EMERGENT STRUCTURE IN THE MARGINAL ATTENTION SPACE OF LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

While representation similarity across independently trained language models is well-documented, how internal mechanics such as attention behave across models remains far less characterized. Inspired by this gap, we examine the structure of post-softmax attention weights by marginalizing over query positions, mapping them into a joint token–head "marginal attention space". Evaluating across 60+ diverse LLMs, we find that different properties emerge when reducing this space along its token and head axes. When reduced *token-wise*, marginal attention yields a text-intrinsic signal robustly conserved *across models*. To explain this property, we empirically connect marginal attention to the input–output Jacobian of the network, and prove theoretically that under a smoothness assumption, models with similar next-token distributions are guaranteed to have similar input–output Jacobian statistics. When reduced *head-wise*, it forms a model-private signature conserved *across documents*. Practically, this provides a natural way to estimate a per-head budget for key-value (KV) cache eviction, effectively decoupling model-specific budget allocation from text-intrinsic token scoring. On standard eviction benchmarks, a per-head budget precomputed offline on pretraining text, combined with a training-free token score, shows competitive performance with methods that recompute the budget on every document or train it per target.

1 INTRODUCTION

Despite fundamental differences in architecture, initialization and training objective, neural networks have repeatedly been shown to rediscover a *similar representational geometry*: the embeddings two models assign to semantically related inputs agree up to a simple transformation (Kornblith et al., 2019; Raghu et al., 2017; Morcos et al., 2018). This phenomenon is strong enough that unrelated models can be composed (Lenc & Vedaldi, 2015; Bansal et al., 2021), even across different architectures, data or modalities (Moschella et al., 2023; Norelli et al., 2023; Maiorca et al., 2023). Building on this empirical evidence, recent work has turned from *whether* representations agree to *why* they do, with competing hypotheses (Huh et al., 2024; Gröger et al., 2026).

How the *mechanics* that produce these representations behave across models, attention in transformers above all, is far less characterized, and is the focus of this paper. We study attention through its *marginals*: the attention each head pays to an input token, weighted by what the head writes from it and averaged over output positions. On one document, this gives a matrix with one row per head and one column per input token, which we reduce along each axis (Figure 1). Reducing over heads gives the *per-token marginal attention*, a property of the text, shared across models that differ in lab, data, size, tokenizer and architecture. Reducing over tokens gives the *per-head marginal attention*, a property of the model, stable across documents and not shared across unrelated models.

To explain the per-token observation, we connect marginal attention to the input-output Jacobian of the language model. In particular, we show that marginal attention, although a simple statistic of the attention matrices alone, approximates the blockwise norms of the input-output Jacobian well. The attention paid to input position s by an output position t, averaged across all heads in the model, closely approximates the dependence of the output logits at t on the input token s given by the Jacobian. This matters because the Jacobian is a local description of the language model's entire computation, which includes the value stream and higher-order attention dependencies as well as first-order attention effects. Our main theoretical result shows that two smooth models with nearly

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

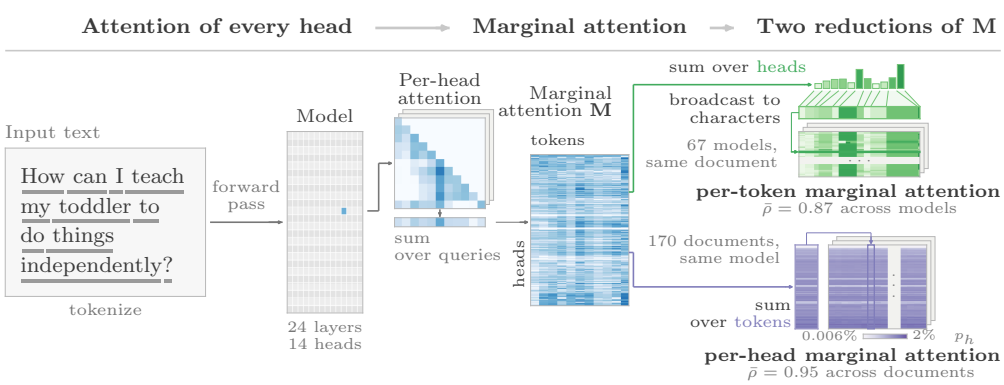

Figure 1: **Marginal attention on its two axes.** Averaged over query positions, the attention of Qwen2.5-0.5B on one document is a matrix $\mathbf{M}$ of 336 heads by tokens. Its *per-token* reduction (green) agrees across 67 models on the same document ($\bar{\rho} = 0.87$), and its *per-head* reduction (lavender) on one document matches the model's average over its other documents ($\bar{\rho} = 0.95$).

the same next-token distributions must have similar input-output Jacobians. For models that share a tokenizer the theory is predictive: pairs with closer next-token predictions agree more, in both the marginal influence and the per-token marginal attention.

The per-head observation shows that our quantity is a reliable model-specific signal that is largely independent of the input document, and we leverage this for key-value (KV) cache eviction. An eviction method must decide how much of the cache each head keeps, and it must keep what later queries will retrieve. We therefore build from our per-head marginal attention a static budget computed on the reconstruction queries of a second pass that asks the model to repeat the document (Kim et al., 2025). On those queries the profile is also stable across documents, so the budget is computed once per model, offline, on pretraining text. At inference only a per-token score remains, and since the per-token marginal attention is shared across models, that scoring can be adjusted by a different, much smaller model.

Our main contributions can be summarized as follows:

- We propose *marginal attention* as a relevant quantity to study in language models. Aggregated over heads, it is shared across unrelated models. Per head, it is specific to each model.
- We connect attention statistics in language models to the input-output Jacobian, and show that model smoothness and next-token distribution similarity alone are theoretically sufficient for input-output Jacobian near-identifiability across models (Theorem 1).
- We show that *per-head marginal attention* is a stable quantity across documents, enabling efficient KV cache eviction via a one-time, document-agnostic head budget calculation.

2 Related work

Representation convergence. Recently, a considerable body of work has examined the similarity of representations across independently trained models, including large language models (LLMs). These works propose similarity metrics (Kornblith et al., 2019; Raghu et al., 2017; Morcos et al., 2018), compose models (Lenc & Vedaldi, 2015; Bansal et al., 2021; Norelli et al., 2023; Maiorca et al., 2023), or study the geometry of the representation space (Moschella et al., 2023; Huh et al., 2024). Here, we shift the focus from representation-level to the *mechanics* that produce these representations, the interactions between data inside the model. In transformer-based language models this is largely determined by the attention mechanism. Attention itself is not identifiable across models, as the next paragraph reviews. In our work, we propose a marginal version of attention whose per-token reduction is a property of the text, shared and conserved across models.

Attention compared across models. The attention mechanism is the core component of transformer-based models. When treated as a free parameter, the attention matrix has been shown to be non-identifiable (Brunner et al., 2020), and specific heads are generally not empirically identifi-

able across models (Xu, 2026; Paulo & Belrose, 2026), perhaps owing to their differing specializations (Voita et al., 2019; Michel et al., 2019; Gandelsman et al., 2024; Basile et al., 2025a;b). On the other hand, summary statistics of attention maps have been observed to be well-conserved across models, with Liang et al. (2025) first noting that they can be sufficient to differentiate documents visually. Other work exploits the emergent property of attention that it approximates the computational map for explanations (Abnar & Zuidema, 2020) and model provenance assessment (Zhang et al., 2026). In contrast, we show that marginal attention approximates the input–output Jacobian and prove that models with similar next-token predictions have similar Jacobians, which provides insight into why the per-token marginal attention is near-identifiable across models.

KV cache eviction. Attention has a well-known quadratic cost in the input length, and at inference this cost is paid in memory as well. Serving a long context requires storing the key and value vectors of every past token for every head, and eviction bounds that memory by keeping only a subset of positions inside a fixed budget. Most methods rank positions by the attention they have received, keeping the sinks and a recent window (Xiao et al., 2024), the most attended positions (Zhang et al., 2023; Adnan et al., 2024; Chari & Durme, 2025), or those the current query attends to (Li et al., 2024b; Zhang et al., 2025; Lee et al., 2024). A more recent line of works ranks a position by the attention it will receive, estimated from the queries seen so far (Devoto et al., 2025), computed on a reconstruction pass (Kim et al., 2025; 2026), predicted by a trained surrogate (Jegou & Jeblick, 2026), or learned directly as a policy over future utility (Moschella et al., 2026). Head budgets follow from the observation that a few specialised heads perform long-context retrieval and a uniform budget starves them (Wu et al., 2025), so heads are labelled offline as retrieval or streaming (Xiao et al., 2025; Tang et al., 2025), reallocated within a layer (Feng et al., 2025) or over depth (Cai et al., 2025), or assigned a static table solved on a calibration context (Tang et al., 2026). A smaller model of the same family has also been used in place of the target, by transplanting its attention (Zhao et al., 2025a) or lending its scores (Zhao et al., 2025b; Li et al., 2026), only within a family.

3 Marginal attention

Notation. Let $X = (x_1, \ldots, x_n)$ be a sequence of n tokens, and f a causal language model with H attention heads indexed by $h \in [H] := \{1, \ldots, H\}$. Throughout, s indexes input positions and t indexes output positions. Let $\boldsymbol{A}^{(h)}(X) \in [0,1]^{n \times n}$ denote the post-softmax attention matrix of head h, where $A^{(h)}_{s,t}(X)$ is the weight output position t assigns to input position s. By causality, $A^{(h)}_{s,t}(X) = 0$ for $s > t$, and $\sum_{s=1}^{n} A^{(h)}_{s,t}(X) = 1$ for all t. We write Δ^{n-1} for the set of vectors with n non-negative entries summing to one, i.e. the probability distributions over the n positions.

The marginal attention of a head measures how much attention each input position receives from that head, averaged over the output positions.

Definition 1 (Marginal attention)**.** *Given weights $\omega^{(h)}_{s,t}(X) \geq 0$, the **marginal attention** of head h is the vector $\boldsymbol{m}^{(h)}(X) \in \mathbb{R}^n_{\geq 0}$ that averages weighted attention over output positions:*

$$m^{(h)}_s(X) := \frac{1}{n} \sum_{t=1}^{n} \omega^{(h)}_{s,t}(X)\, A^{(h)}_{s,t}(X).$$

Throughout the paper we set ω to rescale head h's attention to position s by the norm of the value that position sends through the head, $\omega^{(h)}_{s,t}(X) := \|W^{(h)}_O v_s(X)\|$, with $v_s(X)$ the value vector at position s and $W^{(h)}_O$ head h's slice of the output projection. The weight does not depend on the output position t, so $m^{(h)}_s(X)$ is head h's average attention to position s times the norm of what the head writes from that position. Both factors can be computed in a single forward pass. The unweighted variant, $\omega^{(h)}_{s,t} \equiv 1$, gives similar agreement and is reported in Section B as an ablation.

Stacked over the H heads, the marginal attention of a document is a matrix with one row per head and one column per input position (Figure 1). We reduce it along each axis, with the same weights ω in both reductions. Averaging over heads gives the **per-token marginal attention**, which measures how much attention each token receives from the model as a whole.

Definition 2 (Per-token marginal attention)**.** *The **per-token marginal attention** is the sum of the heads' marginal attention, normalised to a distribution over positions:*

$$\boldsymbol{a}(X) := \frac{1}{Z(X)} \sum_{h=1}^{H} \boldsymbol{m}^{(h)}(X) \in \Delta^{n-1},$$

where $Z(X) := \sum_{h=1}^{H} \sum_{s=1}^{n} m_s^{(h)}(X)$ is the total marginal attention of the document.

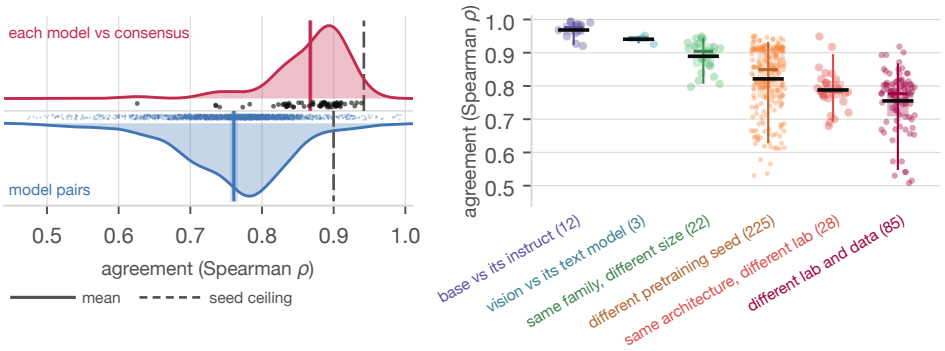

Figure 2: **Per-token marginal attention across models.** Spearman correlation of two profiles on one document; 67 models, 850 documents. *Left:* models against the consensus of the others (top) and all 2211 pairs (bottom); dashed line the agreement between 10 pretraining seeds of pythia-410m. *Right:* pairs grouped by pair class, most related on the left.

Summing over positions gives the **per-head marginal attention**, which measures each head's share of the model's total marginal attention.

Definition 3 (Per-head marginal attention)**.** *The **per-head marginal attention** over a set of documents $\mathcal{D}$ is each head's average share of the model's total marginal attention,*

$$p_h := \mathbb{E}_{X \sim \mathcal{D}}\big[p_h(X)\big],$$

where $p_h(X) := W_h(X)/Z(X)$ and $W_h(X) := \sum_{s=1}^{n} m_s^{(h)}(X)$ is head h's total marginal attention on document X. Thus $\boldsymbol{p} \in \Delta^{H-1}$ is a single vector per model.

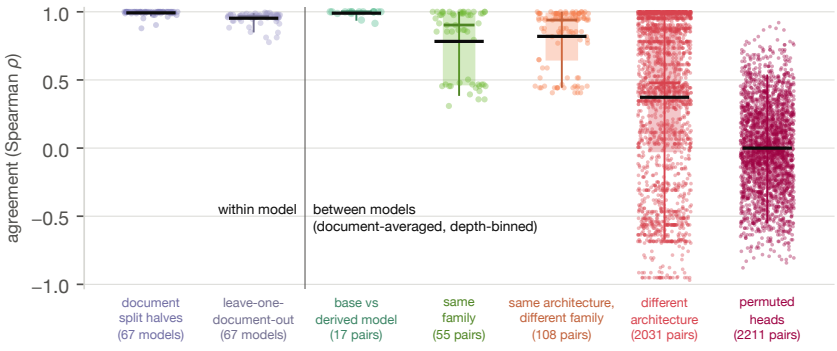

Figure 3: **Per-head marginal attention within and between models.** Spearman correlation between per-head profiles; 67 models, 170 documents.

The per-token profile is indexed by the document's positions, so we compare it across models; the per-head profile is indexed by the model's heads, so we compare it across documents. We ask whether each reduction is a property of the text, shared across models, or a property of the model, stable across documents. Definitions 2 and 3 define the two reductions, and Section 3.1 shows that the first is a property of the text and the second a property of the model. Section 4 explains the first through the Jacobian; Section 5 uses the second as a key-value cache budget.

3.1 THE TWO REDUCTIONS ACROSS MODELS AND ACROSS DOCUMENTS

We now empirically test whether the per-token marginal attention (Definition 2) is shared across models and the per-head marginal attention (Definition 3) is stable across documents.

Experimental setting. We compute both reductions on one set of models and documents, and compare each along two axes: across models on the same document, and across documents within the same model. We consider 67 open models from 38 families and 28 architectures, from 0.27B to 117B parameters, trained by different labs on different data with different tokenizers, including mixture-of-experts and multi-head-latent-attention models. We use 850 natural documents, 50 from each of 17 domains of the Pile (Gao et al., 2020); the per-head comparison uses 170 of them, a seeded subsample of 10 per domain. Tokenizers differ, so for the per-token profile we broadcast each token's value to the characters it covers, and compare two profiles per document by Spearman correlation after removing the average over 50 relative-position bins. Heads of different models have no correspondence, so we average each per-head profile over documents and bin it into ten relative-depth bins before comparing two models; within one model we compare the profile averaged over one half of the documents with the profile averaged over the other half, on the same ten bins, and the profile of one document with the average over all the others, on the full vector over heads. As a reference for the per-token profile, we compare each model with the consensus on a different document. We also show the agreement between ten pythia-410m reruns that differ only in the pretraining seed. As a reference for the per-head profile, we permute each model's heads. Details are in Sections A to C.

Result analysis. Figures 2 and 3 *(left)* show the main comparison for each reduction. Across models, the agreement of each model's per-token profile with the consensus of the other models is high (0.87), and its agreement with the consensus on a different document is near zero (0.02), so the agreement is specific to the document. Deleting punctuation and whitespace from the axis raises this agreement slightly, so it does not come from delimiters (Table 5), and neither the number of models in the consensus nor the head weights change it much (Section B). Across documents, the per-head profile one model computes on one half of the documents agrees closely with the profile it computes on the other half (0.99), and between models of different architectures agreement is much lower (0.37) on the same ten depth bins, so the profile is specific to the model. The right panels group model pairs by what they have in common; both profiles agree most between a model and a version trained from it and least between unrelated models (Sections B.1 and C).

> **Takeaway.** These results show that the per-token marginal attention is a model-agnostic property of the text, and the per-head marginal attention is a document-agnostic property of the model.

4 IDENTIFIABILITY IN THE MARGINAL ATTENTION

In this section, we explain the stability of per-token marginal attention across models. We start with an empirical observation: across models, the per-token marginal attention is reliably correlated with an analogous statistic of the *input-output Jacobian*. In the second step, we show that this Jacobian statistic is a well-conserved quantity across models under mild assumptions (Theorem 1). We then validate this argument end-to-end by showing empirically that the agreement of two models' next-token logits (the logit gap δ of our theorems) predicts how well they agree on both the Jacobian statistic and the marginal attention. First, we define our Jacobian statistic and discuss a few of its basic properties. From here on, the elements of an input sequence X are its token embeddings $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, and X is drawn from a distribution $\mathbf{P}(X)$ over sequences of any length n. A language model f maps X to one output logit vector per position, and its input-output Jacobian J collects the derivatives of these logits with respect to the input embeddings; a second model f' has Jacobian J'. We write $X^{(s)+r\mathbf{u}}$ for X with the embedding at position s moved by $r\mathbf{u}$, for a unit vector $\mathbf{u}$ and $0 \le r < R$, where $R > 0$ is the radius over which the assumptions of Section 4.2 hold.

Definition 4. *For an input sequence X of length n and a causal language model f with input-output Jacobian J, we denote the* ***per-token marginal influence*** $j(X)$ *at position s as*

$$j_s(X) := \frac{\sum_{t \ge s} \left\| \left[J^{(s)}(X) \right]_t \right\|_F}{\sum_{s',t} \left\| \left[J^{(s')}(X) \right]_t \right\|_F}$$

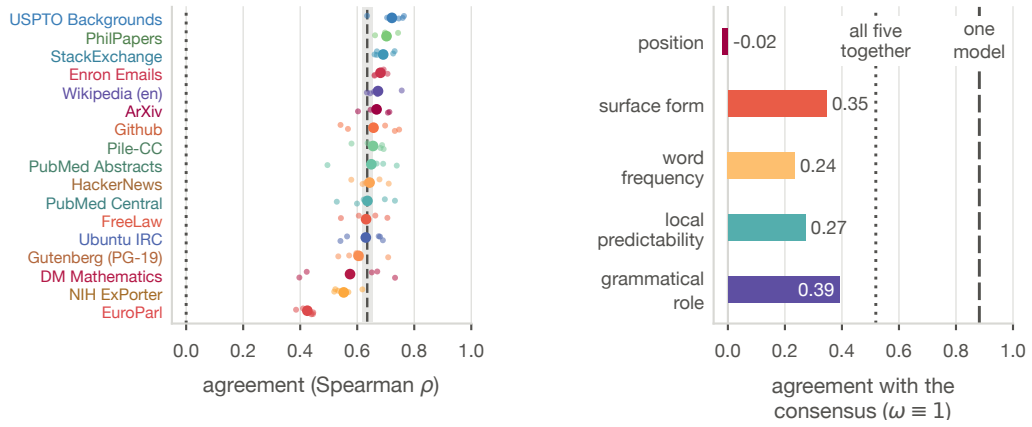

Figure 4: **What the per-token marginal attention tracks.** Spearman correlation per document. *Left:* with the per-token marginal influence of Qwen2.5-0.5B over 85 documents. *Right:* word properties regressing the unweighted ($\omega \equiv 1$) consensus profile over 550 documents.

> *where s, s' index input positions and t indexes output positions, so $[J^{(s)}]_t$ is the block of the Jacobian which locally linearizes the effect of input position s on the logits at output position t.*

By causal masking, we have that $\boldsymbol{j}(X) \in \Delta^{n-1}$, the $(n-1)$-simplex, so $\boldsymbol{j}(X)$ can be seen as a distribution over input positions. Now that it is placed in the same space Δ^{n-1} as the per-token marginal attention $\boldsymbol{a}(X)$ (Definition 2), we can establish the correlation between the two quantities.

4.1 Marginal attention approximates the Jacobian

Experimental evidence. We compare the per-token marginal attention $\boldsymbol{a}(X)$ of Qwen2.5-0.5B, with the weights ω of Section 3, to its per-token marginal influence $\boldsymbol{j}(X)$ on the same document, across 85 natural documents. To approximate the Jacobian we apply Hutchinson's estimator to a random subset of blocks (Section E.1). The two quantities are compared as in Section 3.1: both are projected onto the characters, converted to ranks, residualised against 50 relative-position bins, and compared by Spearman correlation, one document at a time. Figure 4 (left) shows that the per-token marginal attention and the per-token marginal influence agree across all domains (average Spearman ρ of 0.635). Agreement is lowest in the EuroParl domain, which contains substantial non-English text. As a sanity check, agreement of the two for a pair of distinct documents is nearly zero (0.022). As shown in Section E.1, the choice of weights does influence the level of agreement: the simplest setting of uniform weights $\omega \equiv 1$ reduces the agreement to 0.534.

Across models, where Hutchinson's estimate is costly, we approximate the per-token marginal influence by one row of each Jacobian block, the gradient of the log-probability of the observed next token (Section E.1). Over 16 models from 6 families, each model's influence map agrees with the consensus of the other models at 0.772, and the unweighted per-token marginal attention at 0.861 (Table 7). **The link does not need attention at all**: the per-token marginal influence of Mamba (Gu & Dao, 2024) and RWKV-4 (Peng et al., 2023) models, which have no attention mechanism, agrees with the consensus per-token marginal attention of the census similarly to a transformer model (Section E.4).

Derivation. We now motivate this approximation heuristically. For a single causal attention layer with the attention weights held fixed, head h adds $\sum_{s \le t} A^{(h)}_{s,t} W^{(h)}_{OV} \mathbf{x}_s$ to the residual stream at position t, with $W^{(h)}_{OV} := W^{(h)}_O W^{(h)}_V$, so its part of the Jacobian block is $[J^{(s)}_h]_t = A^{(h)}_{s,t} W^{(h)}_{OV}$, and the block $[J^{(s)}]_t$ is the sum of these parts over heads, up to the identity at $s = t$. Applied to $\mathbf{x}_s$, the head's part has norm $A^{(h)}_{s,t} \|W^{(h)}_O v_s\| = \omega^{(h)}_{s,t} A^{(h)}_{s,t}$, exactly the summand of Definition 1. Summing over output positions and heads, up to normalisation,

$$a_s \propto \sum_{t \ge s} \sum_h \left\| [J^{(s)}_h]_t \, \mathbf{x}_s \right\|, \qquad j_s \propto \sum_{t \ge s} \left\| \sum_h [J^{(s)}_h]_t \right\|_F,$$

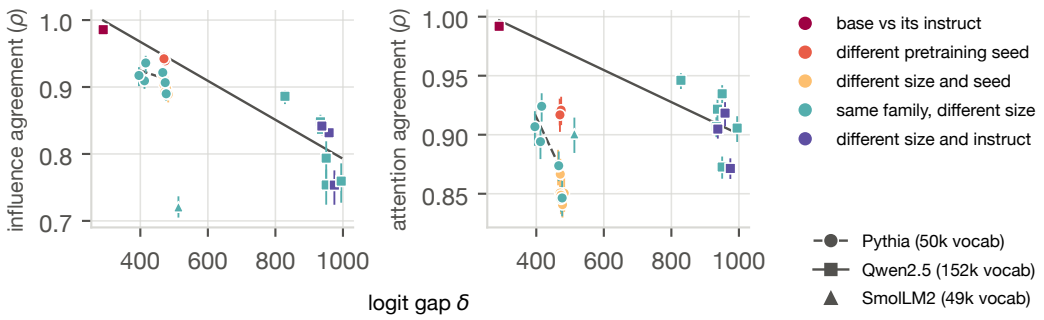

Figure 5: **Within a tokenizer group, pairs with a smaller logit gap δ agree more on per-token marginal influence (left) and per-token marginal attention (right, $\omega \equiv 1$).** Colour: pair class; marker: tokenizer; lines: least-squares fits; bars: 95% intervals. Estimator in Section E.1.

The two differ only in that $\boldsymbol{a}$ sums the heads' norms along $\mathbf{x}_s$ where $\boldsymbol{j}$ takes the Frobenius norm of their sum, so they agree when the heads' contributions do not systematically cancel and $\mathbf{x}_s$ is a generic direction of roughly constant norm. With $\omega \equiv 1$, the first-order term drops each head's gain $\|W_{OV}^{(h)}\mathbf{x}_s\|$, which may explain lower agreement. Section D.1 gives the full argument, and Section D.2 extends it to several layers.

> **Takeaway.** The per-token marginal attention $\boldsymbol{a}(X)$ is an empirically validated approximation to the per-token marginal influence $\boldsymbol{j}(X)$: it is the first-order, per-head part of the Jacobian.

4.2 Jacobian identifiability

Having established that per-token marginal attention $\boldsymbol{a}(X)$ is empirically related to the per-token marginal influence $\boldsymbol{j}(X)$, we now develop a model-agnostic identifiability theory for $\boldsymbol{j}(X)$.

Our theory relies on three assumptions, stated formally in Section G.1 with constants quantified in Section E. *Jacobian smoothness* (Assumption 1) bounds the curvature of the model by a constant M, which can be estimated with Jacobian-vector products. *Preceding-token dependence* (Assumption 2) states that each prediction depends non-trivially on the token immediately before it, and *lag decay* (Assumption 3) states that the dependence of output positions on increasingly "further back" input positions decays toward zero. Our claim is one of statistical near-identifiability (Nelson et al., 2026), with the constants in the hypotheses giving quantitative control over the level of identifiability.

> **Theorem 1** (Near-Identifiability of the Per-Token Marginal Influence, informal)**.** *Let f, f' be two language models satisfying Assumptions 1 to 3. If their next-token logits agree under small perturbations of any single input position,*
>
> $$\left\| f_t(X^{(s)+r\mathbf{u}}) - f'_t(X^{(s)+r\mathbf{u}}) \right\| \le \delta \qquad \text{for all } s \le t,\ 0 \le r < R,\ \|\mathbf{u}\| = 1,$$
>
> *then their per-token marginal influences agree,*
>
> $$\|\boldsymbol{j}(X) - \boldsymbol{j}'(X)\|_1 \le \epsilon,$$
>
> *where ϵ depends on neither n nor d, and $\epsilon \to 0$ as $M, \delta \to 0$.*

Training produces a small logit gap δ: two models fit to the same next-token objective approach the text's conditional distribution as their loss falls. Over a training run, cross-family agreement rises as the loss falls and is near zero until it beats a bigram baseline (Figure 9 and Section E.3).

Experimental evidence. We take the 16 models of Section 4.1 and form 26 tokenizer-sharing pairs: 15 pythia pairs, 10 qwen2.5 pairs and one SmolLM2 pair, evaluated on 100 documents. The logit gap δ of a pair is the norm of the difference of the two models' centred next-token logits, averaged over positions and documents. Agreement (Section 3.1) is computed on each tokenizer group's character axis and applied both to the per-token marginal influence, in the row-gradient approximation of Section 4.1, and to the unweighted ($\omega \equiv 1$) per-token marginal attention. Within a tokenizer group, as the logit gap grows, per-token marginal influence agreement falls, which is consistent with Theorem 1. Validating our empirical argument, per-token marginal attention agreement falls as well.

Table 1: **Eviction scores.** RULER 4096 score over its 13 tasks and documents at 4×, 8× and 16× compression. *Cost:* △ offline work per target, • online work per document (Table 10). *Shade:* rank within a column, darker is better; *Bold:* best method; the uncompressed row is not ranked. *Italics:* our run, where no budget curve or surrogate is published for the target.

	cost		Qwen3-8B			Llama-3.1-8B			phi-4			Qwen3-32B			Mistral-7B		
	offline	online	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×
AdaKV + EA			88.4	72.8	46.5	75.7	45.3	15.2	68.7	44.0	13.8	81.9	63.6	36.4	67.2	34.0	19.9
AdaKV + KVzap	△△		94.0	89.2	49.9	94.5	89.3	63.1	*87.8*	*77.0*	*51.6*	95.4	90.0	50.3	*88.3*	*70.2*	*42.1*
KVzip		••	95.0	**92.3**	59.4	95.3	92.8	**74.2**	91.5	89.6	68.8	**95.6**	**95.6**	**91.8**	91.2	**89.2**	**72.2**
LU-KV + EA	△		*90.6*	*75.2*	*58.4*	95.1	91.8	67.0	***92.4***	*87.9*	*69.5*	*90.2*	*72.6*	*38.2*	91.2	85.1	61.7
HM + EA	△		95.0	91.8	**73.3**	95.1	92.3	66.3	91.5	87.8	68.5	95.2	93.2	79.1	90.5	84.3	65.3
HM + Guide	△	•	**95.1**	91.6	71.3	**95.3**	**93.5**	72.3	92.3	**91.2**	**73.9**	95.4	94.1	79.2	**91.3**	86.7	71.6
Uncompressed			95.4			95.7			91.0			95.7			92.2		

Takeaway. Sufficiently smooth models with similar next-token logits for a given input have their per-token marginal influences guaranteed to be close, under mild assumptions on the dependence patterns in the Jacobian (Theorem 1). Empirically, a smaller logit gap δ implies higher agreement on both per-token marginal influence and per-token marginal attention (Figure 5).

5 Application: KV cache eviction

KV cache eviction bounds the memory of a long context by keeping only part of the stored keys and values (Section 2). Eviction decides how many entries each head keeps and which ones, so we write the retention score of an entry as a per-head budget times a per-entry score:

$$S_{h,s}(X) \;=\; \beta_h \cdot \gamma_{h,s}(X), \tag{1}$$

where $\beta_h \geq 0$ is the budget of head h, the fraction of its entries it keeps, and $\gamma_{h,s}(X) \geq 0$ is the score of entry (h, s) in document X. Head h keeps its $\beta_h n$ highest-scoring entries. Budgets are set per key-value head, and in this section H is the number of key-value heads. We ask whether a static per-model budget can replace document-specific ones, and answer with a budget read from the per-head marginal attention (Section 5.1). To obtain an eviction policy, we pair it with Expected Attention as the score (Section 5.2).

5.1 Head budget from the per-head marginal attention

Head Mass (HM) is the per-head marginal attention of Definition 3 with three choices taken from KVzip, a state-of-the-art eviction method that is expensive at run time, since it pays a second target forward per document. First, Head Mass is read on the **reconstruction queries**, the positions t of a second pass in which the model repeats the context, since eviction must keep what later queries attend to. Second, KVzip's per-entry score replaces the mean over queries of Definition 1: an entry scores the largest value-weighted attention it receives from those queries, normalised by the norm of the attention input, since an entry that any one later query needs must be kept (exact form in Equation (3), Appendix F.1). Third, Head Mass applies KVzip's global cut, which keeps the fraction B of all entries, so head h counts the fraction $m_h(X)$ of its own entries KVzip would keep.

The head share h is $p_h(X) = m_h(X) / \sum_j m_j(X)$. We average it over a set $\mathcal{D}_{\text{pre}}$ of 200 pretraining documents disjoint from benchmarks (Appendix F.1), and the budget spends B with this proportion:

$$\beta_h = H B\, p_h, \qquad p_h = \mathbb{E}_{X \sim \mathcal{D}_{\text{pre}}}\big[p_h(X)\big], \tag{2}$$

clipped to a floor of 0.01 and a cap of 1. The budgets are then rescaled so that the heads spend exactly B on average. The per-head profile is a property of the model (Section 3.1) and stays so on the reconstruction queries (Table 8), so the budget is computed once per model.

5.2 Results

Experimental setting. We evaluate Equation (1) on Qwen3-8B, Qwen3-32B, Llama-3.1-8B-Instruct, phi-4 and Mistral-7B-Instruct-v0.3 (Yang et al., 2025; Grattafiori et al., 2024; Abdin et al.,

2024; Jiang et al., 2023). Each method pairs a budget with a score. AdaKV + EA (Feng et al., 2025) splits the budget inside each layer per document and scores with Expected Attention (EA) (Devoto et al., 2025), a training-free scorer. KVzip (Kim et al., 2025) takes both the budget and the score from a second target forward per document. AdaKV + KVzap (Jegou & Jeblick, 2026) keeps the AdaKV budget and scores with a KVzip surrogate trained per target. LU-KV + EA (Tang et al., 2026) takes its budget from an offline per-head table calibrated per target and scorer, and scores with EA. HM + EA takes the Head Mass budget and scores with EA. HM + Guide keeps the Head Mass budget and tilts EA, head by head, by a linear map fitted once per target (Appendix F.4). The map predicts KVzip's score from the marginal attention of Qwen2.5-0.5B-Instruct, a model about fifteen times smaller than the smallest target. It probes whether a smaller model's attention carries the eviction signal and is not meant as a practical method. We report RULER (Hsieh et al., 2024) (13 tasks, 6500 documents) at context 4096 and compression $4\times$, $8\times$ and $16\times$, the factor by which the cache shrinks. Scores at context 8192 ($8\times$ and $16\times$) and on LongBench (Bai et al., 2024) are in Tables 13 and 14, per-task scores in Section F.5, and costs in Table 10.

Result analysis. *Budget.* Table 1 shows that the HM budget accounts for most of the gap from standard EA to KVzip. In fact, with the same scorer, HM + EA considerably improves on its baseline on every target and setting, closing most of the gap to KVzip without the double-forward on the target. HM + EA is also ahead of AdaKV + KVzap in all but one cell, and its budget is built in minutes without representation access where a KVzap surrogate trains for hours (Table 10) and needs light online computation. Even though it has extra scorer calibration, LU-KV + EA is level with HM + EA on phi-4 and Llama-3.1-8B, close on Mistral-7B, and far behind on both Qwen3 targets. Averaging over pretraining text loses little: the HM budget agrees closely, head for head, with the keep rates KVzip realises on held-out documents (Appendix F.2). The query set is crucial: set by the **reading queries**, the document's own positions, a proportional budget (Equation (4)) consistently loses to the baseline AdaKV + EA with no fixed budget at all (Table 12), so a profile can be stable across documents and still be the wrong budget. The cut matters less (Appendix F.3). *Score.* Tilting EA by the guide's map raises HM + EA in all but two columns, so a smaller model's attention can carry part of the eviction signal.

> **Takeaway.** Which heads matter for eviction is a property of the model: a per-head marginal attention profile computed once recovers most of the gain of per-document signals, and a score tilted by a smaller model can further improve it.

6 Conclusions

We studied attention in language models through its marginals and found that one structure per document carries two different signals. Reduced over heads, the per-token marginal attention is a property of the text: models that differ in lab, data, size, tokenizer and architecture agree on which input tokens are attended. Reduced over tokens, the per-head marginal attention is a property of the model: it is stable across documents and private to each model. The per-token profile is shared because it tracks a Jacobian statistic that similar predictors must share: smooth models with close next-token distributions have close input-output Jacobians, with marginal attention approximating its the blockwise norms. The only data assumption is that its discrete conditional distributions admit a smooth approximation good enough for independently trained models to agree on the next token. To our knowledge this is the first formal result of the kind discussed by the Platonic Representation Hypothesis (Huh et al., 2024), but it does not imply semantic convergence (Nelson et al., 2026): the shared profile separates documents without aligning strongly with classical syntactic structure. The per-head profile is stable enough to be computed once per model, offline, on pretraining text, and used as a budget for KV cache eviction. At inference only a per-token score remains, and since that score is shared across models it can be supplied by a much smaller one. The resulting method is competitive with those that recompute the budget on every document or train it per target.

Limitations. While we have provided empirical evidence for the convergence of marginal attention across documents (per-head) and models (per-token), our theory applies only to the marginal influence. The empirical step linking the marginal influence to the marginal attention remains necessary, and a study of the theoretical assumptions necessary to ground that step would be of significant interest. The census covers prevalently English text of at most 2048 tokens, and the guide tilt costs one extra forward of a small model per document, which standard serving stacks do not provide.

AI USE STATEMENT

We used AI tools (Claude and Claude Code, running Claude Fable 5.1 and Claude Opus 5), during this work. The authors set the questions, the claims and the experiments, reviewed every contribution of the tools before it was kept, and take responsibility for the final content.

Experiments. The authors wrote the core library by hand (with autocompletion). AI tools were used to write scripts and manage experiments. Results and code were reviewed by an author.

Writing and presentation. AI tools were used to integrate raw results, perform bulk edits and layout changes, and edit the paper. Figure 1 was drawn by an AI tool from a sketch by the authors.

Theory. The definitions, hypotheses, theorems and proofs were proposed by the authors. AI tools were used to find gaps in the argument and improve the exposition. Every definition, statement and proof was checked by the authors.

Model citations. Given the large number of models used in this work, AI tools were used to scrape the citations for each model from the corresponding HuggingFace repository. These citations were checked by an author.

Supplementary code. AI tools assisted in compiling the final supplementary material, which mainly involved automated removal of experiments not included in the final paper.

REPRODUCIBILITY STATEMENT

The code is submitted as supplementary material and will be released as a public repository upon acceptance. Every number in the paper can be traced to the script that produced it. One library implements the estimator (the marginal attention hooks, the projection onto the character axis, ranking, positional residualization and the decomposition of Section 3); each experiment is a short script that calls it, reads its configuration from a versioned config file, and writes results on file. Every figure and table is produced from those results files by a notebook committed with its outputs. Each experiment has tests.

All 67 census models are public checkpoints, listed with their sources in Appendix A.2. The corpus is built from public Pile subsets and three synthetic domains by a script that freezes the sampled documents and records their hash (Appendix A.1); the documents will be included with the code. The KV eviction experiments use the kvpress library (Devoto et al., 2025) and its RULER and LongBench harnesses unchanged, with our budget and scorer added as one more press. Section B reports the controls, the seed ceiling and the sensitivity of the headline to the estimator's choices; Section E reports the measured constants behind the assumptions of Section 4, and Section G gives complete proofs with their assumptions stated.

REFERENCES

01.AI, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yanpeng Li, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.AI, 2024. URL https://arxiv.org/abs/2403.04652.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th An-*

nual Meeting of the Association for Computational Linguistics, pp. 4190–4197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385/.

Muhammad Adnan, Akhil Arunkumar, Gaurav Jain, Prashant J. Nair, Ilya Soloveychik, and Purushotham Kamath. Keyformer: Kv cache reduction through key tokens selection for efficient generative inference. In P. Gibbons, G. Pekhimenko, and C. De Sa (eds.), *Proceedings of Machine Learning and Systems*, volume 6, pp. 114–127, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/file/48fecef47b19fe501d27d338b6d52582-Paper-Conference.pdf.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM2: When smol goes big – data-centric training of a small language model, 2025. URL https://arxiv.org/abs/2502.02737.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The Falcon series of open language models, 2023. URL https://arxiv.org/abs/2311.16867.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. 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. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 225–236. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/01ded4259d101feb739b06c399e9cd9c-Paper.pdf.

Lorenzo Basile, Valentino Maiorca, Luca Bortolussi, Emanuele Rodolà, and Francesco Locatello. Residual transformer alignment with spectral decomposition. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=z37LCgSIzI.

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. Head pursuit: Probing attention specialization in multimodal transformers. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 112351–112383. Curran Associates, Inc., 2025b. doi: 10.52202/085713-3752. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a30e060cde25f9c1e73924d2f996542f-Paper-Conference.pdf.

Marco Bellagente, Jonathan Tow, Dakota Mahan, Duy Phung, Maksym Zhuravinskyi, Reshinth Adithyan, James Baicoianu, Ben Brooks, Nathan Cooper, Ashish Datta, Meng Lee, Emad Mostaque, Michael Pieler, Nikhil Pinnaparju, Paulo Rocha, Harry Saini, Hannah Teufel, Niccolo Zanichelli, and Carlos Riquelme. Stable LM 2 1.6B technical report, 2024. URL https://arxiv.org/abs/2402.17834.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

BigScience Workshop, Teven Le Scao, et al. BLOOM: A 176B-parameter open-access multilingual language model, 2023. URL https://arxiv.org/abs/2211.05100.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow, March 2021. URL https://doi.org/10.5281/zenodo.5297715.

Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers, 2020. URL https://arxiv.org/abs/1908.04211.

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

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.

John Dang et al. Aya Expanse: Combining research breakthroughs for a new multilingual frontier, 2024. URL https://arxiv.org/abs/2412.04261.

DeepSeek-AI. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model, 2024. URL https://arxiv.org/abs/2405.04434.

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.

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. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 113152–113188. Curran Associates, Inc., 2025. doi: 10.52202/085713-3773. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a40ff56daab9f4808b1e18350c8a11ce-Paper-Conference.pdf.

Yossi Gandelsman, Alexei Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 18395–18416, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5085a2b8c298edeadc46b9ffe6df5f64-Paper-Conference.pdf.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. URL https://arxiv.org/abs/2101.00027.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Paolo Glorioso, Quentin Anthony, Yury Tokpanov, Anna Golubeva, Vasudev Shyam, James Whittington, Jonathan Pilault, and Beren Millidge. The Zamba2 suite: Technical report, 2024. URL https://arxiv.org/abs/2411.15242.

Granite Team, IBM. Granite 3.0 language models. https://github.com/ibm-granite/granite-3.0-language-models/, October 2024.

Aaron Grattafiori et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Fabian Gröger, Shuo Wen, and Maria Brbic. Revisiting the platonic representation hypothesis: An aristotelian view. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=uz0gAAYydl.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. DeepSeek-Coder: When the large language model meets programming – the rise of code intelligence, 2024. URL https://arxiv.org/abs/2401.14196.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies, 2024. URL https://arxiv.org/abs/2404.06395.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. Qwen2.5-Coder technical report, 2024. URL https://arxiv.org/abs/2409.12186.

Simon Jegou and Maximilian Jeblick. Kvzap: Fast, adaptive, and faithful kv cache pruning, 2026. URL https://arxiv.org/abs/2601.07891.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023. URL https://arxiv.org/abs/2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

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. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 167563–167591. Curran Associates, Inc., 2025. doi: 10.52202/085713-5585. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f4eaa4b8f2d08edb3f0af990d56134ea-Paper-Conference.pdf.

Jang-Hyun Kim, Dongyoon Han, and Sangdoo Yun. Fast kvzip: Efficient and accurate llm inference with gated kv eviction, 2026. URL https://arxiv.org/abs/2601.17668.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Wonbeom Lee, Jungi Lee, Junghwan Seo, and Jaewoong Sim. InfiniGen: Efficient generative inference of large language models with dynamic KV cache management. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pp. 155–172, Santa Clara, CA, July 2024. USENIX Association. ISBN 978-1-939133-40-3. URL https://www.usenix.org/conference/osdi24/presentation/lee.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, Chongyan Zhu, Xiaoyi Ren, Chao Li, Yifan Ye, Peng Liu, Lihuan Zhang, Hanshu Yan, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model, 2024a. URL https://arxiv.org/abs/2410.05993.

Junjie Li, Jiong Lou, and Jie Li. Proxykv: Cross-model proxy pruning for efficient long-context llm inference, 2026. URL https://arxiv.org/abs/2605.16360.

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. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 22947–22970. Curran Associates, Inc., 2024b. doi: 10.52202/079017-0722. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/28ab418242603e0f7323e54185d19bde-Paper-Conference.pdf.

Jinglin Liang, Jin Zhong, Shuangping Huang, Yunqing Hu, Huiyuan Zhang, Huifang Li, Lixin Fan, and Hanlin Gu. Order-level attention similarity across language models: A latent commonality. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 121233–121269. Curran Associates, Inc., 2025. doi: 10.52202/085713-4044. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/af800ac69aea2fb5c4bcf8c6a5f3c79d-Paper-Conference.pdf.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024a. URL https://arxiv.org/abs/2310.03744.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, January 2024b.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric P.

Xing. LLM360: Towards fully transparent open-source LLMs, 2023. URL https://arxiv.org/abs/2312.06550.

Anton Lozhkov et al. StarCoder 2 and The Stack v2: The next generation, 2024. URL https://arxiv.org/abs/2402.19173.

Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 55394–55414. Curran Associates, Inc., 2023. doi: 10.52202/075280-2418. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ad5fa03c906ca15905144ca3fbf2a768-Paper-Conference.pdf.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. SmolVLM: Redefining small and efficient multimodal models, 2025. URL https://arxiv.org/abs/2504.05299.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/2c601ad9d2ff9bc8b282670cdd54f69f-Paper.pdf.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a7a3d70c6d17a73140918996d03c014f-Paper.pdf.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SrC-nwieGJ.

Luca Moschella, Laura Manduchi, and Ozan Sener. Learning to evict from key-value cache, 2026. URL https://arxiv.org/abs/2602.10238.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoE: Open mixture-of-experts language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=xXTkbTBmqq.

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation, 2024. URL https://arxiv.org/abs/2407.14679.

Walter Nelson, Marco Fumero, THEOFANIS KARALETSOS, and Francesco Locatello. Statistical and structural identifiability in representation learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 151985–152020, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/f67e5f99b23b108a3a6665f410034bcd-Paper-Conference.pdf.

Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. Asif: Coupled data turns unimodal models to multimodal without training. In A. Oh, T. Naumann, A. Globerson, K. Saenko,

M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 15303–15319. Curran Associates, Inc., 2023. doi: 10.52202/075280-0673. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3186591903d9db31770ad131adb5ceb4-Paper-Conference.pdf.

NVIDIA. NVIDIA Nemotron Nano 2: An accurate and efficient hybrid Mamba-Transformer reasoning model, 2025. URL https://arxiv.org/abs/2508.14444.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 35748–35760, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/3c1fe56b043848b211030c202764c6a7-Paper-Conference.pdf.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. RWKV: Reinventing RNNs for the transformer era. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14048–14077, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.936. URL https://aclanthology.org/2023.findings-emnlp.936/.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

Sharath Turuvekere Sreenivas, Saurav Muralidharan, Raviraj Joshi, Marcin Chochowski, Ameya Sunil Mahabaleshwarkar, Gerald Shen, Jiaqi Zeng, Zijia Chen, Yoshi Suhara, Shizhe Diao, Chenhan Yu, Wei-Chun Chen, Hayley Ross, Oluwatobi Olabiyi, Ashwath Aithal, Oleksii Kuchaiev, Daniel Korzekwa, Pavlo Molchanov, Mostofa Patwary, Mohammad Shoeybi, Jan Kautz, and Bryan Catanzaro. LLM pruning and distillation in practice: The Minitron approach, 2024. URL https://arxiv.org/abs/2408.11796.

Hanlin Tang, Yang Lin, Jing Lin, Qingsen Han, Danning Ke, Shikuan Hong, Yiwu Yao, and Gongyi Wang. Razorattention: Efficient kv cache compression through retrieval heads. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 16632–16646, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/2a98af4fea6a24b73af7b588ca95f755-Paper-Conference.pdf.

Ziyao Tang, Pengkun Jiao, Xinhang Chen, Wei Liu, Shiyong Li, and Jingjing Chen. Predicting future utility: Global combinatorial optimization for task-agnostic kv cache eviction, 2026. URL https://arxiv.org/abs/2602.08585.

Team GLM et al. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools, 2024. URL https://arxiv.org/abs/2406.12793.

TII Team. The Falcon 3 family of open models. https://huggingface.co/tiiuae/Falcon3-7B-Base, December 2024.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.

Evan Pete Walsh et al. 2 OLMo 2 furious. In *Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=2ezugTT9kU.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 billion parameter autoregressive language model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 62143–62156, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9b77f07301b1ef1fe810aae96c12cb7b-Paper-Conference.pdf.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5e5fd18f863cbe6d8ae392a93fd271c9-Paper-Conference.pdf.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. Duoattention: Efficient long-context llm inference with retrieval and streaming heads. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 37228–37253, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5c1ddd2e59df46fd2aa85c833b1b36ed-Paper-Conference.pdf.

Yongzhong Xu. Pattern selectivity is not task-causal structure: A cross-architecture mechanistic study of composed-task circuits in 1b-class language models, 2026. URL https://arxiv.org/abs/2606.05378.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Haobo Zhang, Zhenhua Xu, Junxian Li, Shangfeng Sheng, Dezhang Kong, and Meng Han. Attndiff: Attention-based differential fingerprinting for large language models, 2026. URL https://arxiv.org/abs/2604.05502.

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/.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. TinyLlama: An open-source small language model, 2024. URL https://arxiv.org/abs/2401.02385.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang "Atlas" Wang, and Beidi Chen. H2o: Heavy-hitter oracle for efficient generative inference of large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34661–34710. Curran Associates, Inc., 2023. doi: 10.52202/075280-1506. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6ceefa7b15572587b78ecfcebb2827f8-Paper-Conference.pdf.

Yi Zhao, Zuchao Li, and Hai Zhao. IAM: Efficient inference through attention mapping between different-scale LLMs. 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. 19522–19533, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.959. URL https://aclanthology.org/2025.acl-long.959/.

Yi Zhao, Yajuan Peng, Nguyen Cam-Tu, Zuchao Li, Xiaoliang Wang, Hai Zhao, and Xiaoming Fu. Smallkv: Small model assisted compensation of kv cache compression for efficient llm inference. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 48637–48665. Curran Associates, Inc., 2025b. doi: 10.52202/085713-1624. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/45790132b9f91b5787232e1cbe0712fc-Paper-Conference.pdf.

Jinguo Zhu et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

A Census details

A.1 Data

The census corpus, *mix1000*, draws 50 documents from each of 20 domains. Most domains are subsets of the Pile (Gao et al., 2020), taken from its release without the copyrighted subsets (monology/pile-uncopyrighted). The other three are generated text with little linguistic structure: one short sentence repeated, a random string over the four DNA letters, and rows of random ASCII-art symbols. The main text uses only the natural documents.

Pile documents shorter than 2500 characters are skipped, and the documents of each domain are drawn by seeded reservoir sampling over one pass of the stream. Each document contributes one window of 12,000 characters, taken from a random position inside it. Extraction stops at 2048 tokens per model, and each document is scored only on the characters that every model covers. Figure 7 shows the length of every document by domain.

A second corpus, *mix10k*, follows the same recipe with ten times as many documents per domain. It samples Pile rows that the first corpus never reads, and its generated domains use a different seed, so the two corpora are practically disjoint. Ubuntu IRC is capped lower than the other domains, because the Pile holds too few of its documents above the length floor.

Table 2 applies the estimator of Section 3.1 to each corpus, on all of its documents, on the natural ones and on the generated ones. Agreement is lower on generated text; on the larger corpus, scored with a twelve-model subset of the roster, it stays high, so the headline does not depend on one draw of documents. Figure 6 shows the agreement of each natural document by domain, and how it relates to document length within each domain. Figure 8 shows the per-domain agreement on the three generated domains.

Table 2: **Per-token agreement on every document set.** Each row applies the same estimator to the value-weighted per-token marginal attention on one set of documents: *mix1000* (20 kinds of text), its natural and generated subsets, and *mix10k*. ± is half the 95% bootstrap interval over documents.

document set	models	documents	consensus of the other models	between two models
mix1000, all domains	67	1000	0.8443 ± 0.0042	0.7253 ± 0.0065
mix1000, natural domains	67	850	0.8673 ± 0.0032	0.7607 ± 0.0051
mix1000, generated domains	67	150	0.6114 ± 0.0136	0.4093 ± 0.0162
mix10k, all domains	12	9792	0.7977 ± 0.0021	0.6787 ± 0.0028
mix10k, natural domains	12	8292	0.8357 ± 0.0012	0.7277 ± 0.0018
mix10k, generated domains	12	1500	0.4990 ± 0.0064	0.3362 ± 0.0060

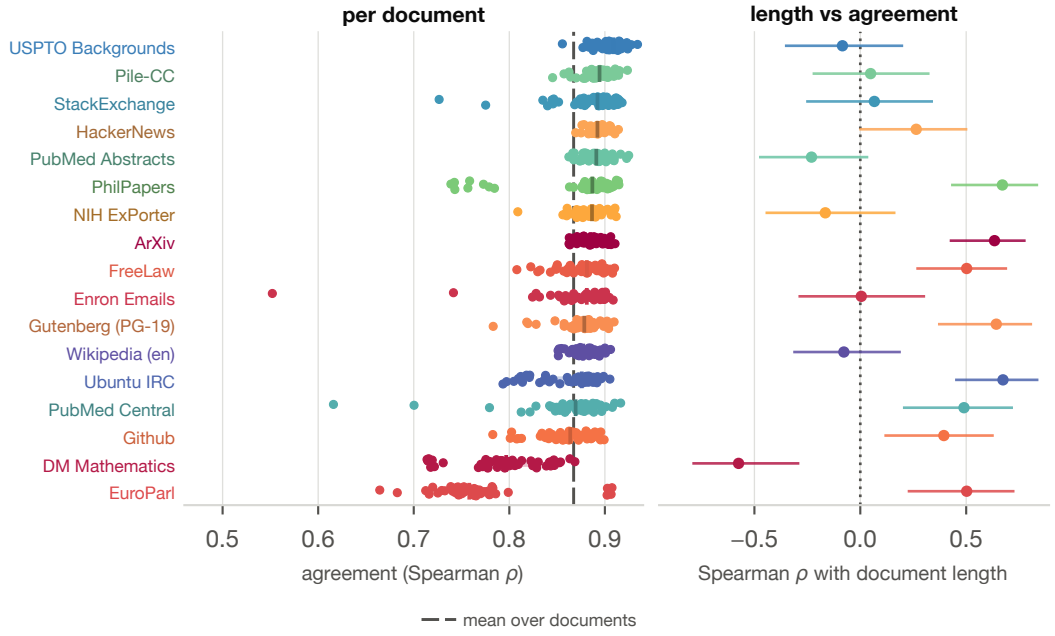

Figure 6: **Agreement is high in every natural domain; its link to length varies.** *Left:* per-document per-token agreement, tick the median, dashed line the mean. *Right:* Spearman correlation of length with agreement, 95% intervals. Setting in Appendix A.1.

A.2 Models

Table 4 lists every model of the census, with its family, architecture and size, and Table 3 gives the census summary.

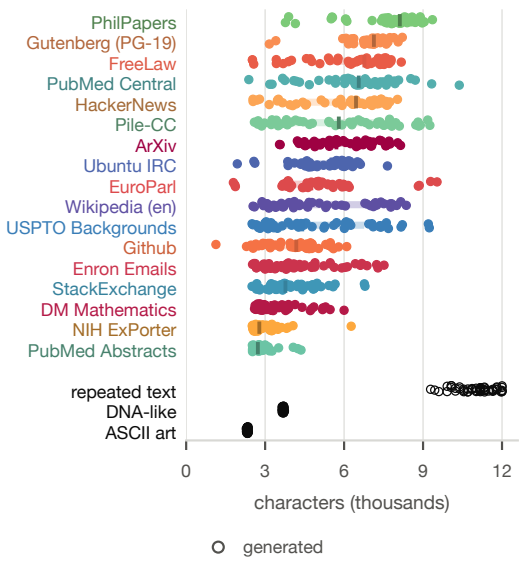

Figure 7: **Document length varies about threefold across domains, from PhilPapers and Gutenberg (PG-19) down to NIH ExPorter and PubMed Abstracts.** Dots are documents, ticks medians, bars interquartile ranges. Setting in Appendix A.1.

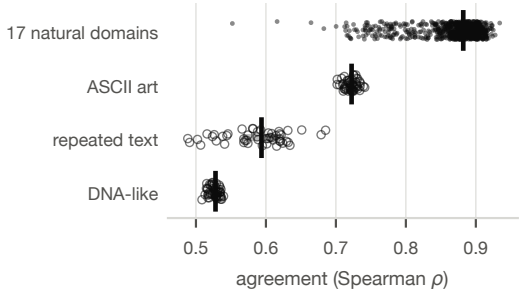

Figure 8: **Models agree less on generated text than on natural text.** Per-document per-token agreement of each model with the consensus of the others, averaged over models; tick the median, open markers generated domains. Rows scored as in Table 2.

Table 3: **Per-token census summary.** 67 models, 850 natural documents. *Upper:* agreement with the other 66 models and between two models, with the different-document control and the pythia-410m seed ceiling. *Lower:* agreement between two models by pair class. On the first row, ± is half the 95% bootstrap interval over documents.

		vs the other 66	*two models*		
whole census	agreement	0.867 ± 0.003	0.761 ± 0.005		
	different document (control)	0.015	0.014		
	10 pythia-410m seeds (ceiling)	0.942	0.900		
		pairs	*mean*	*median*	*5th to 95th*
pair class	base vs its instruct	12	0.968	0.976	[0.924, 0.993]
	vision vs its text model	3	0.941	0.944	[0.928, 0.951]
	same family, different size	22	0.889	0.904	[0.807, 0.945]
	different pretraining seed	225	0.821	0.849	[0.628, 0.933]
	same architecture, different lab	28	0.788	0.784	[0.693, 0.896]
	different lab and data	85	0.755	0.777	[0.547, 0.868]

Table 4: **Every model used.** Repository id, parameter count in billions (B) and the report the model card names (ref.); alphabetical by repository, down the left block and then down the right. 69 models: 67 form both the per-token census and the per-head analyses, and Qwen3-32B and Mistral-7B-Instruct-v0.3 are used only as eviction targets. Eviction targets: Qwen3-8B, Llama-3.1-8B-Instruct, phi-4, Qwen3-32B, Mistral-7B-Instruct-v0.3; guide: Qwen2.5-0.5B-Instruct. The Pythia pretraining-seed pool (10 seeds at each of 5 sizes) is not listed.

model	B	ref.	model	B	ref.
01-ai/Yi-6B	6.1	(01.AI et al., 2024)	meta-llama/Llama-3.1-8B	8.0	(Grattafiori et al., 2024)
01-ai/Yi-6B-Chat	6.1	(01.AI et al., 2024)	meta-llama/Llama-3.1-8B-Instruct	8.0	(Grattafiori et al., 2024)
01-ai/Yi-9B	8.8	(01.AI et al., 2024)	microsoft/phi-4	14.7	(Abdin et al., 2024)
allenai/OLMo-2-0425-1B	1.5	(Walsh et al., 2025)	mistralai/Mistral-7B-Instruct-v0.1	7.2	(Jiang et al., 2023)
allenai/OLMo-2-0425-1B-Instruct	1.5	(Walsh et al., 2025)	mistralai/Mistral-7B-Instruct-v0.2	7.2	(Jiang et al., 2023)
allenai/OLMoE-1B-7B-0924	6.9	(Muennighoff et al., 2025)	mistralai/Mistral-7B-Instruct-v0.3	7.2	(Jiang et al., 2023)
bigcode/starcoder2-3b	3.0	(Lozhkov et al., 2024)	mistralai/Mistral-7B-v0.1	7.2	(Jiang et al., 2023)
bigscience/bloom-1b7	1.7	(BigScience Workshop et al., 2023)	mistralai/Mixtral-8x7B-v0.1	46.7	(Jiang et al., 2024)
CohereLabs/aya-expanse-8b	8.0	(Dang et al., 2024)	nvidia/Mistral-NeMo-Minitron-8B-Base	8.4	(Sreenivas et al., 2024; Muralidharan et al., 2024)
deepseek-ai/deepseek-coder-1.3b-base	1.3	(Guo et al., 2024)	openai/gpt-oss-120b	116.8	(OpenAI, 2025)
deepseek-ai/DeepSeek-V2-Lite	15.7	(DeepSeek-AI, 2024)	openai/gpt-oss-20b	21.5	(OpenAI, 2025)
EleutherAI/gpt-j-6b	6.1	(Wang & Komatsuzaki, 2021)	openbmb/MiniCPM3-4B	4.1	(Hu et al., 2024)
EleutherAI/gpt-neo-1.3B	1.4	(Black et al., 2021)	OpenGVLab/InternVL3-8B-hf	7.9	(Zhu et al., 2025)
EleutherAI/pythia-1.4b	1.5	(Biderman et al., 2023)	Qwen/Qwen2.5-0.5B	0.5	(Qwen et al., 2025)
EleutherAI/pythia-2.8b	2.9	(Biderman et al., 2023)	Qwen/Qwen2.5-0.5B-Instruct	0.5	(Qwen et al., 2025)
EleutherAI/pythia-410m	0.5	(Biderman et al., 2023)	Qwen/Qwen2.5-1.5B	1.5	(Qwen et al., 2025)
EleutherAI/pythia-6.9b	7.0	(Biderman et al., 2023)	Qwen/Qwen2.5-1.5B-Instruct	1.5	(Qwen et al., 2025)
facebook/opt-1.3b	1.3	(Zhang et al., 2022)	Qwen/Qwen2.5-32B	32.8	(Qwen et al., 2025)
facebook/opt-6.7b	6.7	(Zhang et al., 2022)	Qwen/Qwen2.5-32B-Instruct	32.8	(Qwen et al., 2025)
google/gemma-2-2b	2.6	(Gemma Team, 2024)	Qwen/Qwen2.5-3B	3.1	(Qwen et al., 2025)
google/gemma-2-2b-it	2.6	(Gemma Team, 2024)	Qwen/Qwen2.5-7B	7.6	(Qwen et al., 2025)
google/gemma-2-9b	9.2	(Gemma Team, 2024)	Qwen/Qwen2.5-7B-Instruct	7.6	(Qwen et al., 2025)
google/gemma-2-9b-it	9.2	(Gemma Team, 2024)	Qwen/Qwen2.5-Coder-7B	7.6	(Hui et al., 2024)
google/gemma-3-1b-it	1.0	(Gemma Team, 2025)	Qwen/Qwen2.5-VL-7B-Instruct	8.3	(Bai et al., 2025)
google/gemma-3-270m	0.3	(Gemma Team, 2025)	Qwen/Qwen3-32B	32.8	(Yang et al., 2025)
HuggingFaceTB/SmolLM2-1.7B	1.7	(Allal et al., 2025)	Qwen/Qwen3-8B	8.2	(Yang et al., 2025)
HuggingFaceTB/SmolLM2-1.7B-Instruct	1.7	(Allal et al., 2025)	rhymes-ai/Aria	25.3	(Li et al., 2024a)
HuggingFaceTB/SmolLM2-360M	0.4	(Allal et al., 2025)	stabilityai/stablelm-2-1_6b	1.6	(Bellagente et al., 2024)
HuggingFaceTB/SmolLM2-360M-Instruct	0.4	(Allal et al., 2025)	THUDM/glm-4-9b-hf	9.4	(Team GLM et al., 2024)
HuggingFaceTB/SmolVLM-Instruct	2.2	(Marafioti et al., 2025)	tiiuae/falcon-7b	7.2	(Almazrouei et al., 2023)
ibm-granite/granite-3.1-2b-base	2.5	(Granite Team, IBM, 2024)	tiiuae/Falcon3-7B-Base	7.5	(TII Team, 2024)
ibm-granite/granite-3.1-8b-base	8.2	(Granite Team, IBM, 2024)	tiiuae/Falcon3-7B-Instruct	7.5	(TII Team, 2024)
ibm-granite/granite-3.1-8b-instruct	8.2	(Granite Team, IBM, 2024)	TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T	1.1	(Zhang et al., 2024)
llava-hf/llava-v1.6-mistral-7b-hf	7.6	(Liu et al., 2024a;b)	Zyphra/Zamba2-1.2B	1.2	(Glorioso et al., 2024)
LLM360/Amber	6.7	(Liu et al., 2023)			

B PER-TOKEN MARGINAL ATTENTION: ADDITIONAL EVIDENCE

B.1 AGREEMENT BY PAIR CLASS

The right part of Figure 2 groups model pairs by what they share and gives the mean agreement of the per-token profile in each class (850 natural documents, after removing position). A base model and its instruct version agree at 0.9684 (12 pairs), and a vision-language model and the checkpoint it was built from at 0.9406 (3 pairs). Two models of the same family and data at different sizes agree at 0.8894 (22 pairs), and two pythia models of the same size that differ only in the pretraining seed at 0.8213 (225 pairs), so post-training moves the per-token profile less than a change of seed; the seed pairs are scored on a longer axis (median 5676 characters, compared with 4784.5), so their level is not directly comparable. The ten pythia-410m reruns agree with the average of the other nine at 0.9421 (Section B.5). The same architecture from different labs agrees at 0.7880 (28 pairs) and different labs, data and tokenizers at 0.7551 (85 pairs), a gap smaller than the spread within either class (standard deviation 0.060 and 0.088). Over all pairs of the census the mean agreement is 0.7607 $[0.7556, 0.7658]$, below the 0.8673 of one model compared with the consensus of the others, since averaging over models removes model-specific deviations.

B.2 AGREEMENT SATURATES AT ABOUT TWENTY MODELS

Agreement with a consensus of k models rises quickly with k and then saturates. On the 850 natural documents, after removing position, the per-token marginal attention of a model agrees with a consensus of k other models at 0.7593 for $k = 1$, 0.8432 for $k = 5$, 0.8567 for $k = 10$, 0.8630 for $k = 20$, and 0.8673 with the whole roster. The size of the roster is therefore not what produces the headline.

B.3 A DIFFERENT DOCUMENT GIVES NEAR-ZERO AGREEMENT

Scored against the consensus computed on a different document, the per-token marginal attention agrees at 0.0155 with the other models and at 0.0139 between two models (natural documents; 0.0128 and 0.0114 for the unweighted variant, $\omega \equiv 1$). The agreement of the main text is about the document at hand, not a generic shape shared by all attention profiles.

B.4 ROBUSTNESS: PUNCTUATION, WHITESPACE AND HEAD WEIGHTING

Table 5 removes classes of characters from the axis before ranking and scores the per-token marginal attention on the 850 natural documents again. Deleting punctuation raises agreement from 0.8673 to 0.8705, and deleting punctuation and whitespace raises it to 0.8734 with 87.6% of the axis left. Punctuation alone, 5.0% of the axis, gives 0.8648. Whitespace alone is the only low row, at 0.7336, and it covers 2.3% of the axis. The unweighted variant ($\omega \equiv 1$) gives 0.8605 on the same documents. On the 1000-document census with the unweighted variant, deleting punctuation and whitespace also raises agreement (0.8459, compared with 0.8396 undeleted), but deleting punctuation alone lowers it slightly (0.8361) and keeping punctuation alone raises it (0.8524).

B.5 THE SEED CEILING

The seed ceiling is measured on 50 pythia pretraining-seed reruns, 10 at each of 5 sizes, never pooled across sizes. On the 850 natural documents, after removing position, each rerun agrees with the average of the other nine of its size at 0.7810 (14M), 0.8674 (31M), 0.9140 (70M), 0.9491 (160M) and 0.9421 (410M). The census headline of 0.8673 should be read against the two largest pools, not against 1.

B.6 THE SCRAMBLING LADDER

Within each tokenizer group, agreement orders the conditions as real text $>$ shuffled words $>$ shuffled characters $>$ random tokens (1000-document census, unweighted variant): in the pythia group 0.8277 on real text, 0.7201 on shuffled words and 0.5974 on random token ids. The conditions are not length-matched, since the 2048-token cap spans fewer characters on scrambled text, so the gap

Table 5: **Deleting punctuation and whitespace from the character axis does not lower agreement.** Each row deletes one class of characters from the axis, or keeps only that class, and compares the per-token marginal attention of 67 models on 850 natural documents: one model with the consensus of the other 66, and two models with each other. Whitespace only is the one low row; it keeps a character only where every model's token is whitespace, which leaves 2.3% of the axis and no character at all in 205 of the 850 documents. The two rows under the rule are references: a different document, and the unweighted variant ($\omega \equiv 1$) on all characters.

characters on the axis	axis (%)	documents	vs the other 66	two models
all characters	100.0	850	0.8673	0.7607
punctuation deleted	90.8	850	0.8705	0.7664
punctuation and whitespace deleted	87.6	850	0.8734	0.7711
punctuation only	5.0	848	0.8648	0.7601
whitespace only	2.3	645	0.7336	0.6169
a different document (control)	100.0	850	0.0155	0.0139
unweighted ($\omega \equiv 1$), all characters	100.0	850	0.8605	0.7491

sizes below the shuffled-words condition are provisional; the ordering of real text above shuffled words is not affected.

C PER-HEAD MARGINAL ATTENTION: ADDITIONAL EVIDENCE

Agreement by pair class. The right part of Figure 3 groups model pairs by what they share, after averaging each per-head profile over the 170 documents and binning it into ten relative-depth bins. Pairs in the class *base vs derived model* (base and instruct, or a vision model and its backbone) agree at 0.9900 (17 pairs), *same family* (a different size or version) at 0.7829 (55 pairs), *same architecture, different family* at 0.8198 (108 pairs), and *different architecture* at 0.3728 (2031 pairs), compared with -0.0003 for *permuted heads*, where each model's heads are permuted before binning (2211 pairs; Table 6). Same architecture agrees more than same family here, while for the per-token profile the order is reversed (Section B.1); the two sets of classes are built from different pair lists, so only their order is compared. Within one model (*leave-one-document-out*), one document agrees with the average of the other 169 at 0.9529 on the full vector over heads, without depth binning.

Two related statistics. The per-head write in Table 6 is approximate: the stored data hold one value norm per layer, so every head of a layer is given the layer's mean value norm. Two further columns check this approximation from either side. The write per layer is exact but coarser, since it sums the heads of each layer; the output norm, the norm of each head's output, is exact per head but measures a different quantity. All three columns show a profile that is stable inside a model and much lower between architectures. The output norm is missing for one model, so its model and pair counts, given in parentheses, are smaller.

Table 6: **Every head-wise statistic is stable inside one model and much lower between models of different architecture.** Rows are the comparisons of Figure 3; entries are the mean Spearman correlation between per-head profiles over n models or model pairs (in parentheses where the output norm has fewer), on 170 documents. Setting in Section C.

comparison	n	write per head	write per layer	output norm
document split halves	67	0.992	0.992	0.992 (66)
leave-one-document-out	67	0.953	0.978	0.934 (66)
base vs derived model	17	0.990	0.990	0.987
same family	55	0.783	0.783	0.884
same architecture, different family	108	0.820	0.820	0.841
different architecture	2031	0.373	0.373	0.436 (1965)
permuted heads	2211	0.000	-0.005	0.001 (2145)

D WHY MARGINAL ATTENTION APPROXIMATES THE JACOBIAN

D.1 A SINGLE ATTENTION LAYER

We now motivate this approximation heuristically. For a single causal attention layer, ignore the dependence of the attention weights on the input. Head h adds $\sum_{s\le t} A^{(h)}_{s,t}\, W^{(h)}_{OV}\mathbf{x}_s$ to the residual stream at position t, with $W^{(h)}_{OV} := W^{(h)}_O W^{(h)}_V$, so its part of the Jacobian block is $[J^{(s)}_h]_t = A^{(h)}_{s,t} W^{(h)}_{OV}$, and the block is $[J^{(s)}]_t = \sum_h [J^{(s)}_h]_t$ up to the identity at $s = t$. Applied to the input it acts on, the head's part has norm

$$\left\|[J^{(s)}_h]_t\,\mathbf{x}_s\right\| = A^{(h)}_{s,t}\,\|W^{(h)}_O v_s\| = \omega^{(h)}_{s,t}\,A^{(h)}_{s,t},$$

which is exactly the summand of Definition 1. Summing over output positions and heads and normalising, Definitions 2 and 4 become

$$a_s \propto \sum_{t\ge s}\sum_h \left\|[J^{(s)}_h]_t\,\mathbf{x}_s\right\|, \qquad j_s \propto \sum_{t\ge s}\left\|\sum_h [J^{(s)}_h]_t\right\|_F,$$

with both normalisers summing the same expression over s. Per-token marginal attention and per-token marginal influence differ in two places. First, a sums the heads' norms where j takes the norm of their sum, so they agree when the heads' contributions do not systematically cancel, for example when the heads write into linearly independent subspaces. Second, a measures each head's part along the one direction $\mathbf{x}_s$ where j takes its Frobenius norm over all directions; the two are

proportional, with a factor that does not depend on s, when $\mathbf{x}_s$ is a generic direction for $W_{OV}^{(h)}$ and $\|\mathbf{x}_s\|$ is roughly constant, as it is after the normalisation that precedes attention. With $\omega \equiv 1$, the first-order term instead drops each head's gain $\|W_{OV}^{(h)}\mathbf{x}_s\|$ and keeps only the attention. This perhaps explains the lower correlation between the two quantities.

D.2 Several layers

Section D.1 argues for a single attention layer. To understand why this might continue to hold across multiple layers, write the input-output Jacobian as

$$J = \prod_{\ell=1}^{L}(I + J_\ell) = I + \sum_{\ell=1}^{L} J_\ell + (\text{higher-order terms})$$

where J_ℓ denotes the nontrivial Jacobian contribution of layer ℓ. This expansion can be interpreted as a sum over paths through token space: the first-order terms are direct paths from s to t consisting of a single attention edge, and thus inherit the structure of the attention matrices, while terms such as $J_k J_\ell$ correspond to two-hop paths $s \to u \to t$, weighted by attention coefficients, with successively higher-order terms corresponding to progressively longer paths. Under the assumption that these multi-hop paths distribute attention somewhat evenly, this leads to

$$J = I + \sum_{\ell=1}^{L}\sum_{h} A^{(\ell,h)} \otimes W_{OV}^{(\ell,h)} + (\text{multi-hop terms})$$

with the first-order argument of Section 4.1 applying to each layer's heads, and the per-token marginal attention summing their terms over all layers.

Concretely, for the two-layer case $s \to u \to t$, consider the distribution over (s, u) induced by the first layer's attention pattern, $\mathbf{P}(s, u) \propto A_{s,u}^{(1)}$. Then, "attention distributed evenly" means $\mathbb{E}_u[a_u^{(2)} \mid s] = 1$ a.s., where $a_u^{(2)} := \sum_t A_{u,t}^{(2)}$ is the total attention paid to u in the second layer, n times its marginal attention with $\omega \equiv 1$. The condition is impossible under causal masking. However, violations are plausibly monotonic in position due to causal masking, perhaps explaining why higher-order terms disagree little with one another.

Heuristically, the multi-hop gain is also small. In general, only $\|W_{OV}^{(j)}W_{OV}^{(k)}\|_F \le \|W_{OV}^{(j)}\|_F\|W_{OV}^{(k)}\|_{\text{op}}$ holds. Suppose instead that the column space of $W_{OV}^{(k)}$ is randomly oriented relative to the row space of $W_{OV}^{(j)}$, that is, $W_{OV}^{(k)}$ is replaced by $QW_{OV}^{(k)}$ with Q a uniformly random rotation of $\mathbb{R}^{d_{\text{model}}}$. Then $\mathbb{E}_Q\|W_{OV}^{(j)}QW_{OV}^{(k)}\|_F^2 = \|W_{OV}^{(j)}\|_F^2\|W_{OV}^{(k)}\|_F^2/d_{\text{model}}$, and since $W_{OV}^{(k)}$ has rank at most d_{head}, this is at most $(d_{\text{head}}/d_{\text{model}})\|W_{OV}^{(j)}\|_F^2\|W_{OV}^{(k)}\|_{\text{op}}^2$. Under this approximation, the two-hop gain is smaller than the general bound by a factor of about $\sqrt{d_{\text{head}}/d_{\text{model}}}$, which plausibly further reduces the weight of the multi-hop terms in the Jacobian norm. Trained heads that compose are not randomly oriented, so this is not a bound.

E Theory: measurements

Table 7 compares, over 16 models, how well the per-token marginal influence and the per-token marginal attention of each model agree with the consensus of the other models.

E.1 Estimators of the marginal influence

Two estimators of the per-token marginal influence appear in Section 4. The **block norm** is the object of Definition 4: the Frobenius norm of each block of the Jacobian of the output logits with respect to the input embeddings, summed over output positions. We compute it only for Qwen2.5-0.5B, with Hutchinson's estimator on a sample of output positions in each document. Its agreement with the per-token marginal attention is moderate (0.64), and lower with uniform weights $\omega \equiv 1$ (0.53).

The **row gradient** replaces the block norm where it is too costly to compute for every model. The influence of input position s on output position t is the norm of the gradient of the log-probability

Table 7: **Marginal influence is shared across models, marginal attention more so.** 16 models from 6 families, 100 documents. Spearman correlation per document of one model's map with the consensus of the other models, with its own family left out of the consensus, and between two models, each beside a different-document reference. Rows: the unweighted per-token marginal attention and three row-gradient estimates of the per-token marginal influence (Section E.1): gradient length, substitution sensitivity, gradient times input.

	vs all other models		vs other families		
map	agreement	different document	agreement	different document	two models
unweighted profile	0.8612	0.0334	0.8076	0.0321	0.7652
influence, gradient length	0.7722	0.0336	0.6768	0.0316	0.6295
influence, substitution	0.7759	0.0328	0.6828	0.0308	0.6349
influence, gradient $\times$ input	0.6568	0.0344	0.5797	0.0318	0.4771

of the observed next token with respect to the embedding at s, which is one row of the block $[J^{(s)}]_t$. Each row is normalised to sum to one over input positions, and the rows are averaged over a sample of output positions. It is a proxy for the block norm, not the quantity of Definition 4; Table 7 also reports two variants of it. Most of the 16 models belong to two families, pythia and Qwen2.5, so this comparison covers fewer families than the census.

E.2 Assumption constants

We estimate the constants of the three assumptions and of the logit gap on the tokenizer-sharing pairs and the transformers of Section 4.2, with the same documents. Smoothness and the current-token constant are computed on the Jacobian blocks themselves; the lag profile, which is too costly as a block for every model, uses the row gradient.

Logit gap. The logit gap δ of a pair is the norm of the difference between the two models' centred next-token logits. As it grows, the agreement of the two models' per-token marginal influence falls, and it orders the 26 pairs tightly (Spearman -0.88). Pooled over tokenizer groups it does not order the agreement of the per-token marginal attention, which falls with the gap only inside each group. The theorem asks for a bound at every position, and the largest gap in a document is less than twice its typical gap, so the average gap is a fair summary of the bound.

Smoothness. The constant M of Assumption 1 bounds how fast a Jacobian block changes when one input embedding moves. We move one embedding by a small step in a random direction and divide the relative change of the block by the relative size of the step. This ratio is of order one for every model and changes little with the step for most of them, so the block changes in proportion to the step, as the assumption requires. It is lower for larger models (Spearman -0.62 with parameter count), so larger models are smoother. For the base and instruct versions of Qwen2.5-0.5B, the bound of Lemma 1 at these constants is about four times the typical block norm, so the theorem explains the direction of the effect rather than its size.

Current-token dependence. The constant c of Assumption 2 bounds from below the Jacobian block that links each prediction to the last token it reads, the token at its own position. The block is above zero at every sampled position of every model, and it holds a sizeable share of the block norm of its row (median 0.22 to 0.48 across models). The same holds for the row gradient of the models without attention (Section E.4). The constant is in logit units and differs across models by orders of magnitude, but the theorem only needs each model's own c to be above zero.

Lag decay. Assumption 3 asks that the dependence of a prediction on earlier tokens fall toward zero with their distance. This dependence decays as a power law of the distance, with an exponent of about one half to two thirds, and the models without attention decay in the same way. Larger models decay more slowly. The decay is slow: between a fifth and a third of the dependence lies more than 512 tokens back, so the envelope ψ of the theorem falls slowly and the bound is loose for long documents.

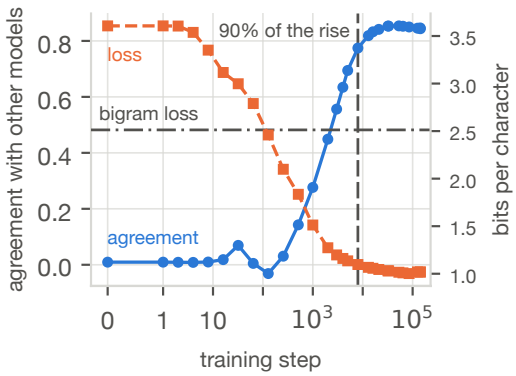

Figure 9: **Agreement with other models rises as the loss falls.** Pythia-410M over training: agreement of its per-token marginal attention with four trained models (left axis) and loss in bits per character (right axis); dash-dotted line: a bigram model's loss. Setting in Section E.3.

E.3 Emergence during training

Experimental setting. We follow one model through training to ask whether its agreement with other models grows as its language-modelling loss falls. The subject is Pythia-410M at 28 public training checkpoints, from random initialisation (step 0) to the final model (step 143,000), on 60 Pile passages cut at 512 tokens (median scored length 2192 characters). At each checkpoint we compute the unweighted ($\omega \equiv 1$) per-token marginal attention and score it with the estimator of Section 3.1: projection onto the characters, ranks, residualisation on 50 relative-position bins, and Spearman correlation per document. The reference is the mean of the same profile over four fully trained models of other families (Qwen2.5-0.5B, Qwen2.5-1.5B, SmolLM2-360M and OLMo-2-1B), so Pythia is never part of its own target. Agreement is this correlation averaged over documents, and the loss is the checkpoint's bits per character on the same passages, from the same forward pass. As a loss reference, we fit a bigram model with absolute discounting, backed off to a unigram model, on the other 440 passages of the corpus, and score it on the same 60 passages with the same tokenizer. We ran no bootstrap, so we report no interval.

Result analysis. Agreement is near zero at random initialisation and high at the final step (0.84), so the model acquires the shared profile during training and does not have it at the start. Most of the rise is early: it is nearly complete by step 8,000, while the loss is still falling. Agreement stays near zero as long as the loss is above the bigram model's 2.51 bits per character, and rises only after the loss drops below it, so the model shares the profile only once it predicts better than token-pair statistics. The evidence is narrow: one model, one training seed, and 60 documents of at most 512 tokens. The loss also stands in for the logit gap δ of Theorem 1, which we do not compute between this model and the four reference models.

E.4 Models without attention

Experimental setting. We ask whether the shared per-token marginal attention comes from the attention mechanism or from next-token prediction, by computing the per-token marginal influence of models that have no attention. The subjects are four Mamba models (130M to 1.4B parameters; Gu & Dao, 2024) and two RWKV-4 models (430M and 1.5B; Peng et al., 2023), with Qwen2.5-0.5B as the one transformer reference. The influence is the object of Definition 4, estimated with Hutchinson's method (8 probes) on 128 sampled output positions per document, all at position 64 or later, at 2048 tokens; RWKV-4 reads at most 1024 tokens, so its rows use the first 1024. The reference is the consensus of the unweighted ($\omega \equiv 1$) per-token marginal attention of the census models on the same document, and each influence map is compared with it using the estimator of Section 3.1; Qwen2.5-0.5B's own attention is part of that consensus. We use 100 documents of the census corpus, five per domain, the three generated domains included, and Figure 10 shows bootstrap intervals over documents.

Result analysis. The influence of the four Mamba models agrees with the attention consensus nearly as well as the transformer's does (0.42 to 0.47, and 0.50 for Qwen2.5-0.5B), so the shared profile

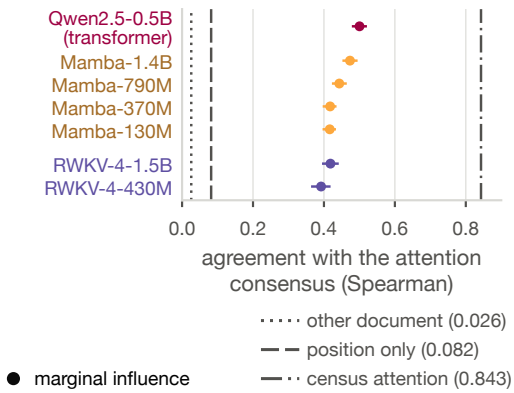

Figure 10: **Per-token marginal influence of models without attention.** Spearman correlation with the consensus per-token marginal attention of the census; one model per row; dashed: references. Setting in Section E.4.

does not require the attention mechanism. Every model stays far above a map of position alone and above its own influence map compared with a different document. The evidence is narrow: one transformer reference, documents that include generated text, and no logit gap δ for any model without attention, so this is not a test of Theorem 1.

E.5 What the shared per-token profile encodes

Experimental setting. To ask what the shared per-token marginal attention encodes, we predict it from properties of the text alone. The target is the consensus per-token profile of the census, unweighted ($\omega \equiv 1$), per word, ranked and residualised on 50 relative-position bins as in Section 3.1. The predictors are five groups of word properties: position in the document; surface form (length, punctuation, capitalisation, sentence edges); word frequency (external and in-corpus counts); local predictability (surprisal and entropy at the word and the next token); and grammatical role (parse depth, number of dependents, part of speech, dependency label). We fit a ridge regression on each group alone and on all five together, out of fold by document (5 folds), and score it by Spearman correlation per document over 550 documents of 11 English-prose domains. The reference is one model's own per-token profile compared with the consensus of the other 66, averaged over the 67 census models.

Result analysis. The five groups together predict the consensus profile only in part (0.52), well below a model's own unweighted per-token profile (0.88). Alone, grammatical role predicts it best, then surface form, local predictability and word frequency, and position predicts nothing (right panel of Figure 4). Removing one group from the five costs the most for grammatical role and local predictability, little for surface form and word frequency, and nothing for position.

F KV eviction details

F.1 Head budget details

The set $\mathcal{D}_{\text{pre}}$ of Equation (2) holds 200 documents from the Nemotron pretraining sample (NVIDIA, 2025), the calibration corpus of KVzap (Jegou & Jeblick, 2026), and is disjoint from every benchmark. KVzip (Kim et al., 2025) scores every entry (h, s) of a document by the largest value-weighted attention it receives from the queries $\mathcal{R}(X)$ of its reconstruction pass, normalised by the attention input h_t, and keeps the fraction B of all entries above one global cut $\tau_B(X)$. The fraction $m_h(X)$

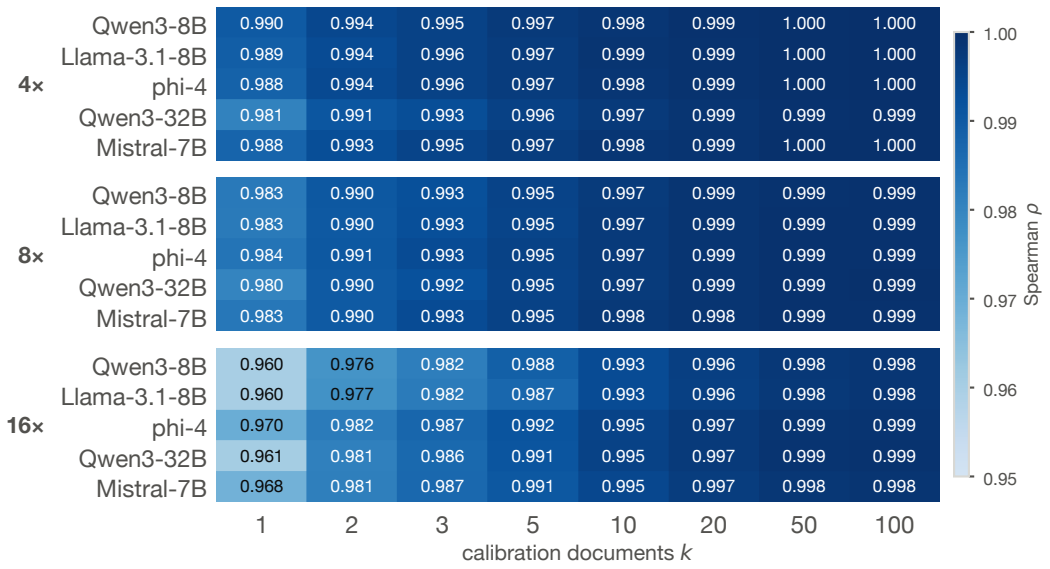

Figure 11: **Agreement of Head Mass tables from k and $200 - k$ documents.** One panel per compression ratio, one row per target, one column per k. Each cell is the median, over 100 random draws of k of the 200 pretraining documents, of the Spearman correlation over the target's key-value heads between the per-head marginal attention of the drawn documents and of the rest. At $k = 1$ the median is 0.980 to 0.984 at 8×.

behind the share $p_h(X)$ of Equation (2) counts head h's own entries above that cut:

$$m^{\mathrm{rec}}_{h,s}(X) = \max_{t \in \mathcal{R}(X)} \frac{A^{(h)}_{s,t}(X)\,\|W^{(h)}_O v_s(X)\|}{\|h_t\|}, \qquad \frac{1}{Hn}\sum_{j=1}^{H}\sum_{s=1}^{n} \mathbf{1}\big[m^{\mathrm{rec}}_{j,s}(X) > \tau_B(X)\big] = B,$$
$$m_h(X) = \frac{1}{n}\sum_{s=1}^{n} \mathbf{1}\big[m^{\mathrm{rec}}_{h,s}(X) > \tau_B(X)\big]. \tag{3}$$

Table 8: **Head Mass tables from disjoint sets of pretraining documents agree at a median Spearman of 0.98 or higher.** Spearman correlation over each target's key-value heads between the per-head marginal attention (Equation (2)) of two disjoint sets of the 200 pretraining documents, read on KVzip's reconstruction queries at 8×. *1 vs 199* is the median (5th percentile) over documents, *5 vs 195* and *20 vs 180* the median of 100 draws, and *100 vs 100* the even documents compared with the odd ones. The *top decile* column is the Jaccard overlap of the top tenth of heads between even and odd documents, and the *prefill queries* column is the 1 vs 199 median on the document's own queries.

target	key-value heads	1 vs 199	5 vs 195	20 vs 180	100 vs 100	top decile	prefill queries
Qwen3-8B	288	0.9828 (0.9704)	0.9952	0.9985	0.9995	1.0000	0.9624
Llama-3.1-8B	256	0.9830 (0.9744)	0.9951	0.9985	0.9994	0.9259	0.9394
phi-4	400	0.9836 (0.9706)	0.9954	0.9986	0.9992	0.9512	0.9681
Qwen3-32B	512	0.9815 (0.9512)	0.9952	0.9986	0.9993	1.0000	0.9656
Mistral-7B	256	0.9829 (0.9729)	0.9954	0.9984	0.9994	1.0000	0.9666

F.2 Head budget maps, one per target

One figure per target, at 8×: the Head Mass table, LU-KV's calibrated budget, KVzip's realised per-head shares, the proportional budget (Equation (4)) and ordinary prefill attention, head by head. The Head Mass table agrees closely, head for head, with the keep rates KVzip realises on held-out documents (Spearman 0.91 to 0.93, Table 9), so averaging over pretraining text loses little. The same recipe applied to the reading queries, the document's own positions, agrees with it far less, as the maps show.

Table 9: **LU-KV, KVzip (realised) and Proportional agree with Head Mass at Spearman 0.80 to 0.98; prefill attention agrees at 0.13 to 0.42.** Spearman correlation with the Head Mass table over each target's (layer, key-value head) grid at 8×, for the tables of Appendix F.2. The last row is the mean over targets.

target	LU-KV	KVzip (realised)	Proportional	Prefill attention	key-value heads
Qwen3-8B	0.81	0.92	0.96	0.13	288
Llama-3.1-8B	0.80	0.91	0.95	0.38	256
phi-4	0.81	0.92	0.97	0.30	400
Qwen3-32B	0.80	0.91	0.98	0.17	512
Mistral-7B	0.81	0.93	0.96	0.42	256
mean over targets	0.81	0.92	0.96	0.28	–

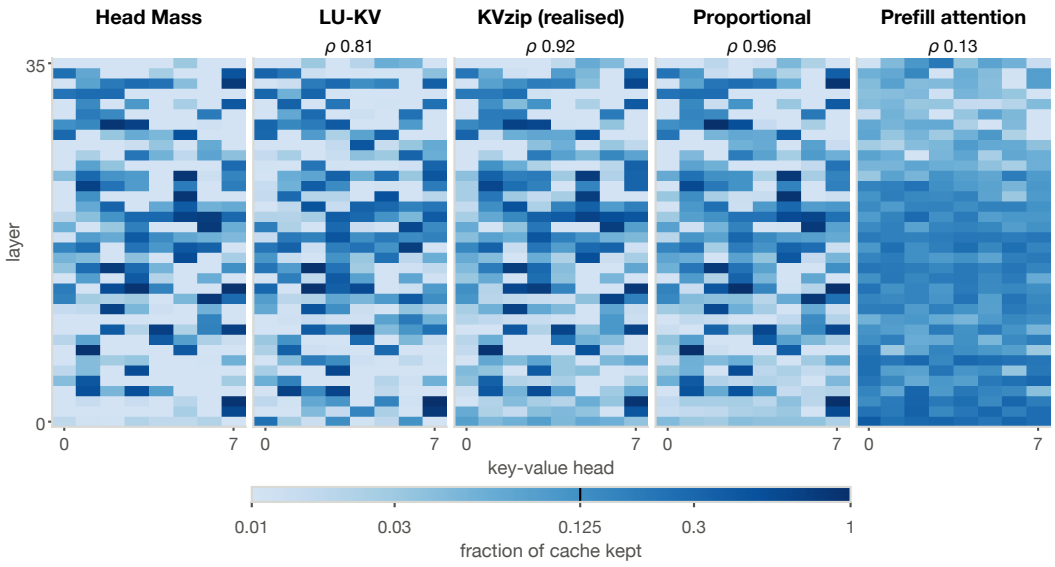

Figure 12: **Per-head budgets at 8×, Qwen3-8B.** One cell per key-value head, layer 0 at the bottom; colour is the fraction of the head's cache kept (log scale, uniform 0.125). Head Mass averages 200 pretraining documents; LU-KV is its budget curve at compression 0.88, hence a mean of 0.12; KVzip is the realised keep rate of its global cut on 12 held-out documents, with no floor (values under 0.01 drawn at 0.01); Proportional is Equation (4); Prefill attention applies the Head Mass recipe to the document's own prefill queries. ρ is the Spearman correlation with Head Mass.

F.3 Proportional budget ablation

The ablation replaces Head Mass's threshold share with a share proportional to each head's summed KVzip score on the reconstruction pass,

$$\beta_h \;\propto\; \mathbb{E}_{X\sim\mathcal{D}_{\text{pre}}}\Big[\sum_s \max_t A^{(h)}_{s,t}(X)\,\|W_O^{(h)}v_s\|/\|h_t\|\Big], \tag{4}$$

max-pooled over each key-value group, then clipped and rescaled as in Equation (2). With EA as the scorer it stays within half a point of HM + EA on Qwen3-8B at every setting, trails it by about one to three points on the other four targets at most compression ratios, and trails it by more than ten points on Qwen3-32B at 16× (Table 11). On Qwen3-8B the cut costs under half a point, while reading the same budget on the reading queries instead of the reconstruction pass costs over 30 points at every setting (Table 12).

F.4 The guided score

The guided score keeps EA and raises the tokens that a small guide model, Qwen2.5-0.5B-Instruct (0.5B parameters), marks as important. The guide reads the context the target compresses, without the question, once on the reading queries and once in a reconstruction pass. For each guide head it

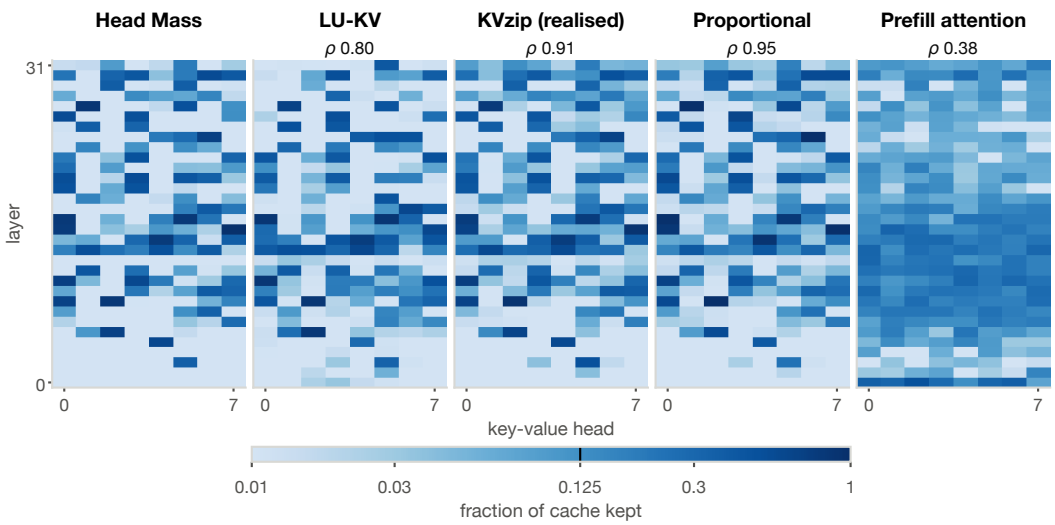

Figure 13: **Per-head budgets at 8×, Llama-3.1-8B.** Layout as in Figure 12.

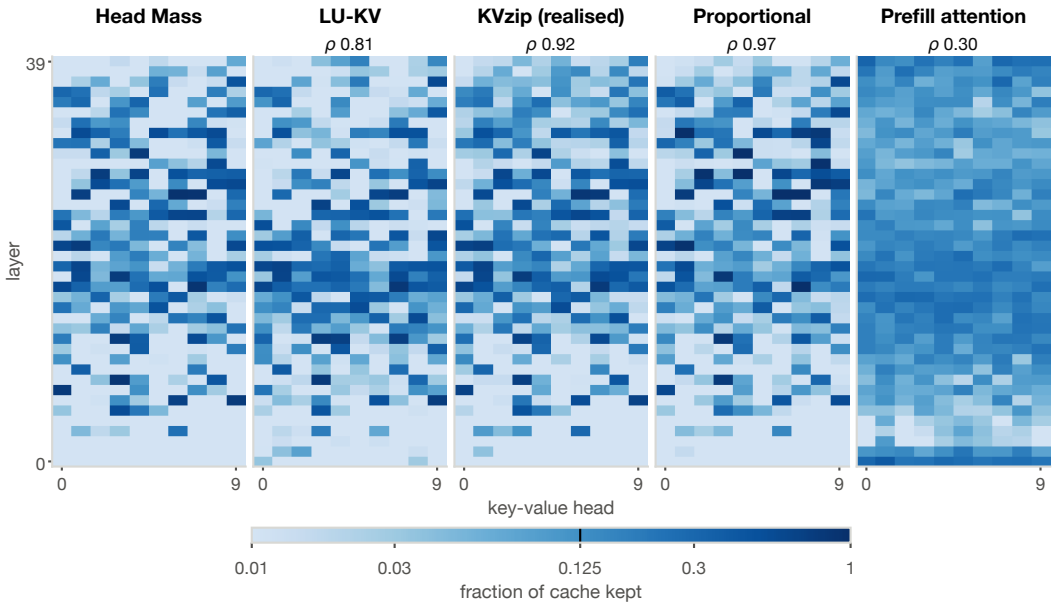

Figure 14: **Per-head budgets at 8×, phi-4.** Layout as in Figure 12.

takes two profiles over positions: the attention each token receives, summed over the reading queries without value weights, and the largest attention it receives from the reconstruction queries. Note that neither profile is strictly the base Definition 1: the reading profile is the marginal attention with $\omega_{s,t}^{(h)} \equiv 1$, without the value weight (the factor $1/n$ does not change ranks), and the reconstruction profile takes the maximum over queries instead of the mean. Each profile is carried to the target's tokens through the characters they cover and ranked within the document. A linear map, fitted once per target by ridge regression on 48 documents from the same Nemotron sample as the budget (Appendix F.1), turns these profiles into a prediction $\hat{g}_{h,s}(X)$ of KVzip's log score for every key-value head h of the target. The prediction tilts EA head by head:

$$\gamma_{h,s}(X) = \gamma^{\mathrm{EA}}_{h,s}(X)\,\big(1 + \lambda\, g_{h,s}(X)\big), \qquad g_{h,s}(X) = \frac{\hat{g}_{h,s}(X) - \min_{s'} \hat{g}_{h,s'}(X)}{\max_{s'} \hat{g}_{h,s'}(X) - \min_{s'} \hat{g}_{h,s'}(X)}, \tag{5}$$

where $\gamma^{\mathrm{EA}}_{h,s}(X)$ is the EA score of entry (h, s), $g_{h,s}(X)$ rescales the prediction to $[0, 1]$ within each head, and $\lambda = 2$ sets the strength of the tilt; $\lambda = 0$ gives EA. The budget β_h is the Head Mass table and the tilt leaves it unchanged.

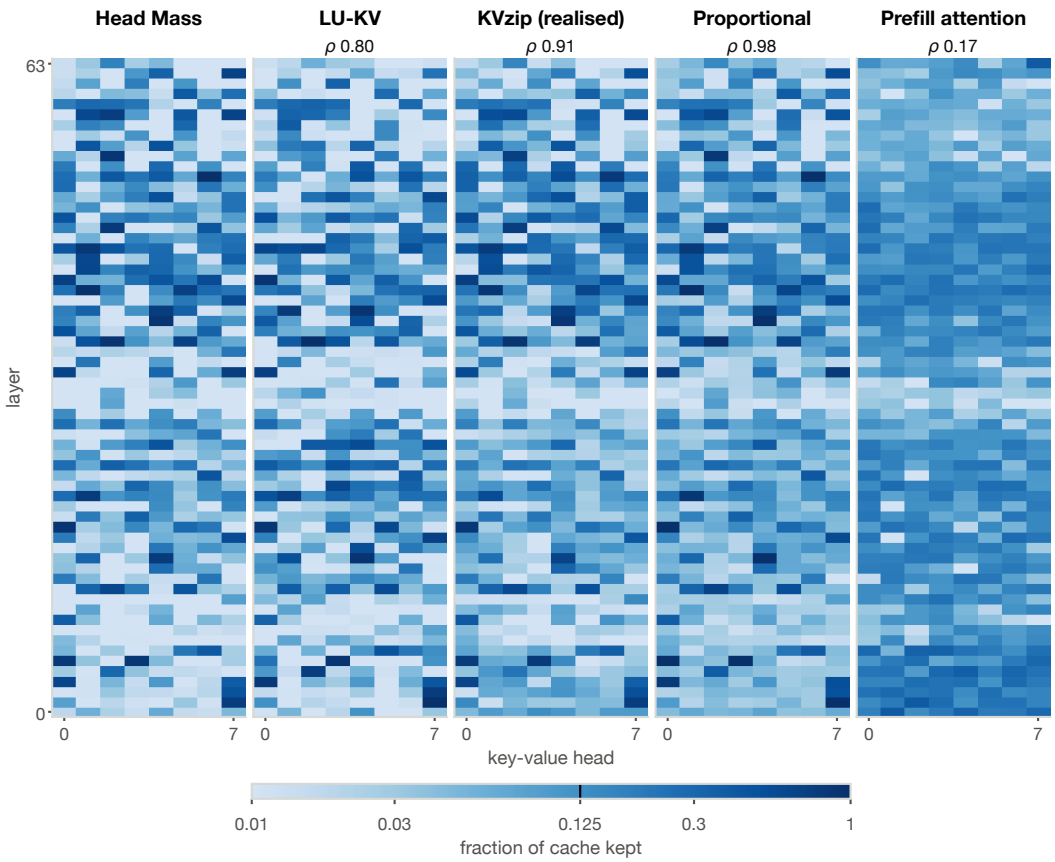

Figure 15: **Per-head budgets at 8×, Qwen3-32B.** Layout as in Figure 12.

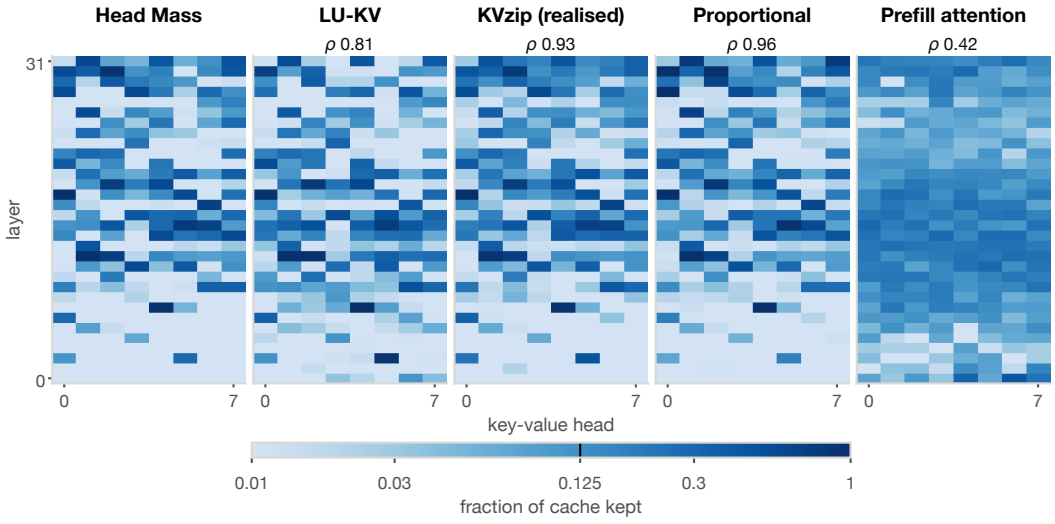

Figure 16: **Per-head budgets at 8×, Mistral-7B.** Layout as in Figure 12.

Table 10: **HM + EA needs one offline budget table per target and no work per document.** Work each method needs beyond ordinary serving. △ marks work done once per target and • work per document; two marks denote training a surrogate or an extra target forward. HM is Head Mass and EA is Expected Attention. The last column is the wall time of the offline step on Mistral-7B-Instruct-v0.3 (SLURM accounting).

		offline, per target	**online, per document**	**offline time, Mistral-7B**
AdaKV + EA		none	none	none
AdaKV + KVzap	△△	surrogate training on KVzip's score	surrogate call, not counted	4 h 14 m, one H100
KVzip	••	none	one extra target forward	none
LU-KV + EA	△	per-head budget curve calibration	none	published curve
HM + EA	△	per-head budget table, 200 pretraining documents	none	34 min
HM + Guide	△•	the same table and a guide map	one Qwen2.5-0.5B-Instruct forward	36 min

Table 11: **With EA as the scorer, the proportional per-head budget (Equation (4)) is within half a point of HM + EA at every setting only on Qwen3-8B; elsewhere it trails by up to 3.3 points, except on Qwen3-32B at 16×, where it trails by more than 11.** Each cell is the RULER score of the proportional budget with EA; in parentheses, its paired difference from HM + EA, bold where the 95% interval excludes zero. Both budgets come from KVzip's reconstruction pass on the same pretraining documents.

	RULER, context 4096			RULER, context 8192	
Target	4×	8×	16×	8×	16×
Qwen3-8B	94.8 (−**0.2**)	91.9 (0.2)	73.1 (−0.2)	89.1 (0.1)	69.5 (0.4)
Llama-3.1-8B	95.0 (−0.1)	90.8 (−**1.5**)	65.1 (−**1.2**)	86.3 (−**1.7**)	60.7 (−**2.1**)
phi-4	90.8 (−**0.6**)	84.8 (−**3.0**)	66.8 (−**1.7**)	82.8 (−**2.1**)	67.3 (−**1.1**)
Qwen3-32B	95.0 (−**0.2**)	92.0 (−**1.2**)	67.6 (−**11.5**)	90.9 (−**1.1**)	66.5 (−**11.4**)
Mistral-7B	89.3 (−**1.2**)	83.2 (−**1.1**)	63.2 (−**2.2**)	79.7 (−**0.7**)	57.8 (−**3.3**)

Table 12: **On Qwen3-8B, the proportional per-head budget (Equation (4)) read on the reconstruction queries is within 0.25 points of HM + EA; read on the document's own queries it scores below AdaKV + EA at every compression ratio.** Over heads, the Spearman correlation between the reading-pass budget and Head Mass is −0.11, −0.15 and −0.10 at 4×, 8× and 16×. The upper rows are RULER-4096 scores (mean over 13 tasks, 6500 documents); the lower rows are paired differences with 95% bootstrap intervals.

	4×	8×	16×
AdaKV + EA	88.4	72.8	46.5
HM + EA	95.0	91.8	73.3
HM + KVzap	94.9	92.0	73.0
Proportional (reconstruction) + EA	94.8	91.9	73.1
Proportional (reading) + EA	61.1	53.0	37.7
paired difference, with its 95% interval			
Proportional (reconstruction) − HM + EA	−0.21 [−0.40, −0.04]	0.15 [−0.22, 0.50]	−0.25 [−1.01, 0.47]
Proportional (reading) − HM + EA	−33.94 [−34.52, −33.31]	−38.72 [−39.43, −38.00]	−35.59 [−36.47, −34.68]
Proportional (reading) − AdaKV + EA	−27.37 [−28.07, −26.62]	−19.78 [−20.58, −19.06]	−8.73 [−9.64, −7.82]

Table 13: **At context 8192, HM + EA scores above AdaKV + EA in all 10 columns.** Mean string-match score over the 13 RULER tasks at context 8192, 6500 documents and one run per cell, for five targets at 8× and 16× compression. Paired 95% bootstrap half-widths of the differences from HM + Guide are 0.33 to 1.09 points. Methods, shading, cost symbols and italics as in Table 1; the uncompressed row is not ranked.

	cost		Qwen3-8B		Llama-3.1-8B		phi-4		Qwen3-32B		Mistral-7B	
	offline	online	8×	16×	8×	16×	8×	16×	8×	16×	8×	16×
AdaKV + EA			71.7	50.9	41.2	15.8	41.7	13.6	64.1	38.4	33.5	20.5
AdaKV + KVzap	△△		86.6	46.5	86.7	60.5	*70.8*	*43.9*	88.5	50.3	*67.0*	*40.2*
KVzip		••	**89.8**	55.6	**90.1**	**73.6**	**88.0**	64.2	**95.0**	**92.1**	**82.5**	**66.3**
LU-KV + EA	△		*72.9*	*55.7*	87.7	62.4	*86.5*	*68.8*	*73.3*	*38.6*	78.8	58.3
HM + EA	△		89.0	**69.1**	88.0	62.8	84.9	68.4	92.0	77.9	80.5	61.1
HM + Guide	△	•	89.1	67.9	89.9	68.3	87.7	**72.9**	93.1	78.7	80.2	60.1
Uncompressed			93.9		94.6		92.1		95.6		89.8	

Table 14: **On LongBench, LU-KV + EA and HM + EA differ by at most 2.5 points over eight columns.** Unweighted mean over the 16 English LongBench subsets at their full official split (3750 documents), each scored by its official metric, one run per cell, for four targets at 8× and 16× compression; phi-4 was not run. No interval is given, since the mean mixes metrics. Methods, shading, cost symbols and italics as in Table 1; the uncompressed row is not ranked.

	cost		Qwen3-8B		Llama-3.1-8B		Qwen3-32B		Mistral-7B	
	offline	online	8×	16×	8×	16×	8×	16×	8×	16×
AdaKV + EA			42.1	33.1	36.4	28.4	46.0	39.2	31.0	27.6
AdaKV + KVzap	△△		34.0	23.9	42.0	34.4	36.8	19.0	*28.5*	*20.0*
KVzip		••	44.7	28.0	44.5	33.8	**50.7**	**45.2**	**43.8**	38.2
LU-KV + EA	△		***46.3***	***41.1***	**45.8**	**43.0**	*48.0*	*43.4*	42.6	38.8
HM + EA	△		46.2	39.9	45.6	40.5	49.9	44.8	42.3	**39.5**
HM + Guide	△	•	46.3	38.0	44.9	37.1	49.8	44.3	42.3	36.4
Uncompressed			47.3		45.9		50.9		42.9	

F.5 PER-TASK BREAKDOWN

Each figure in this subsection shows one benchmark at one compression ratio on one target: RULER at context 4096 or 8192, or LongBench. Each group of bars is one task, or one subset for LongBench, and the bars in a group are the methods of Table 1, in the same order and under the same names: AdaKV + EA, AdaKV + KVzap, KVzip, LU-KV + EA, HM + EA and HM + Guide. The black line is the uncompressed model, so the distance from a bar down to the line is what a method loses on that task. The figures show which tasks make the differences between methods, which a single aggregate score hides. The aggregate scores in the main table (Table 1) and in Tables 13 and 14 are the means of these per-task scores.

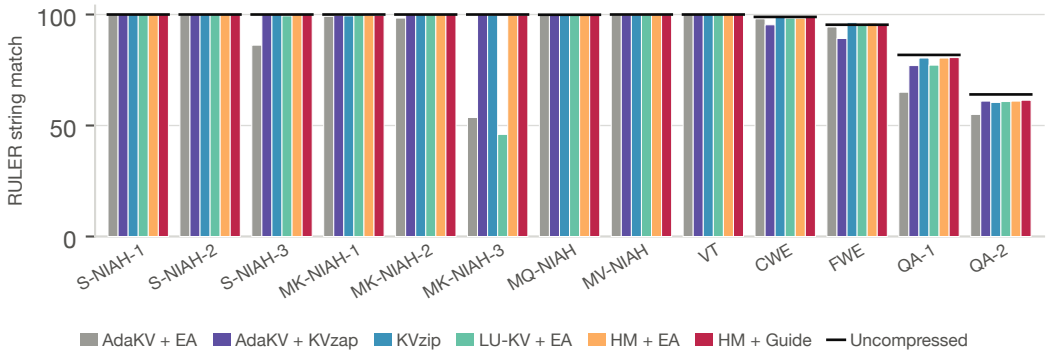

Figure 17: **Per-task scores, RULER-4096, 4×, Qwen3-8B.**

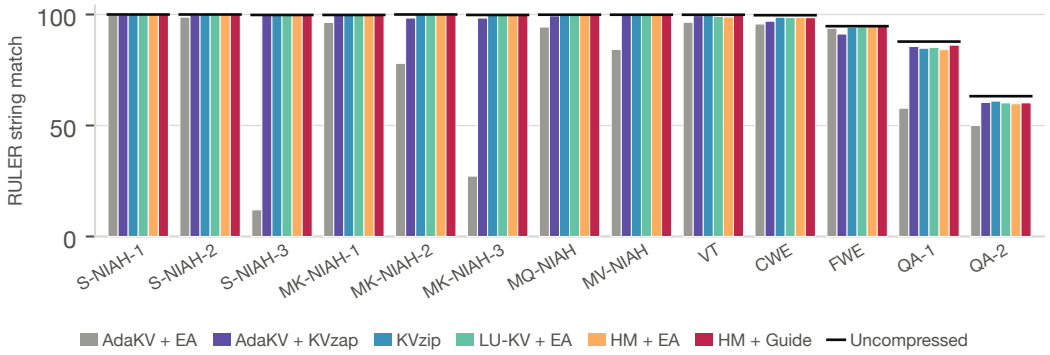

Figure 18: **Per-task scores, RULER-4096, 4×, Llama-3.1-8B.**

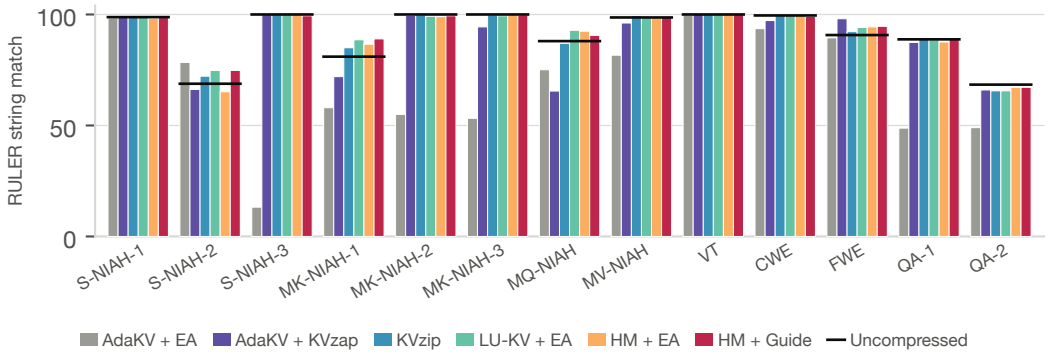

Figure 19: **Per-task scores, RULER-4096, 4×, phi-4.**

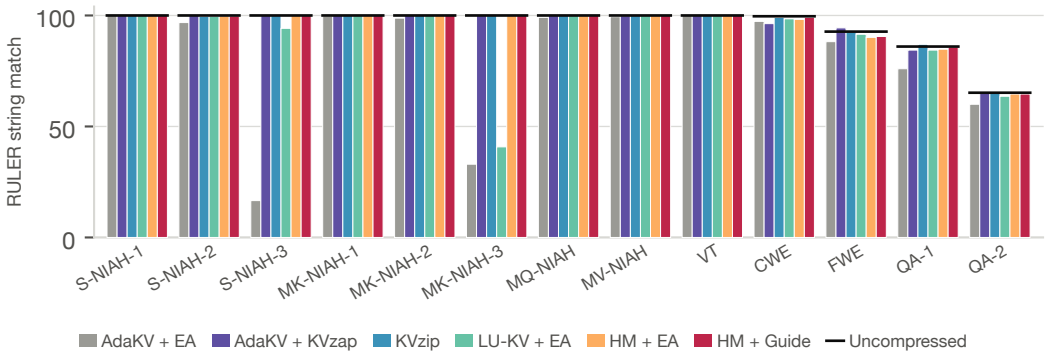

Figure 20: **Per-task scores, RULER-4096, 4×, Qwen3-32B.**

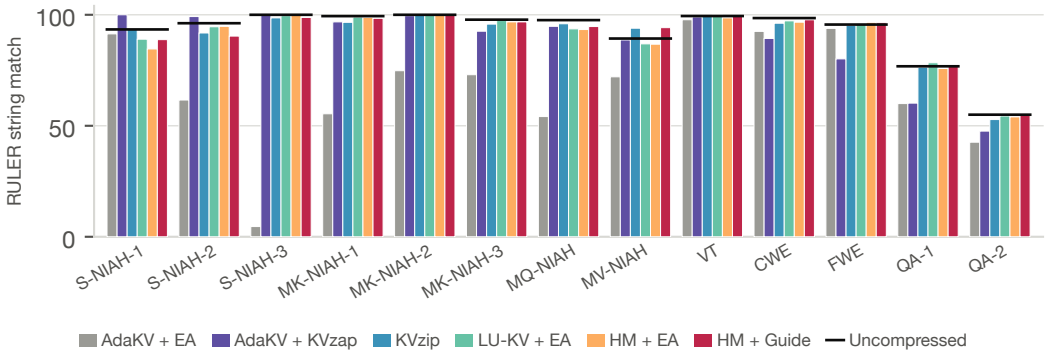

Figure 21: **Per-task scores, RULER-4096, 4×, Mistral-7B.**

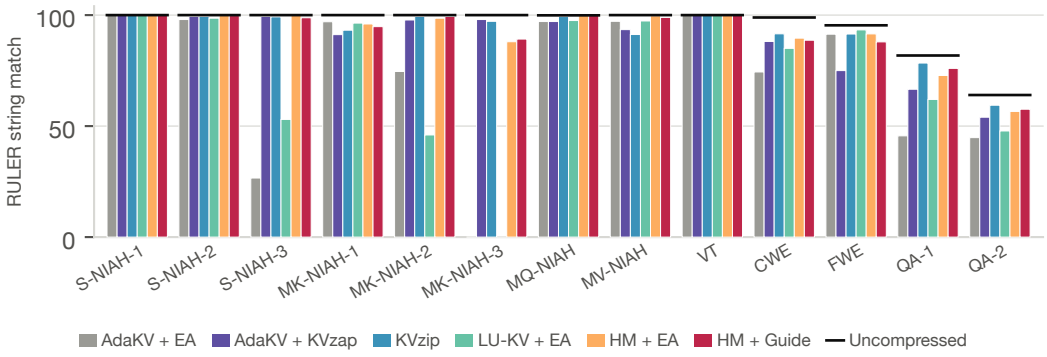

Figure 22: **Per-task scores, RULER-4096, 8×, Qwen3-8B.**

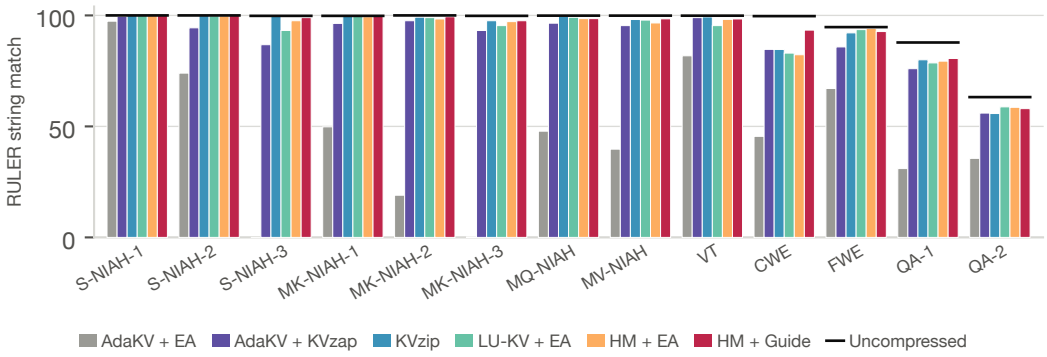

Figure 23: **Per-task scores, RULER-4096, 8×, Llama-3.1-8B.**

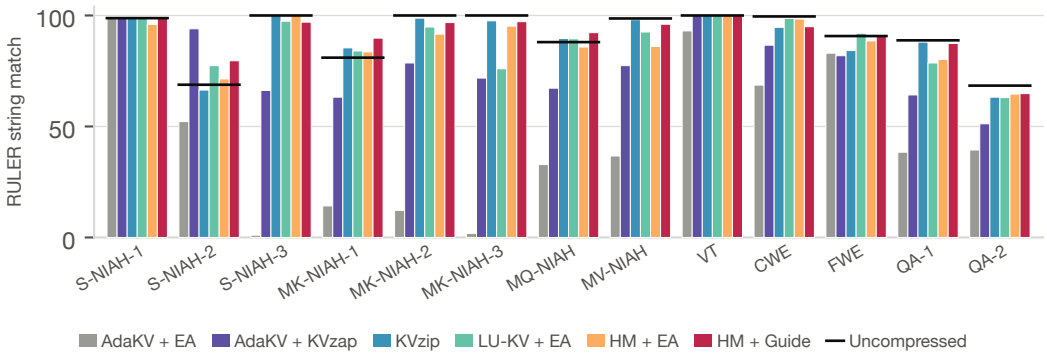

Figure 24: **Per-task scores, RULER-4096, 8×, phi-4.**

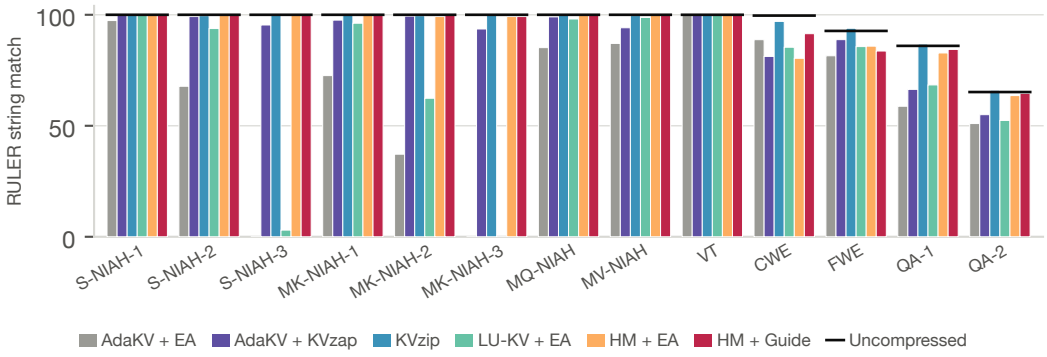

Figure 25: **Per-task scores, RULER-4096, 8×, Qwen3-32B.**

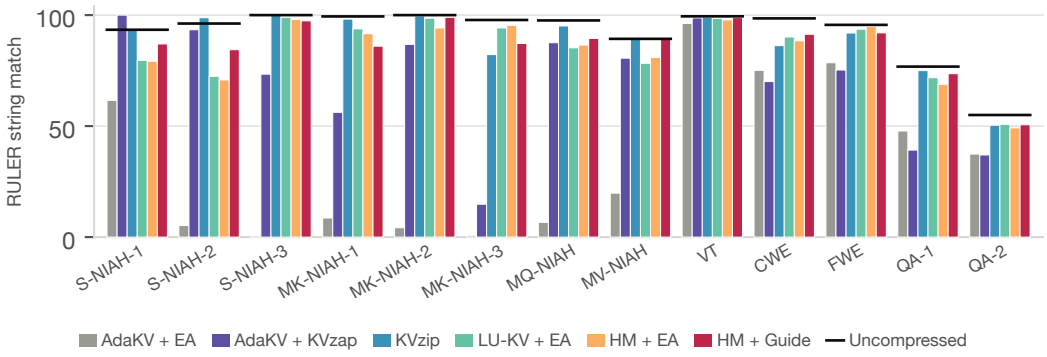

Figure 26: **Per-task scores, RULER-4096, 8×, Mistral-7B.**

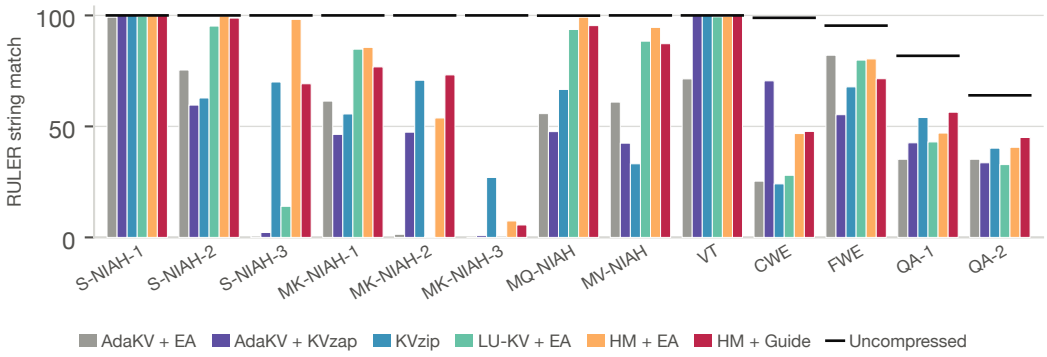

Figure 27: **Per-task scores, RULER-4096, 16×, Qwen3-8B.**

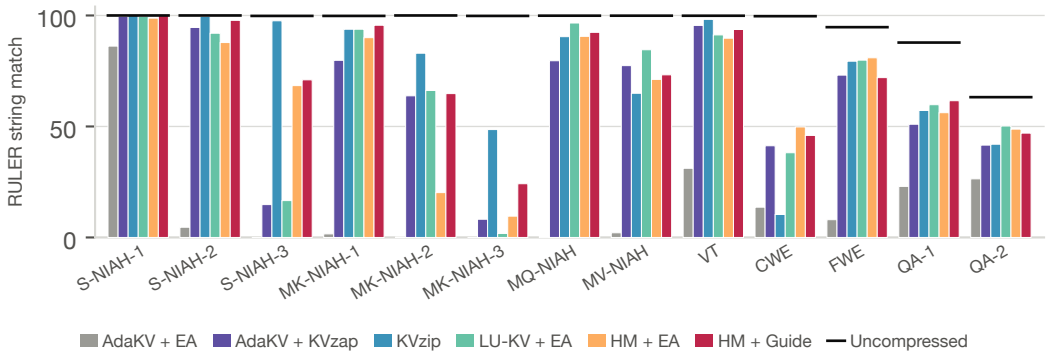

Figure 28: **Per-task scores, RULER-4096, 16×, Llama-3.1-8B.**

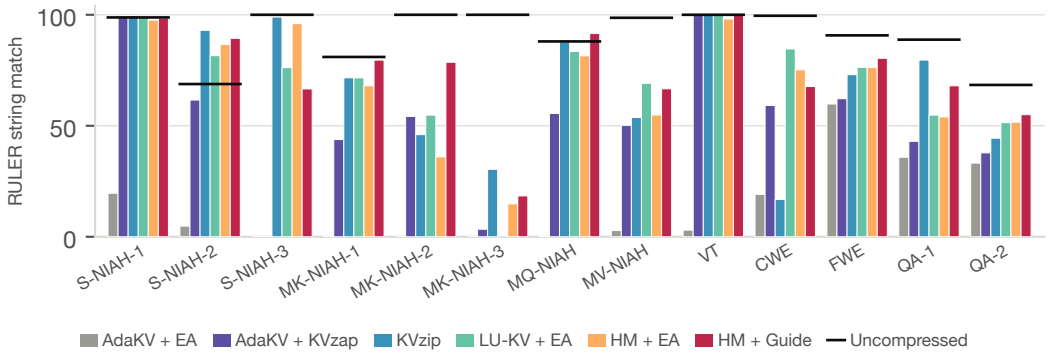

Figure 29: **Per-task scores, RULER-4096, 16×, phi-4.**

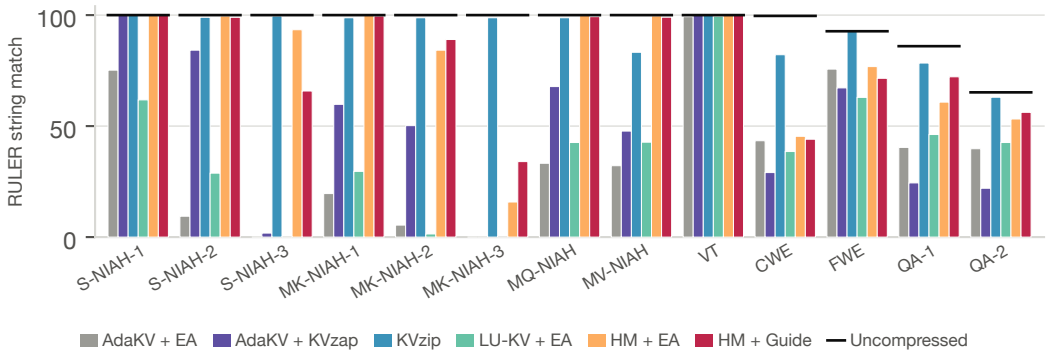

Figure 30: **Per-task scores, RULER-4096, 16×, Qwen3-32B.**

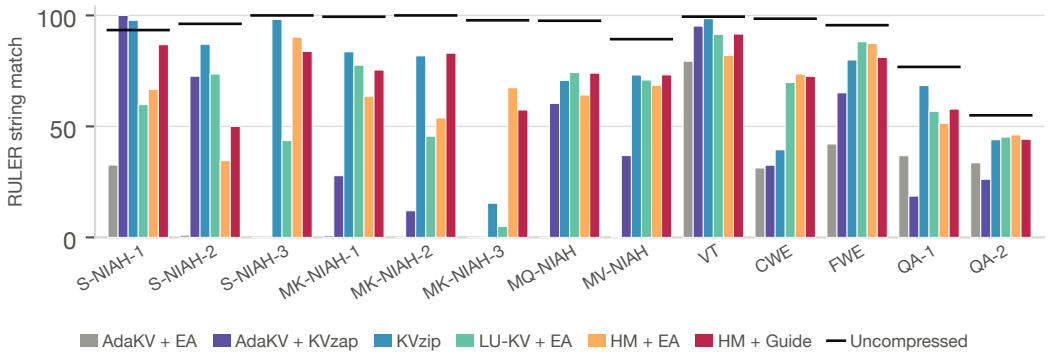

Figure 31: **Per-task scores, RULER-4096, 16×, Mistral-7B.**

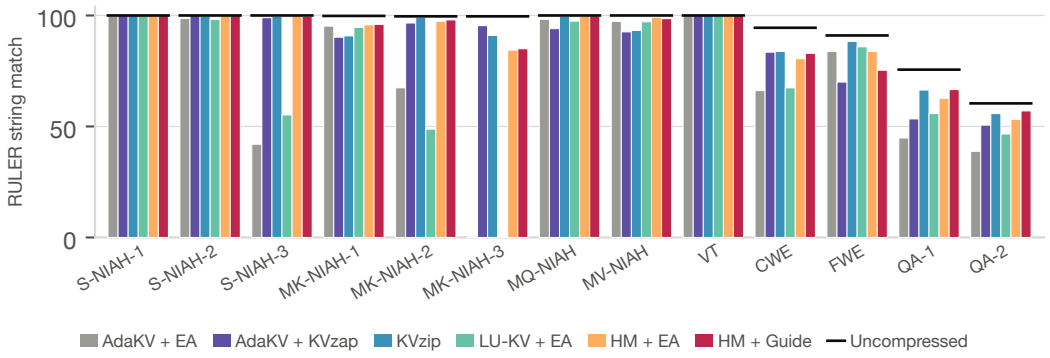

Figure 32: **Per-task scores, RULER-8192, 8×, Qwen3-8B.**

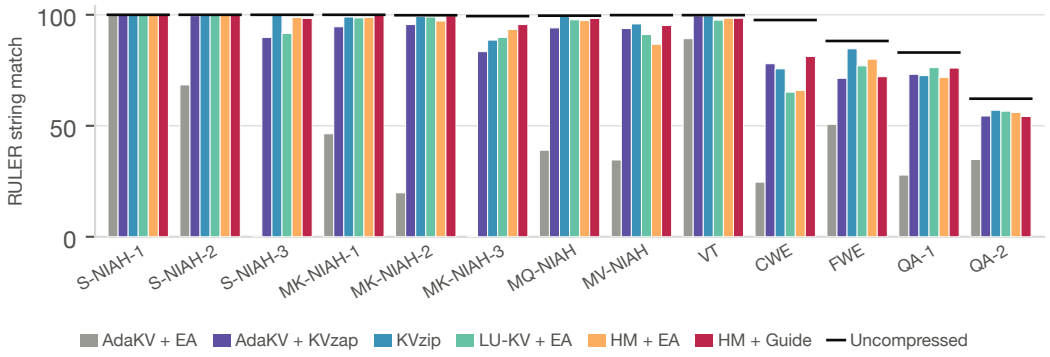

Figure 33: **Per-task scores, RULER-8192, 8×, Llama-3.1-8B.**

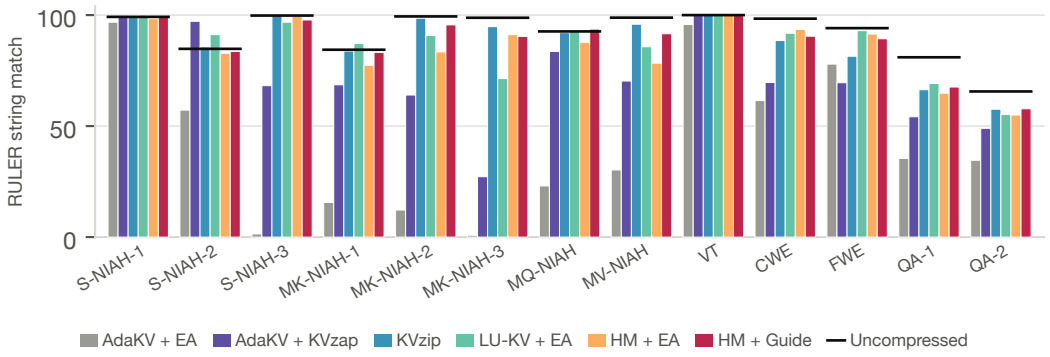

Figure 34: **Per-task scores, RULER-8192, 8×, phi-4.**

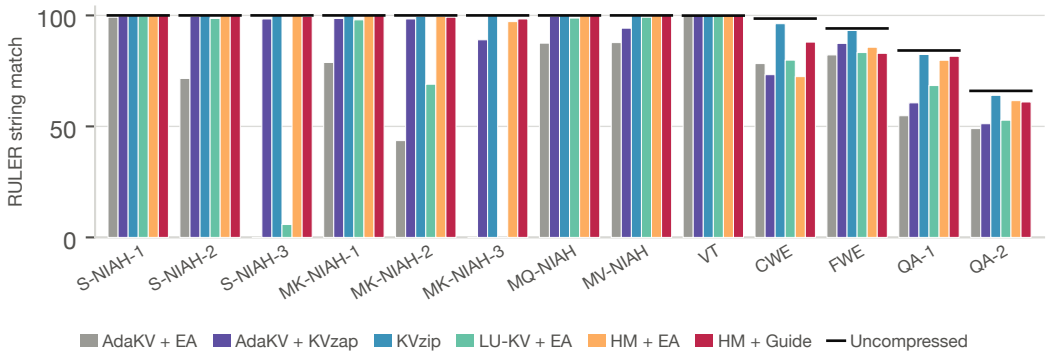

Figure 35: **Per-task scores, RULER-8192, 8×, Qwen3-32B.**

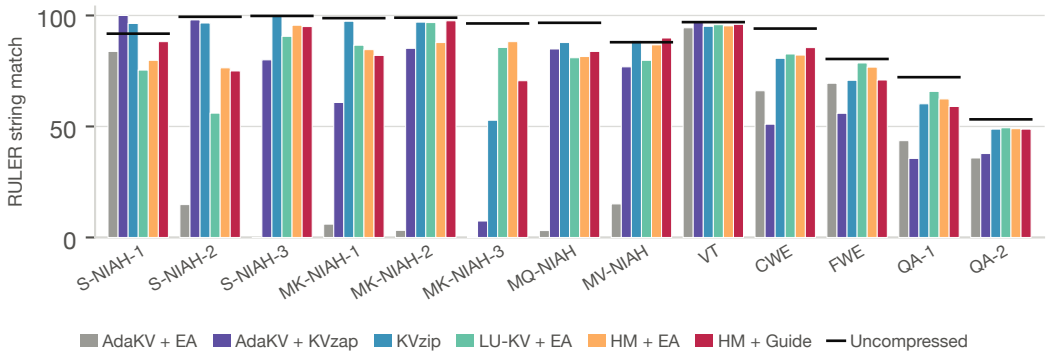

Figure 36: **Per-task scores, RULER-8192, 8×, Mistral-7B.**

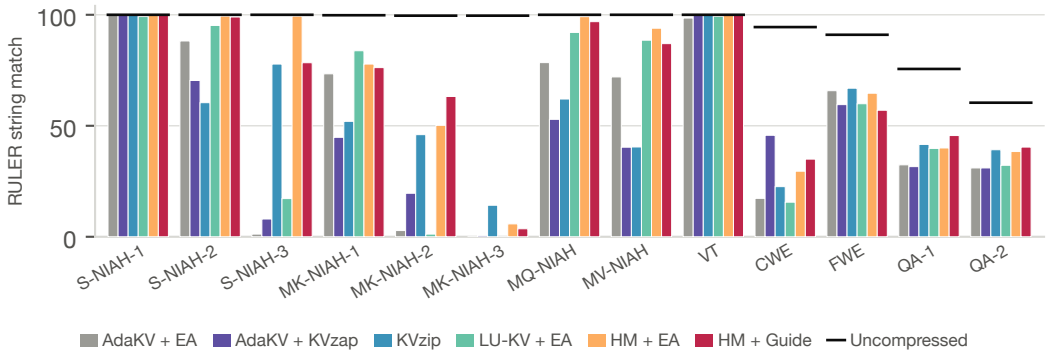

Figure 37: **Per-task scores, RULER-8192, 16×, Qwen3-8B.**

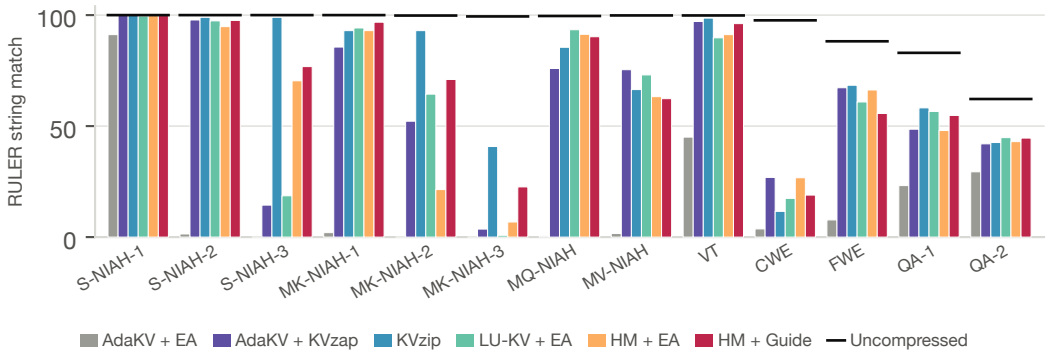

Figure 38: **Per-task scores, RULER-8192, 16×, Llama-3.1-8B.**

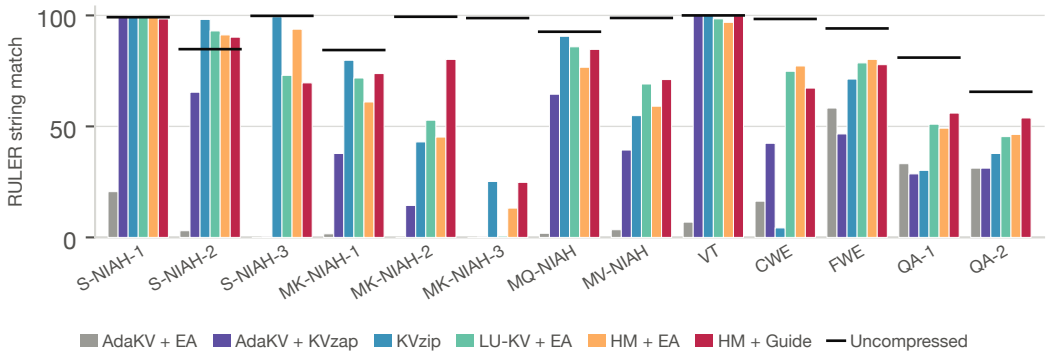

Figure 39: **Per-task scores, RULER-8192, 16×, phi-4.**

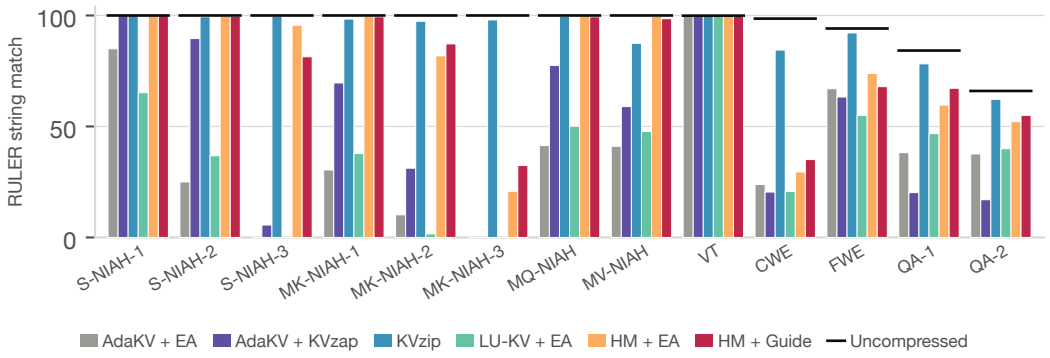

Figure 40: **Per-task scores, RULER-8192, 16×, Qwen3-32B.**

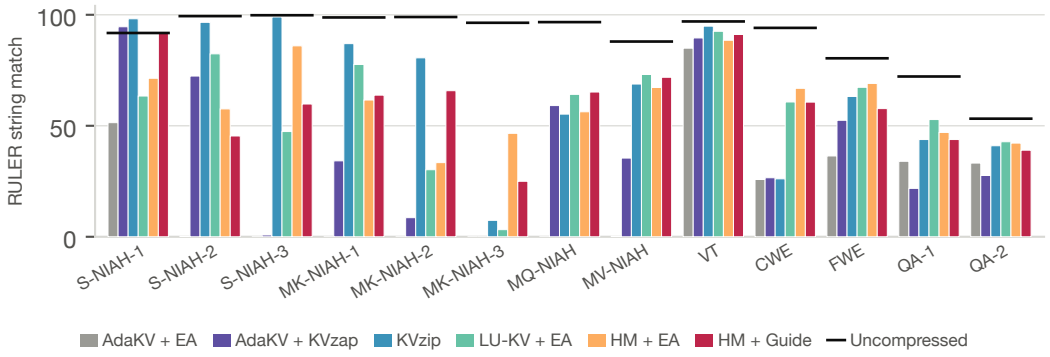

Figure 41: **Per-task scores, RULER-8192, 16×, Mistral-7B.**

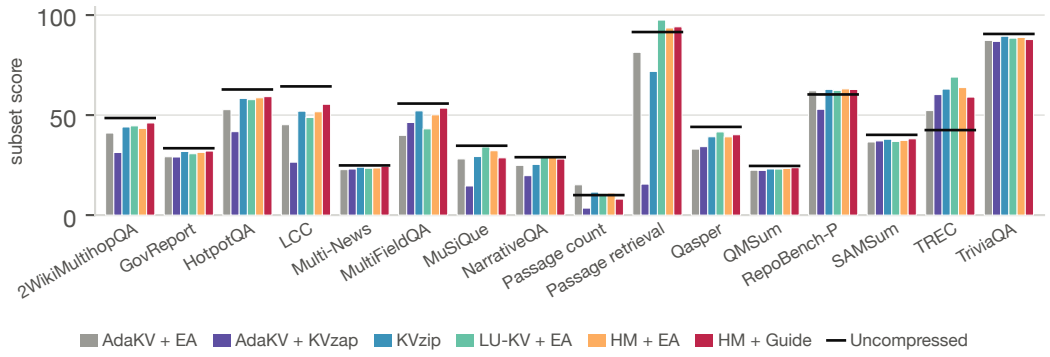

Figure 42: **Per-subset scores, LongBench, 8×, Qwen3-8B.**

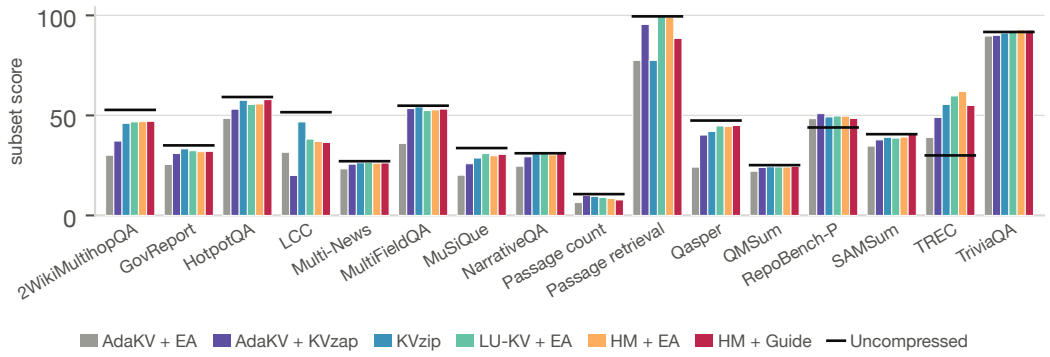

Figure 43: **Per-subset scores, LongBench, 8×, Llama-3.1-8B.**

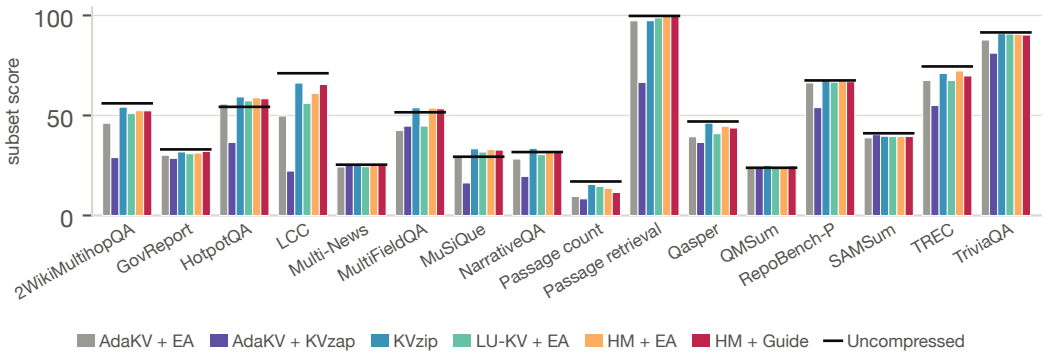

Figure 44: **Per-subset scores, LongBench, 8×, Qwen3-32B.**

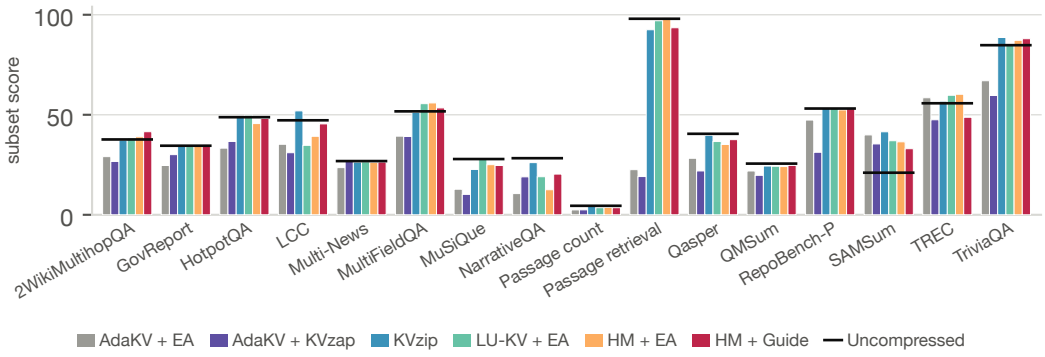

Figure 45: **Per-subset scores, LongBench, 8×, Mistral-7B.**

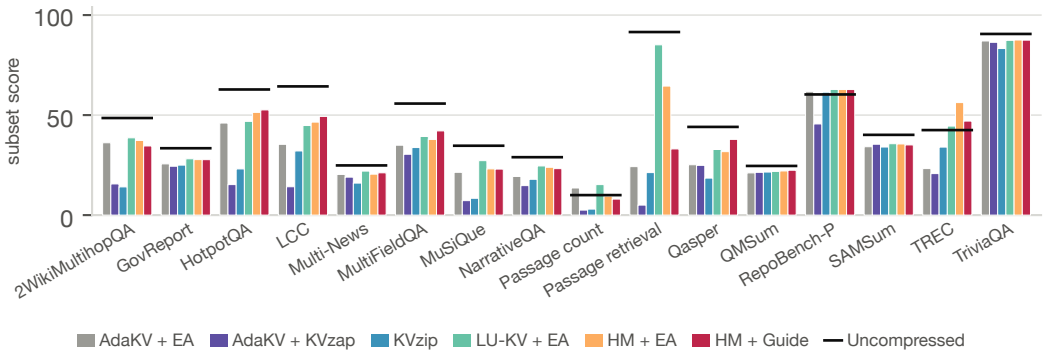

Figure 46: **Per-subset scores, LongBench, 16×, Qwen3-8B.**

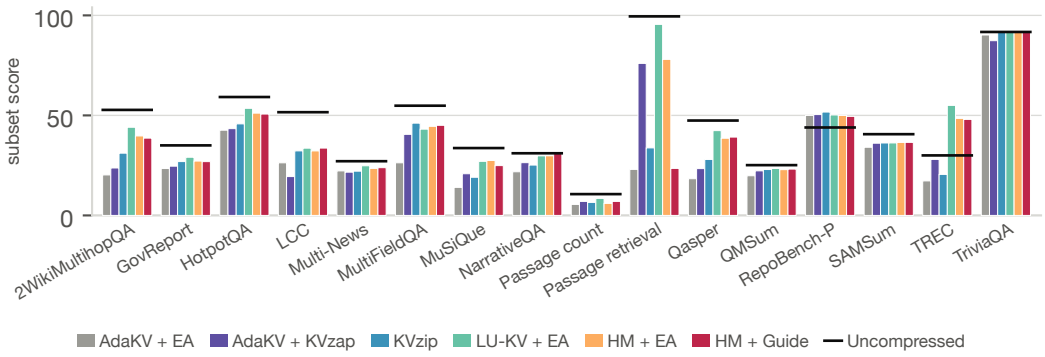

Figure 47: **Per-subset scores, LongBench, 16×, Llama-3.1-8B.**

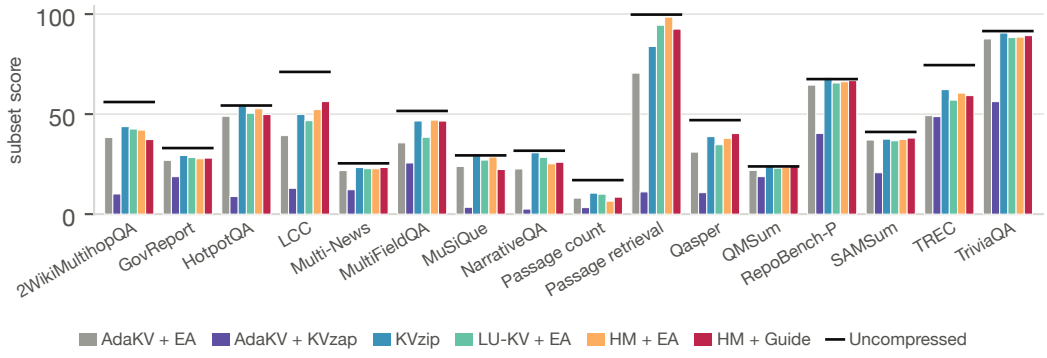

Figure 48: **Per-subset scores, LongBench, 16×, Qwen3-32B.**

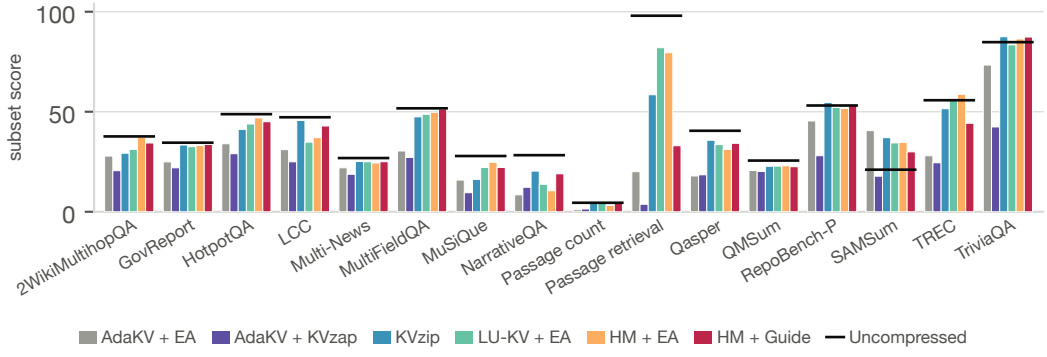

Figure 49: **Per-subset scores, LongBench, 16×, Mistral-7B.**

G PROOFS

Throughout this appendix we retain the notation of the main text. Recall that n refers to the length of X (the number of input positions, indexed by s and output positions, indexed by t), that $J^{(s)}$ without argument refers to $J^{(s)}(X)$ and J refers to $J(X)$, and that $\eta := \inf_{0<r<R} \left(\frac{2\delta}{r} + Mr\right)$.

Remark. In practice, the next-token distribution is blind to a fixed constant being added to each logit, but this is trivially ignored by applying the theorem to the centered logits.

G.1 ASSUMPTIONS

Theorem 1 relies on three assumptions, which we state formally here. The first, smoothness, is all that Lemma 1 needs.

Assumption 1 (Jacobian Smoothness)**.** *Each of f and f' is continuously differentiable in each input position, and there are constants $M \geq 0$ and $R > 0$ such that*

$$\left\| \left[J(X^{(s)+r\mathbf{u}})\right]_t - [J(X)]_t \right\|_F, \left\| \left[J'(X^{(s)+r\mathbf{u}})\right]_t - [J'(X)]_t \right\|_F \leq Mr$$

$\mathbf{P}$-almost everywhere in X, for all $0 \leq r < R$ and unit directions $\mathbf{u}$, and for all $s \leq t$ over all t.

Assumption 1 is a fairly standard assumption on the curvature of the model f. It is quantifiable: M can be estimated efficiently using Jacobian-vector products. Empirically, we find that larger models are smoother (Section E), which is consistent with larger models agreeing more with the consensus in Section 3.1.

The other two constrain the dependence structure of the Jacobian. For $k \in \mathbb{N}$, write $\Psi_k(X) = \frac{1}{n\sqrt{d}} \sum_{(s,t):\, s \leq t-k} \|[J^{(s)}]_t\|_F$ for the lag-k tail of the Jacobian block norms of f, and $\Psi'_k(X)$ for the same quantity for f'.

Assumption 2 (Preceding-Token Dependence)**.** *There is a constant $c > 0$ such that*

$$\left\| \left[J^{(t)}(X)\right]_t \right\|_F^2, \left\| \left[J'^{(t)}(X)\right]_t \right\|_F^2 \geq cd$$

$\mathbf{P}$-almost everywhere in X, for all output positions t.

Assumption 2 states that every prediction depends, by a margin bounded away from zero, on the last token it reads, the one at its own position. We find this block above zero at every sampled position of every model we examine (Section E.2).

Assumption 3 (Lag Decay)**.** *There is a non-increasing function $\psi : \mathbb{N} \to [0, \infty)$, independent of n and d, with $\psi(k) \to 0$ as $k \to \infty$, such that*

$$\Psi_k(X), \Psi'_k(X) \leq \psi(k)$$

$\mathbf{P}$-almost everywhere in X, for all $k \in \mathbb{N}$.

Assumption 3 is the one that depends most on the data: it states that the dependence of a prediction on earlier tokens falls toward zero with their distance, at a rate that does not depend on the length of the sequence. We observe a clear but slow decay, close to a power law, and larger models decay more slowly (Section E.2). The assumption is needed only because of our choice of units: an alternative proof with squared block norms gives a bound independent of n without it. Whether a milder assumption suffices for this statistic, and whether its normalisation is the right one, is an open question.

G.2 PROOFS

Lemma 1 (Near-Identifiability of the Jacobian Blocks)**.** *Let $\mathbf{P}(X)$ be a distribution supported on a space of canonical input sequences $\mathcal{X}$, where each element of the sequence is an embedding in $\mathbb{R}^d$ and sequences may be of any length n. Let f, f' be the functions of two large language models mapping sequences of input embeddings to sequences of output logit vectors, each of whose Jacobians is smooth in each input position in the sense of Assumption 1, with constants M and R as given there. Suppose that the next-token logits implied by f and f' are similar for*

each output position, in the sense that

$$\left\| f_t(X^{(s)+r\mathbf{u}}) - f'_t(X^{(s)+r\mathbf{u}}) \right\| \le \delta$$

$\mathbf{P}$*-almost everywhere in* X*, for some* $\delta \ge 0$*, for all* $0 \le r < R$ *and unit directions* $\mathbf{u}$*, and for all* $s \le t$ *over all* t*. Then, writing* $\eta := \inf_{0<r<R} \left(\frac{2\delta}{r} + Mr \right)$*, we have*

$$\left\| \left[J^{(s)}(X) \right]_t - \left[J'^{(s)}(X) \right]_t \right\|_F \le \sqrt{d}\,\eta$$

$\mathbf{P}$*-almost everywhere in* X*, for all* $s \le t$ *over all* t*. Moreover,* $\eta \le \frac{4\delta}{R} + \frac{MR}{2}$*, and* $\eta = 2\sqrt{2M\delta}$ *whenever* $2\delta \le MR^2$*; in particular,* η *depends on neither* n *nor* d*, and* $\eta \to 0$ *as* $M, \delta \to 0$*.*

Proof of Lemma 1. **Claim.** For all $0 \le r < R$, unit vectors $\mathbf{u}$, and positions $s \le t$, we have $\|f_t(X^{(s)+r\mathbf{u}}) - f_t(X) - r \left[J^{(s)} \right]_t \mathbf{u}\| \le \frac{1}{2} M r^2$. The same bound holds for f' and $[J'^{(s)}]_t$.

Proof of claim. Write $g(\rho) = f_t(X^{(s)+\rho\mathbf{u}})$ for $\rho \in [0, r]$. By Assumption 1, g is continuously differentiable, and by the chain rule $g'(\rho) = \left[J^{(s)}(X^{(s)+\rho\mathbf{u}}) \right]_t \mathbf{u}$. By the fundamental theorem of calculus,

$$\begin{aligned} \left\| f_t(X^{(s)+r\mathbf{u}}) - f_t(X) - r \left[J^{(s)} \right]_t \mathbf{u} \right\| &= \left\| g(r) - g(0) - r \left[J^{(s)} \right]_t \mathbf{u} \right\| \\ &= \left\| \int_0^r \left(\left[J^{(s)}(X^{(s)+\rho\mathbf{u}}) \right]_t - \left[J^{(s)} \right]_t \right) \mathbf{u}\, d\rho \right\| \\ &\le \int_0^r M\rho\, d\rho = \frac{1}{2} M r^2 \end{aligned}$$

where the last line follows from the triangle inequality for integrals together with Assumption 1, using $\|A\mathbf{u}\| \le \|A\|_{\mathrm{op}} \le \|A\|_F$ for unit $\mathbf{u}$. The identical argument applies to f' and J'.

Claim. For all positions $s \le t$ and every unit vector $\mathbf{u}$, we have $\left\| \left(\left[J^{(s)} \right]_t - \left[J'^{(s)} \right]_t \right) \mathbf{u} \right\| \le \eta$.

Proof of claim. Let $0 < r < R$ be arbitrary. By the previous claim, we have

$$f_t(X^{(s)+r\mathbf{u}}) - f_t(X) - r \left[J^{(s)} \right]_t \mathbf{u} - \zeta = 0$$

for some $\|\zeta\| \le \frac{1}{2} M r^2$, and similarly for f', $[J'^{(s)}]_t$, and ζ'. Subtracting the two equations and re-arranging yields

$$\begin{aligned} \left\| r \left[J^{(s)} \right]_t \mathbf{u} - r \left[J'^{(s)} \right]_t \mathbf{u} \right\| &= \left\| f_t(X^{(s)+r\mathbf{u}}) - f'_t(X^{(s)+r\mathbf{u}}) - f_t(X) + f'_t(X) - \zeta + \zeta' \right\| \\ &\le 2\delta + Mr^2 \end{aligned}$$

where the final bound follows by the triangle inequality and the near-agreement of the next-token logits hypothesized in the lemma, the latter applied at r and at 0. Dividing by r and taking the infimum over $0 < r < R$ yields the claim.

It remains to pass to the Frobenius norm and to estimate η. Write $A := \left[J^{(s)} \right]_t - \left[J'^{(s)} \right]_t \in \mathbb{R}^{V \times d}$. Taking the supremum over unit $\mathbf{u}$ in the previous claim gives $\|A\|_{\mathrm{op}} \le \eta$, and since $\mathrm{rank}(A) \le d$ we have $\|A\|_F^2 = \sum_i \sigma_i(A)^2 \le d\|A\|_{\mathrm{op}}^2 \le d\eta^2$. For the explicit estimates, taking $r = R/2$ in the definition of η gives $\eta \le \frac{4\delta}{R} + \frac{MR}{2}$, while the unconstrained minimizer of $r \mapsto \frac{2\delta}{r} + Mr$ over $r > 0$ is $r^* = \sqrt{2\delta/M}$, attaining the value $2\sqrt{2M\delta}$, which is admissible precisely when $r^* \le R$. Both estimates depend only on M, δ and R, and hence on neither n nor d. □

Theorem 1 (Near-Identifiability of the Per-Token Marginal Influence, formal)**.** *Let* $\mathbf{P}(X)$ *be a distribution supported on a space of canonical input sequences* $\mathcal{X}$*, where each element of the sequence is an embedding in* $\mathbb{R}^d$ *and sequences may be of any length* n*. Let* f, f' *be the functions of two large language models mapping sequences of input embeddings to sequences of output logit vectors. Assume that (1) the prediction of the next token depends sufficiently on the immediately preceding token (Assumption 2), (2) each model's Jacobian is smooth in each input position (Assumption 1), and (3) the blockwise Jacobian norms of each model decay with lag (Assumption 3), with constants* c*,* M *and* R *and envelope* ψ *as given there. Suppose moreover*

that the next-token logits implied by f and f' are similar for each output position, in the sense that

$$\left\| f_t(X^{(s)+r\mathbf{u}}) - f'_t(X^{(s)+r\mathbf{u}}) \right\| \le \delta$$

P-*almost everywhere in X, for some $\delta \ge 0$, for all $0 \le r < R$ and unit directions $\mathbf{u}$, and for all $s \le t$ over all t. Then, we have*

$$\|\boldsymbol{\jmath}(X) - \boldsymbol{\jmath}'(X)\|_1 \le \epsilon$$

P-*almost everywhere in X, where $\boldsymbol{\jmath}, \boldsymbol{\jmath}'$ are the per-token marginal influences of f and f' respectively, and where ϵ depends on neither n nor d, and $\epsilon \xrightarrow{M,\delta\to 0} 0$. Explicitly, $\epsilon := \min\{2, \frac{2}{\sqrt{c}} \inf_{k\in\mathbb{N}}(k\eta + 2\psi(k))\}$, which depends on c and ψ.*

Proof of Theorem 1. The hypotheses of the theorem include those of Lemma 1, which therefore gives $\|[J^{(s)}]_t - [J'^{(s)}]_t\|_F \le \sqrt{d}\,\eta$ for all $s \le t$.

Claim. We have $Z(X) \ge n\sqrt{cd}$ and $Z'(X) \ge n\sqrt{cd}$; equivalently, $\Psi_0(X), \Psi'_0(X) \ge \sqrt{c}$. In particular, $\boldsymbol{\jmath}(X)$ and $\boldsymbol{\jmath}'(X)$ are well defined.

Proof of claim. Every term of $Z(X) = \sum_{s\le t}\|[J^{(s)}]_t\|_F$ is non-negative, so discarding all off-diagonal terms and applying Assumption 2 to each of the n diagonal blocks gives

$$Z(X) \ge \sum_{t=1}^{n} \left\| \left[J^{(t)}\right]_t \right\|_F \ge n\sqrt{cd},$$

and identically for f'. Since $c > 0$, both normalizations are strictly positive.

Claim. Let $F, F' \in [0,\infty)^n$ with $Z := \|F\|_1 > 0$ and $Z' := \|F'\|_1 > 0$. Then

$$\left\| \frac{F}{Z} - \frac{F'}{Z'} \right\|_1 \le \frac{2\|F - F'\|_1}{\max(Z, Z')}.$$

Proof of claim. Decompose

$$\frac{F}{Z} - \frac{F'}{Z'} = \frac{F - F'}{Z} + F'\left(\frac{1}{Z} - \frac{1}{Z'}\right).$$

The second term has ℓ^1 norm $\|F'\|_1 \frac{|Z'-Z|}{ZZ'} = \frac{|Z-Z'|}{Z}$, and $|Z - Z'| = \big|\|F\|_1 - \|F'\|_1\big| \le \|F - F'\|_1$ by the reverse triangle inequality. Hence the total is at most $2\|F - F'\|_1/Z$. Exchanging the roles of F and F' gives the same bound with Z' in place of Z, so we may take whichever of the two is larger.

Claim. For every $k \in \mathbb{N}$, we have $\|F(X) - F'(X)\|_1 \le n\sqrt{d}\,(k\eta + 2\psi(k))$.

Proof of claim. By the triangle inequality applied to each $F_s(X) - F'_s(X)$,

$$\|F(X) - F'(X)\|_1 \le \sum_{s\le t} \left| \left\| \left[J^{(s)}\right]_t \right\|_F - \left\| \left[J'^{(s)}\right]_t \right\|_F \right|.$$

Partition $\{(s,t) : s \le t\}$ into the recent pairs $\mathcal{R}_k := \{(s,t) : t - k < s \le t\}$ and the distant pairs $\mathcal{D}_k := \{(s,t) : s \le t - k\}$. On $\mathcal{R}_k$, the reverse triangle inequality and Lemma 1 give

$$\sum_{(s,t)\in\mathcal{R}_k} \left| \left\| \left[J^{(s)}\right]_t \right\|_F - \left\| \left[J'^{(s)}\right]_t \right\|_F \right| \le \sum_{(s,t)\in\mathcal{R}_k} \left\| \left[J^{(s)}\right]_t - \left[J'^{(s)}\right]_t \right\|_F \le nk\sqrt{d}\,\eta,$$

since each output position t contributes at most k recent pairs, so $|\mathcal{R}_k| \le nk$. On $\mathcal{D}_k$, we discard the cancellation via $|a - b| \le a + b$ and recognize the two resulting sums as the lag-k tails $\Psi_k(X)$ and $\Psi'_k(X)$, so that by Assumption 3,

$$\sum_{(s,t)\in\mathcal{D}_k} \left(\left\| \left[J^{(s)}\right]_t \right\|_F + \left\| \left[J'^{(s)}\right]_t \right\|_F \right) = n\sqrt{d}\,(\Psi_k(X) + \Psi'_k(X)) \le 2n\sqrt{d}\,\psi(k).$$

Adding the two contributions yields the claim. For $k = 0$ the set $\mathcal{R}_0$ is empty and the bound holds trivially.

Claim. Define

$$\epsilon := \min\left\{2,\ \frac{2}{\sqrt{c}} \inf_{k\in\mathbb{N}} (k\eta + 2\psi(k))\right\}.$$

Then ϵ depends on neither n nor d, and $\epsilon \to 0$ as $M, \delta \to 0$.

Proof of claim. Independence of n and d is immediate, since η depends only on M, δ, R, the function ψ is independent of n and d by Assumption 3, and c is a fixed constant. For the limit, fix $\varepsilon > 0$. As $\psi(k) \to 0$, choose k_0 with $4\psi(k_0)/\sqrt{c} < \varepsilon/2$. Holding k_0 fixed, we have $\eta \le \frac{4\delta}{R} + \frac{MR}{2} \to 0$ as $M, \delta \to 0$, so $2k_0\eta/\sqrt{c} < \varepsilon/2$ for all sufficiently small M, δ, whence $\epsilon \le \frac{2}{\sqrt{c}}(k_0\eta + 2\psi(k_0)) < \varepsilon$.

Combining the three preceding claims, for every $k \in \mathbb{N}$ we have

$$\|\boldsymbol{j}(X) - \boldsymbol{j}'(X)\|_1 \le \frac{2\|F(X) - F'(X)\|_1}{\max(Z(X), Z'(X))} \le \frac{2n\sqrt{d}\,(k\eta + 2\psi(k))}{n\sqrt{cd}} = \frac{2}{\sqrt{c}}\,(k\eta + 2\psi(k)),$$

while $\|\boldsymbol{j}(X) - \boldsymbol{j}'(X)\|_1 \le 2$ unconditionally, both $\boldsymbol{j}(X)$ and $\boldsymbol{j}'(X)$ being non-negative with unit ℓ^1 norm. Taking the infimum over $k \in \mathbb{N}$ gives $\|\boldsymbol{j}(X) - \boldsymbol{j}'(X)\|_1 \le \epsilon$. □