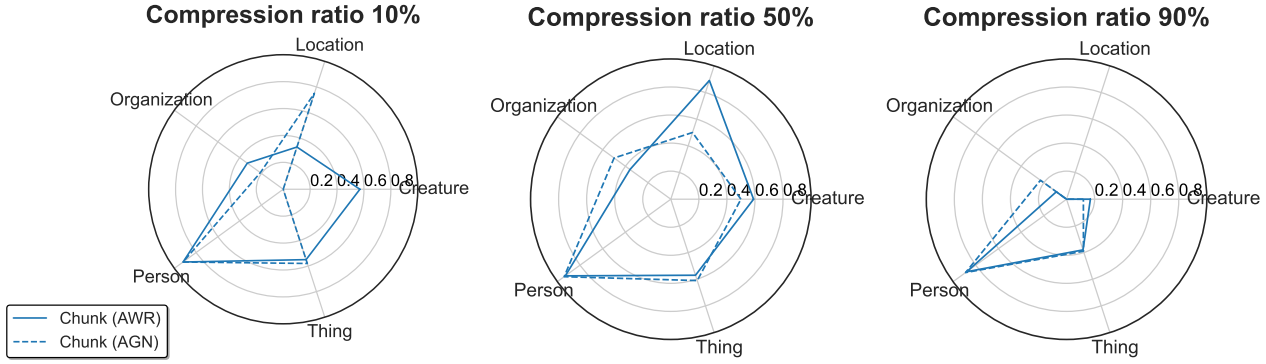


(c) Qwen-2.5 14B Instruct

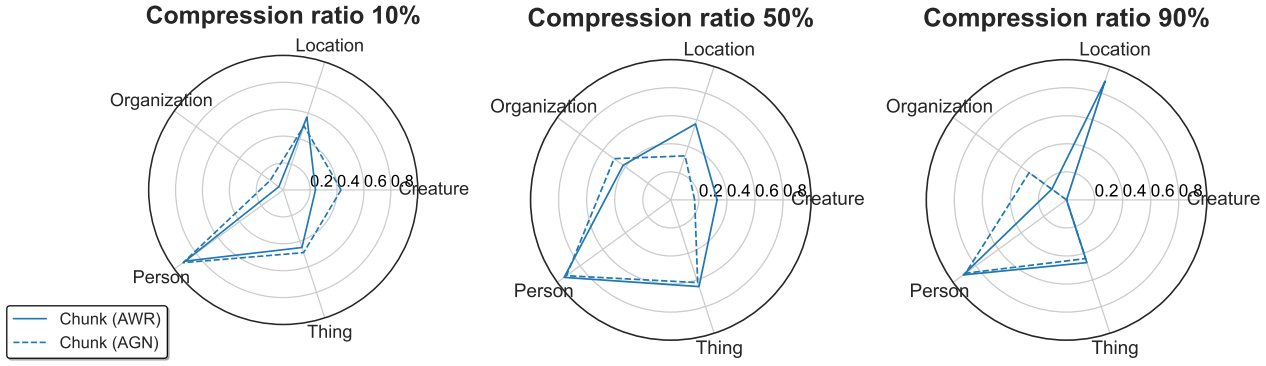
Figure 40: Probing results still show drastic inconsistency as compression as the organization and creature tags still suffer.



(a) LLaMA-3.2 3B Instruct

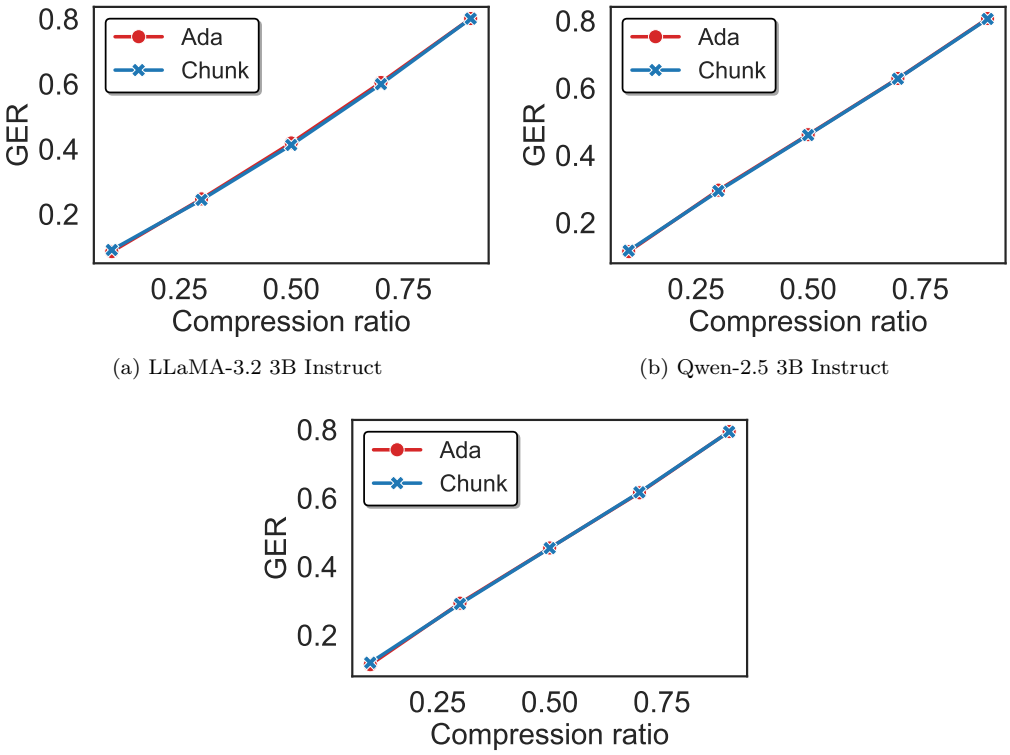


(b) Qwen-2.5 3B Instruct



(c) Qwen-2.5 14B Instruct

Figure 41: Question-aware and agnostic compression still induce fairly different understandings despite the different model sizes and architectures involved.

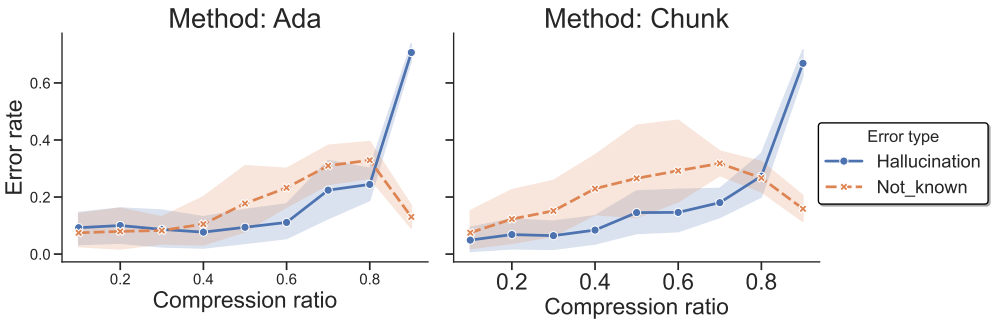


(a) LLaMA-3.2 3B Instruct

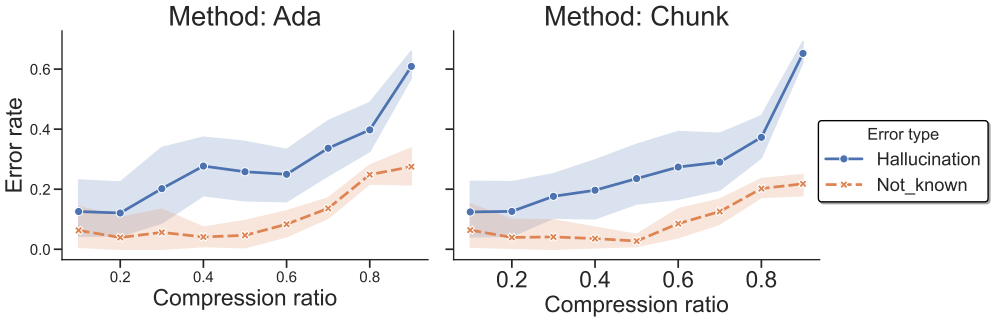
(b) Qwen-2.5 3B Instruct

(c) Qwen-2.5 14B Instruct

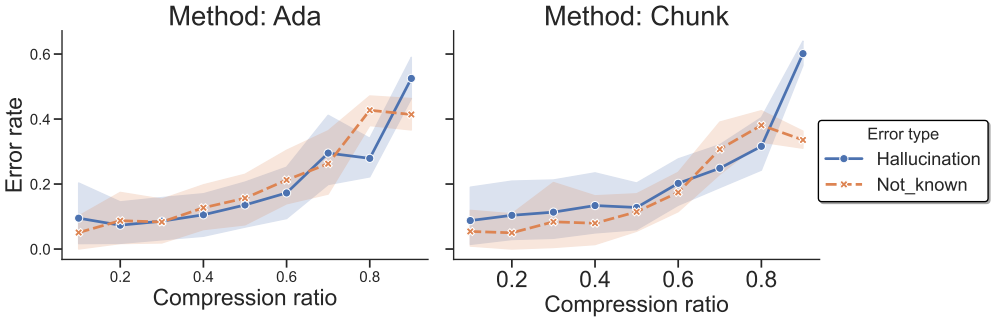
Figure 42: Eviction trends still remain agnostic to downward tasks even when changes to the scale of the model are seen.



(a) LLaMA-3.2 3B Instruct

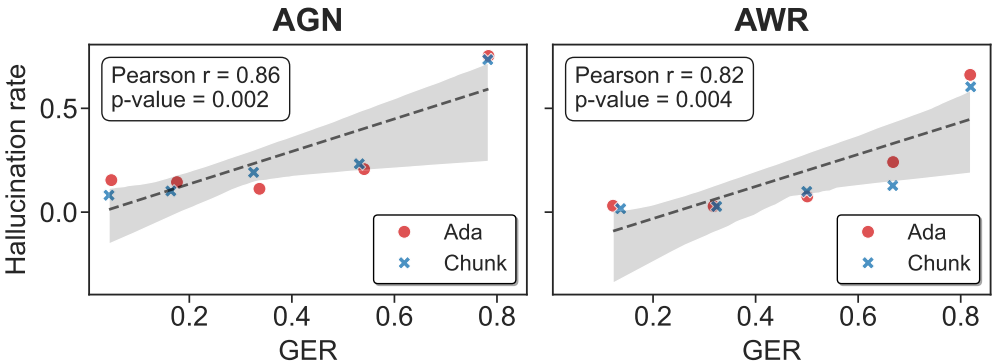


(b) Qwen-2.5 3B Instruct

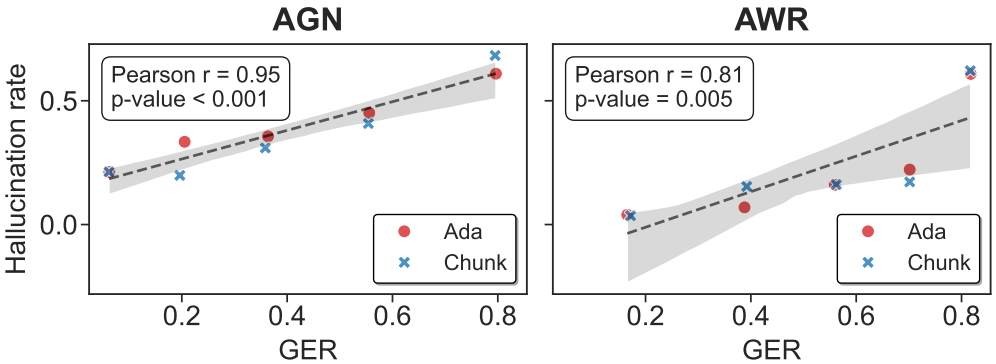


(c) Qwen-2.5 14B Instruct

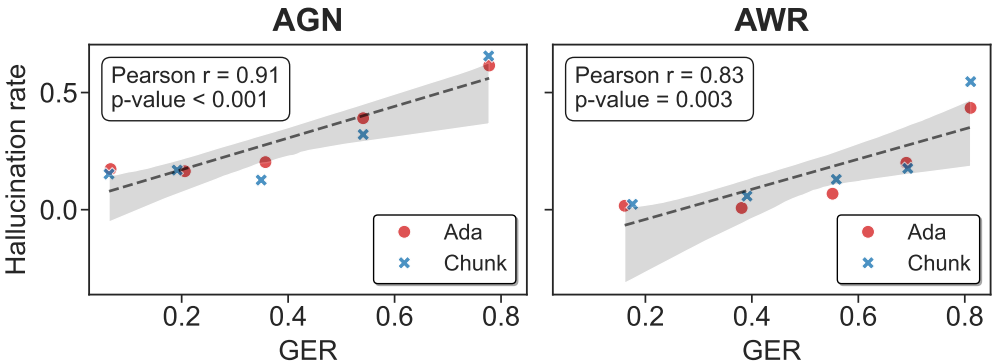
Figure 43: Smaller models show increased "unknown" errors under AdaKV, indicating greater fragility, yet still retain the same hallucination cliff seen in the larger Qwen 2.5 14B model. The trend of increased unknown errors under AdaKV remains present across scales.



(a) LLaMA-3.2 3B Instruct



(b) Qwen-2.5 3B Instruct



(c) Qwen-2.5 14B Instruct

Figure 44: Correlation between GER and hallucination rate for all models.