

Attention Plasticity and the Geometry of Long-Context Failure

Viktor Shcherbakov
2026

University of Geneva



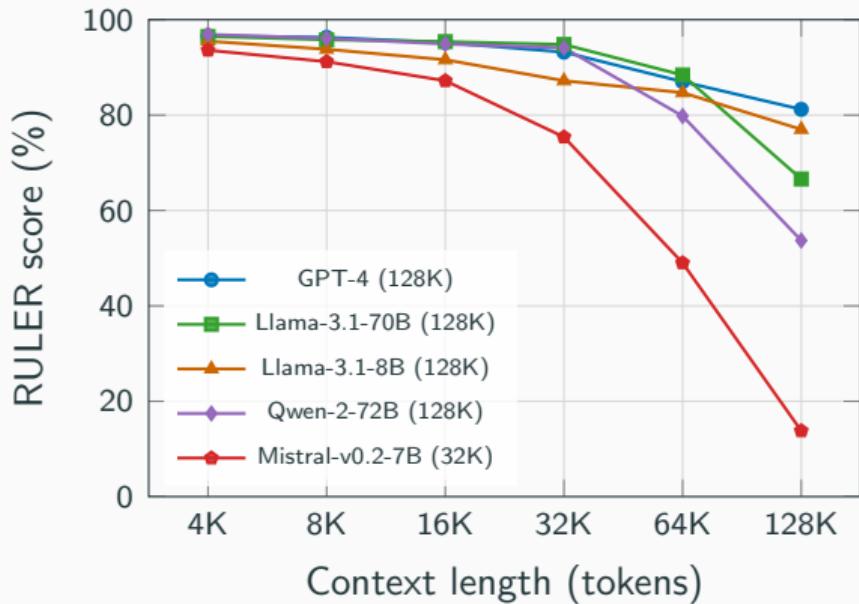
**UNIVERSITÉ
DE GENÈVE**

The Problem: Claimed vs. Effective Context

Models advertise 128K+ token windows, yet **effective** context falls far short.

Benchmarks detect the gap but cannot explain *why*.

This thesis: trace the gap to the geometry of attention heads.



Hypothesis: Attention as Content-Based Reranking

Attention computes a **ranking** over keys via dot-product scores:

$$\text{score}(q, k_i) = q^\top k_i = \underbrace{q_{\text{content}}^\top k_{\text{content}}}_{\text{content relevance}} + \underbrace{q_{\text{pos}}^\top k_{\text{pos}}}_{\text{positional bias}}$$

When position dominates \Rightarrow ranking becomes **rigid**: the model ranks by position rather than by content relevance.

Hypothesis: This rigidity is the mechanistic pathway from architecture to effective context failure.

Research Questions

RQ1 How does position information manifest in the geometry of attention heads?

→ PCA decomposition + Planar rotation model

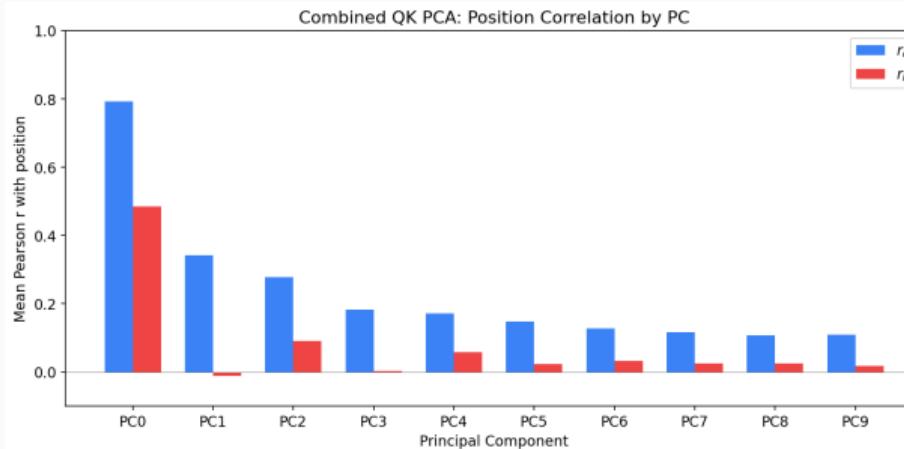
RQ2 Does positional bias functionally constrain attention?

→ Attention plasticity metric + decay theorem

RQ3 Do plasticity profiles correspond to behavioral performance?

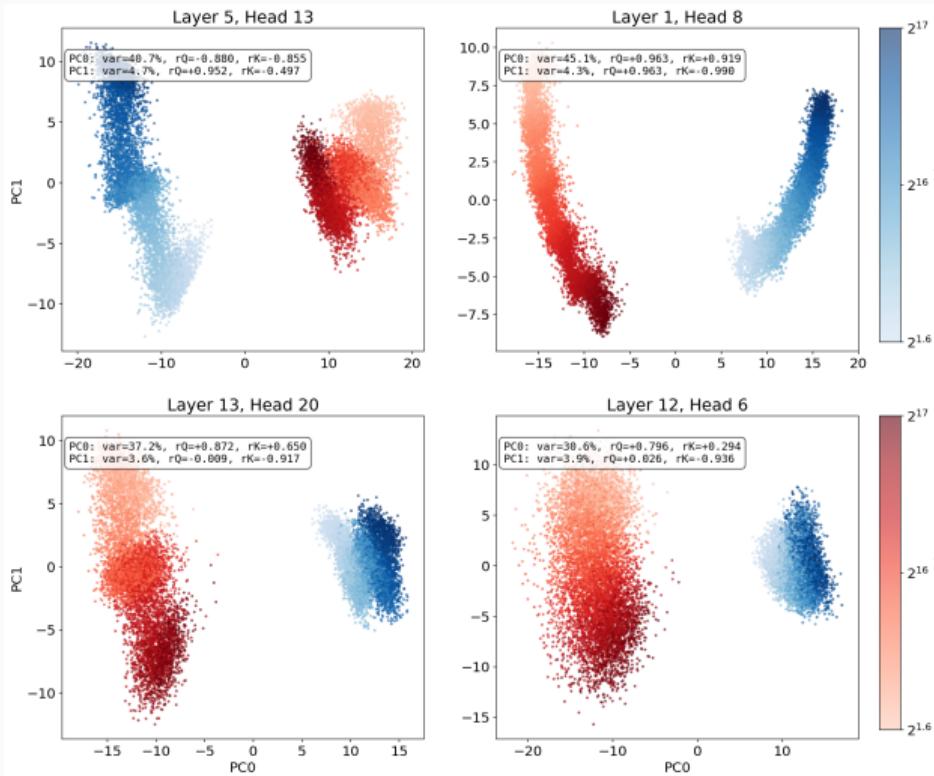
→ Cross-model benchmark correlation + training dynamics

PCA: Position Dominates Q/K Variance

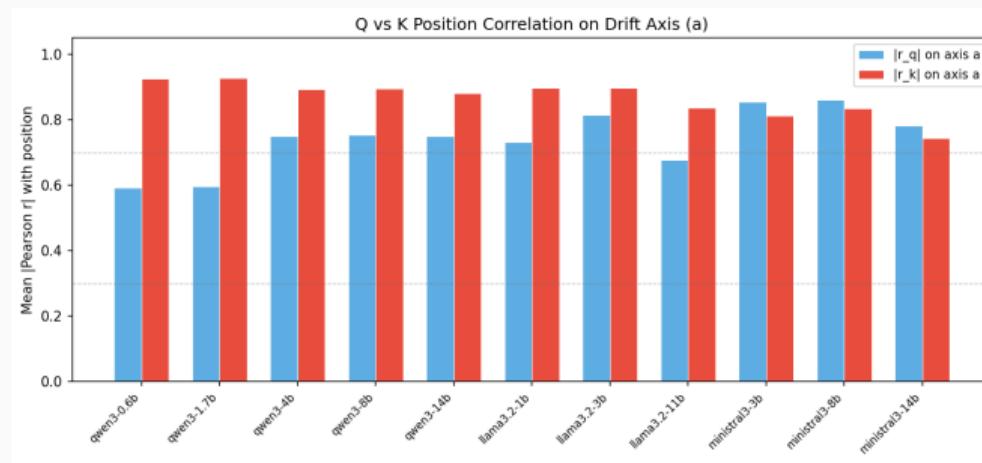


- PC0 captures $\sim 34\%$ of Q+K variance ($4 \times$ PC1). 23–32% of query and 9–20% of key variance is **linear in position**.
- On PC0: $|r_q| \approx 0.80 > |r_k| \approx 0.49$ — but **this is a confound** (PC0 mixes position with Q/K identity).

PCA: What the First Two Components Look Like



Rotation: Isolating the Bias Mechanism



- On drift axis a: **asymmetry reverses** — $|r_k^{(a)}| \approx 0.87 > |r_q^{(a)}| \approx 0.74$. Keys encode position more strongly than queries.
- Parametric bias: $\text{bias_str} = \mu_Q^a \times \alpha_K$. 99% of 3,239 heads show recency bias; tight within families.

Attention Plasticity: The Functional Test

Given query q and two keys k_i, k_j , does *content* determine which key ranks higher?

Score difference $D = q^\top (k_i - k_j) \sim \mathcal{N}(\mu, v)$.

$$p = \Phi\left(\frac{\mu}{\sqrt{v}}\right), \quad \text{PP} = 4p(1-p)$$

- PP $\rightarrow 1$: content determines the ranking (plastic)
- PP $\rightarrow 0$: position locks the ordering (rigid)

Theorem: Under linear positional drift, $\text{AP}_t \rightarrow 0$ as $t \rightarrow \infty$. How much it drops *within* the context window is model-dependent.

Experimental Setup

Cross-model study (13 models)

Family	#	Context
Minstral-3	3B/8B/14B	256K
Qwen-3	0.6B–14B	128K
Llama-3.2	1B/3B/11B	128K
Llama-3.1	8B	128K
Mistral-v0.2	7B	32K

Training dynamics

- SmoILM3-3B
- 10 checkpoints
- Pre-train → anneal → LC ext.
- 4K → 32K → 64K context

Benchmarks

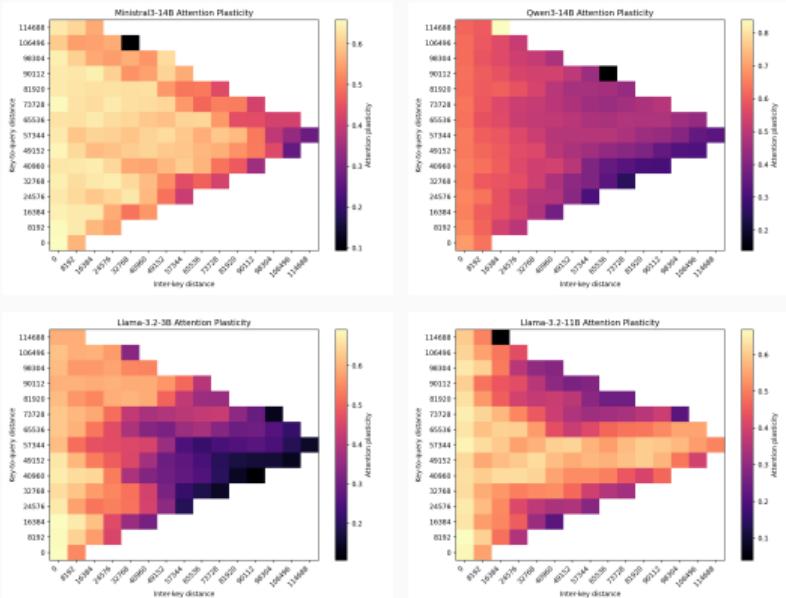
- LongBench-Pro (7 models)
- RULER (2 predecessor models)

Result: Plasticity Profiles Separate Families

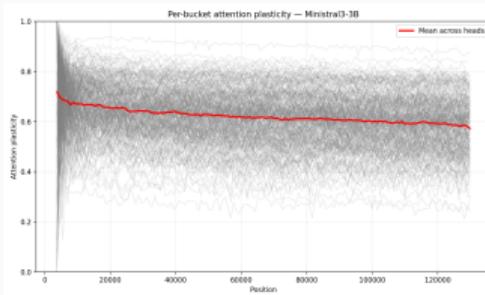
Family	0–20%	80–100%	AP_{drop}
Minstral-3	.655–.664	.571–.592	.07–.08
Qwen-3	.680–.702	.512–.540	.16–.19
Llama-3.2	.658–.703	.456–.489	.17–.23

Every model declines. The **rate** separates families:

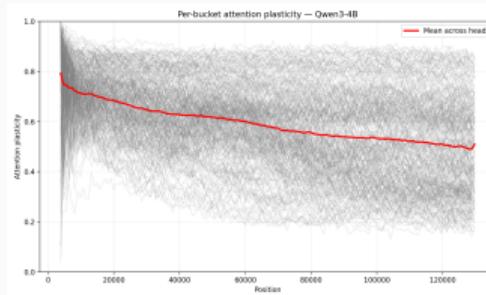
- Minstral: gradual, near-linear
- Qwen: steeper, accelerates
- Llama: steepest overall



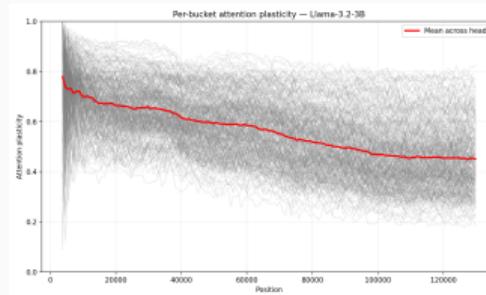
Result: Per-Head Strategies Differ by Family



Minstral-3-3B
Tight, homogeneous bundle



Qwen-3-4B
Large spread

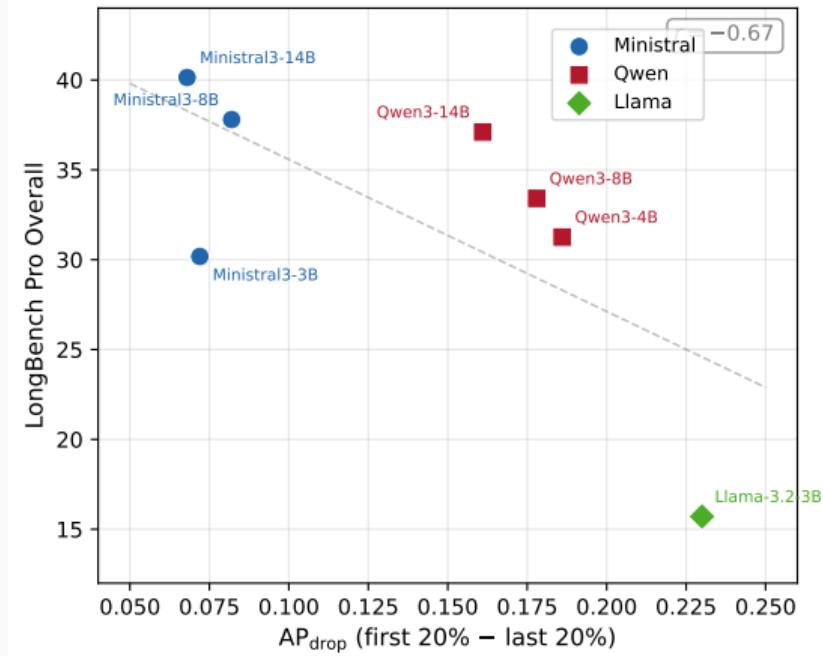


Llama-3.2-3B
Steepest decline

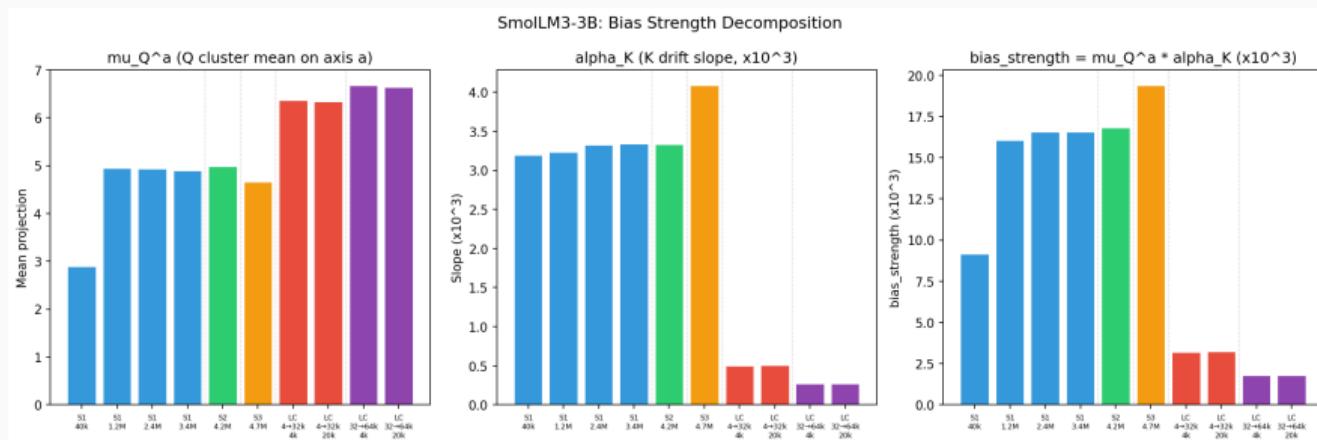
Result: AP_{drop} Predicts LongBench-Pro

Family	AP _{drop}	LBP
Minstral	~0.07	30–40
Qwen	~0.17	31–37
Llama	0.23	16

Diagnostic outlier: Minstral-3-3B — lowest AP_{drop} (0.072), highest aggregate AP (0.622), but LBP only 30.18.
⇒ Context preservation is good; base capability is the bottleneck.



Result: Training Dynamics — Three Phases



Phase 1 — Pre-training

Bias doubles ($0.009 \rightarrow 0.019$).
 $AP_{drop} \approx 0.06$ (mild).

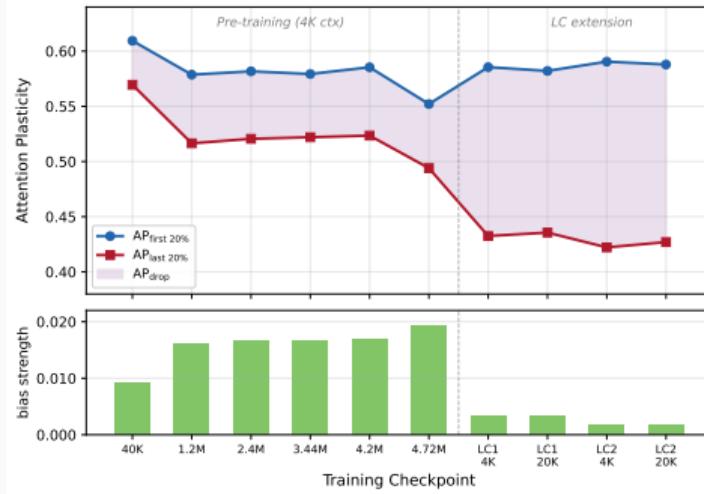
Phase 2 — LC 4K to 32K

Bias collapses $6\times$ via α_K flattening. Short-context AP recovers.

Phase 3 — LC 32K to 64K

Bias halves again (total $10\times$).
But AP_{drop} triples to 0.16.

Result: Plasticity Trajectory



Short-context plasticity ($AP_{\text{first } 20\%}$) **recovers** during LC extension.

Long-context plasticity ($AP_{\text{last } 20\%}$) **continues to decline**.

The gap (AP_{drop}) triples despite 10 \times bias reduction.

Limitations

- **Observational, not causal.** Associations between geometry and performance, not interventions.
- **3 families, 1 training trajectory.** Limited generalization to MoE, hybrids, or larger scales.
- **Pairwise ranking, not softmax weights.** Captures ranking quality, not weight allocation.
- **Base vs. instruct confound.** Mechanistic metrics on base models; benchmarks on instruct variants.
- **Linear drift / Gaussian assumptions.** Empirically supported but approximate; non-linear RoPE structure not captured.

Contributions

1. **Geometric framework.** PCA → Rotation → Plasticity — each resolving what the previous leaves open.
2. **Plasticity decay theorem.** Formal proof + Gaussian closed form decomposing decay into positional and content components.
3. **Cross-model validation.** AP_{drop} separates families in benchmark order across 13 models.
4. **Bias ≠ context.** 10× bias collapse but AP_{drop} triples. Content signal decay is the underexplored dimension.

Future Work

Interventional Ablate low-plasticity heads, measure retrieval accuracy at distance.

Content decay Which mechanism? RoPE rotation accumulation, attention sinks, or feature drift?

Broader models MoE, attention-SSM hybrids, 70B+ scales.

Per-length validation Full per-length LBP/RULER correlation across all matched models.

Thank you

Questions?