UNIVERSITÉ
DE GENÈVE

Université de Genève

Faculty of Science

Department of Computer Science

Attention Plasticity and the Geometry of Long-Context Failure

Master Thesis

presented by

Viktor Shcherbakov

Machine Learning Group

Supervisor:
Prof. François Fleuret

Geneva, February 16, 2026

Post tenebras lux.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Prof. François Fleuret, for his guidance, patience, and unwavering support throughout this project. His willingness to engage in open discussions and his flexibility during the many shifts in research direction were invaluable to the completion of this work.

I am equally grateful to Prof. Martin Jaggi for hosting me at the Machine Learning and Optimization (MLO) laboratory at EPFL during my academic exchange. His insightful discussions and constructive feedback significantly shaped the direction of this thesis. I also wish to thank the members of the MLO lab for the productive discussions following my presentation of this work, which helped refine many of the ideas presented here. In particular, I am grateful to Diba Hashemi and Amirkeivan Mohtashami for insightful and fruitful conversations that improved this thesis.

The supportive and collaborative environment provided by both professors made navigating the challenges of this research a rewarding experience.

Geneva, February 16, 2026 Viktor Shcherbakov

Abstract

Large language models advertise context windows of 128K tokens or more, yet their effective context length—the range over which they reliably use information—often falls far short. Behavioral benchmarks detect this gap but cannot explain it. We develop a mechanistic framework that traces long-context degradation to the geometry of query and key representations inside attention heads.

Our method captures post-RoPE query and key vectors and applies three complementary analyses. PCA decomposition reveals that position encoding dominates variance structure across heads. A planar rotation model isolates the asymmetric positional drift direction and yields a scalar bias strength per head. Attention plasticity then measures the functional consequence: the probability that a random query flips the preference ordering of two keys. We prove that plasticity decays with query position under linear positional drift, and derive a Gaussian closed form that decomposes the decay into positional and content components.

We analyze 13 models from three families (Ministral-3, Qwen-3, Llama-3.2) and track 10 training checkpoints of SmolLM3-3B through pre-training and long-context extension. Plasticity drop—the degradation from early to late context positions—separates model families in the same order as LongBench-Pro scores: Ministral (∼0.07) outperforms Qwen (∼0.17) outperforms Llama (0.23). Training dynamics reveal that RoPE frequency rescaling collapses positional bias but does not prevent plasticity decay at distant positions, indicating that bias reduction is necessary but not sufficient for effective long context.

Résumé

Les grands modèles de langage annoncent des fenêtres de contexte de 128K tokens ou plus, mais leur longueur de contexte effective—la portée sur laquelle ils utilisent l'information de manière fiable—reste souvent bien en deçà. Les benchmarks comportementaux détectent cet écart sans pouvoir l'expliquer. Nous développons un cadre mécaniste qui retrace la dégradation en contexte long à la géométrie des représentations de requête et de clé au sein des têtes d'attention.

Notre méthode capture les vecteurs de requête et de clé après l'application de RoPE et applique trois analyses complémentaires. La décomposition en composantes principales révèle que l'encodage positionnel domine la structure de variance des têtes. Un modèle de rotation planaire isole la direction de dérive positionnelle asymétrique et produit une force de biais scalaire par tête. La plasticité attentionnelle mesure ensuite la conséquence fonctionnelle : la probabilité qu'une requête aléatoire inverse l'ordre de préférence entre deux clés. Nous prouvons que la plasticité décroît avec la position de la requête sous une dérive positionnelle linéaire, et dérivons une forme fermée gaussienne qui décompose cette décroissance en composantes positionnelle et sémantique.

Nous analysons 13 modèles issus de trois familles (Ministral-3, Qwen-3, Llama-3.2) et suivons 10 points de contrôle d'entraînement de SmolLM3-3B à travers le pré-entraînement et l'extension de contexte long. La chute de plasticité—la dégradation entre les positions de contexte proches et lointaines—sépare les familles de modèles dans le même ordre que les scores LongBench-Pro : Ministral (~0.07) surpasse Qwen (~0.17) qui surpasse Llama (0.23). La dynamique d'entraînement révèle que le recalibrage des fréquences RoPE effondre le biais positionnel mais n'empêche pas la décroissance de la plasticité aux positions lointaines, indiquant que la réduction du biais est nécessaire mais non suffisante pour un contexte long effectif.

Contents

Chapter 1

Introduction

1.1 Context and Motivation

Transformer-based language models are increasingly deployed on tasks that require processing long inputs: summarizing legal documents, answering questions over codebases, and reasoning across multi-document evidence [9, 10]. These applications depend on the model's ability to attend to, retrieve, and integrate information from distant positions in the input. Accordingly, recent models advertise context windows of 128K tokens or more [25, 32, 33].

Yet a growing body of evaluation work shows that *claimed* context length and *effective* context length can diverge substantially. Models that accept 128K tokens may degrade on tasks requiring retrieval or reasoning beyond 32K [3, 10]. Performance often drops gradually with input length, in a pattern that varies across tasks, models, and position within the context [9, 26]. Behavioral benchmarks can detect this degradation, but they cannot explain it: they measure *that* a model fails at long context without revealing *why*.

Understanding why requires looking inside the model. The attention mechanism—the core component that routes information between positions—computes a relevance score between each query and key vector, producing a ranking over the context. These vectors carry both content information (from learned projections) and position information (from rotary positional encoding [34]). When position information dominates the attention score, the model's ranking of keys becomes rigid: it ranks by position rather than by content relevance. We hypothesize that this rigidity is a mechanistic pathway to effective context failure.

1.2 Problem Statement

The gap between claimed and effective context length is well-documented behaviorally, but the internal mechanisms that produce this gap remain poorly understood. Existing mechanistic studies have identified individual failure pathways—positional bias in attention patterns [26], attention sinks [19], fragile retrieval heads [38]—but no existing method provides a unified, quantitative framework that connects internal attention geometry to behavioral long-context performance across models.

This thesis addresses the question: *can the geometry of query and key representations inside attention heads predict and explain effective context length?*

1.3 Research Questions

We pursue three research questions, each addressed by a corresponding analysis:

RQ1. *How does position information manifest in the geometry of attention heads?* We apply PCA to post-RoPE query and key vectors and a targeted planar rotation to isolate the positional drift axis, yielding a parametric bias strength per head.

RQ2. *Does positional bias functionally constrain attention?* We define attention plasticity—the probability that a random query flips the preference ordering of two keys—and prove that it decays with query position under linear positional drift.

RQ3. *Do plasticity profiles correspond to behavioral long-context performance?* We correlate plasticity profiles with LongBench-Pro [9] and RULER [10] benchmark scores across model families, and track how plasticity evolves during training through SmolLM3 checkpoints [5].

1.4 Contributions

This thesis makes the following contributions:

1. **A geometric framework for attention head analysis.** We develop three complementary analyses—PCA decomposition, planar rotation model, and attention plasticity—that operate on the same captured post-RoPE query and key vectors. PCA provides a model-agnostic structural fingerprint. The rotation model isolates the positional drift direction and parameterizes bias strength. Attention plasticity measures the functional consequence of that bias.

2. **A formal characterization of plasticity decay.** We prove that attention plasticity decays with query position under linear positional drift and derive a Gaussian closed form that decomposes the decay into positional and content components (Theorem 4.4).

3. **Cross-model empirical validation.** We analyze 13 models from three families and show that plasticity drop—the degradation from early to late context positions—separates model families in the same order as LongBench-Pro benchmark scores [9].

4. **Training dynamics analysis.** We track 10 checkpoints of SmolLM3-3B [5] through pre-training and long-context extension, showing that RoPE frequency rescaling collapses positional bias but does not prevent plasticity decay at distant positions. This demonstrates that bias reduction is necessary but not sufficient for effective long context.

1.5 Thesis Outline

The remainder of this thesis is organized as follows:

Chapter 2 introduces the foundational concepts: effective context length, the transformer attention mechanism, rotary positional encoding, and long-context architectural variants.

Chapter 3 reviews behavioral evaluation of long-context models, mechanistic interpretability of attention patterns, context-extension methods, and positions this thesis within the literature.

Chapter 4 develops the three-analysis framework: Q/K capture protocol, PCA decomposition, the planar rotation model, attention plasticity with its formal decay theorem, and the connections between analyses.

Chapter 5 describes the experimental setup: the 13-model cross-model study, the SmolLM3 training dynamics study, benchmark selection, and implementation details.

Chapter 6 presents the experimental results across models, training checkpoints, and benchmark correlations.

Chapter 7 interprets the findings, discusses limitations, and suggests directions for future work.

Chapter 8 concludes the thesis.

Chapter 2

Background

2.1 Effective Context Length

Modern transformer-based language models advertise context windows ranging from 32K to over 1M tokens. However, the *claimed context length*—the maximum input length the model accepts—is not the same as the *effective context length* (ECL): the longest input at which the model still performs acceptably on a given task [10, 28]. Recent evaluations consistently show that effective context falls substantially short of claimed context [3, 9, 10].

ECL is not a single model constant. The same model may effectively use 128K tokens for simple retrieval (e.g., needle-in-a-haystack) yet degrade at 32K on tasks requiring multi-hop reasoning or global aggregation [9, 10]. Two factors make ECL task-conditional:

- **Task complexity**: retrieval from a specific position is easier than reasoning across distributed evidence. Simple retrieval benchmarks overestimate ECL relative to tasks requiring content integration [10].
- **Performance threshold**: ECL depends on the criterion used to declare failure. A model that scores 90% at 4K and 70% at 128K has different ECL depending on whether the threshold is set at 80% or 60%.

The gap between claimed and effective context length means that applications relying on long-context models—document summarization, multi-document QA, repository-scale code analysis—may silently degrade as inputs grow. Behavioral benchmarks can detect this gap but cannot explain it [3]. Understanding *why* effective context falls short requires examining the internal mechanisms that determine whether a model can attend to, retrieve from, and reason over information at distant positions. This mechanistic perspective motivates the analyses developed in Chapter 4.

2.2 Transformer Attention

The transformer architecture [35] processes sequences through layers of self-attention and feed-forward networks. This section defines the attention mechanism and its variants relevant to this thesis.

2.2.1 Scaled Dot-Product Attention

Given an input sequence of n tokens with embedding dimension d, each attention head computes query, key, and value matrices via learned projections:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \tag{2.1}$$

where $X \in \mathbb{R}^{n \times d_{\text{model}}}$ is the input and $W_Q, W_K \in \mathbb{R}^{d_{\text{model}} \times d}$, $W_V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ are learned projection matrices. The attention output is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{2.2}$$

The score $s(q_t, k_i) = q_t^\top k_i / \sqrt{d}$ measures the relevance of position i to query position t. The softmax normalizes these scores into a probability distribution over positions, and the output is the probability-weighted sum of value vectors.

2.2.2 Causal Masking

In autoregressive language models, position t may only attend to positions $i \leq t$. This is enforced by setting $s(q_t, k_i) = -\infty$ for $i > t$ before the softmax, producing zero attention weight for future positions [35]. Under causal masking, the set of positions competing for attention at query position t grows linearly with t—an asymmetry that underlies the context-length dependence studied in this thesis.

2.2.3 Multi-Head Attention

Rather than a single attention function with d_{model}-dimensional keys, the transformer uses n_h parallel attention heads, each operating on $d = d_{\text{model}}/n_h$ dimensions. Multi-head attention allows different heads to attend to different positions and learn different patterns [35]. The outputs are concatenated and projected:

$$\text{MHA}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_{n_h})\, W_O \tag{2.3}$$

Different heads may develop specialized roles: some encode position, others perform content-based retrieval, and some attend primarily to early tokens regardless of content [19, 38]. This specialization is a key observation exploited by the per-head analyses in Chapter 4.

2.2.4 Grouped-Query Attention

Standard multi-head attention assigns independent K and V projections to each head, which dominates inference memory cost due to the key-value cache. Grouped-query attention (GQA) reduces this cost by sharing a single key-value head across a group of $g = n_q / n_{kv}$ query heads [2]:

$$h_k = \lfloor h_q / g \rfloor \tag{2.4}$$

where h_q is the query head index and h_k is the corresponding key-value head. All models examined in this thesis use GQA (Table 5.1). The sharing means that query heads within the same group operate on identical keys but with different learned query projections, so each produces distinct attention patterns despite sharing the same key geometry.

2.3 Positional Encoding

Self-attention (Equation 2.2) is permutation-equivariant: without positional information, the attention score $q_t^\top k_i$ depends only on the token identities at positions t and i, not on t and i themselves. Language modeling requires position sensitivity—word order carries meaning—so transformers inject position information into their representations [35].

2.3.1 Rotary Position Embeddings

Rotary Position Embeddings (RoPE) encode position by rotating query and key vectors in 2D subspaces [34]. The head dimension d is partitioned into $d/2$ pairs, and each pair (x_{2k}, x_{2k+1}) is rotated by an angle proportional to the token position t:

$$\mathrm{RoPE}(x, t)_{2k:2k+1} = \begin{pmatrix} \cos(t \cdot \theta_k) & -\sin(t \cdot \theta_k) \\ \sin(t \cdot \theta_k) & \cos(t \cdot \theta_k) \end{pmatrix} \begin{pmatrix} x_{2k} \\ x_{2k+1} \end{pmatrix} \tag{2.5}$$

where $\theta_k = \theta_{\text{base}}^{-2k/d}$ are frequency parameters derived from a base frequency θ_{base} (typically 10,000 [34] or 500,000 [16]). The key property is that the dot product of RoPE-transformed vectors depends only on content and the *relative* position $t - i$:

$$\mathrm{RoPE}(q, t)^\top \mathrm{RoPE}(k, i) = f(q, k, t - i) \tag{2.6}$$

This relative-position property means that the attention score encodes how far apart two tokens are, not their absolute positions. However, post-RoPE vectors carry both content and position information in a coupled form: the rotation entangles the learned content representation with the position angle. This entanglement is what the decomposition methods in Chapter 4 aim to disentangle.

2.3.2 RoPE Frequency Structure

The frequencies θ_k span several orders of magnitude [34]. Low-frequency components (θ_k near $\theta_{\text{base}}^{-1}$) rotate slowly with position, encoding coarse position information over long ranges. High-frequency components rotate rapidly, encoding fine-grained local position. This multi-scale structure means that different dimension pairs carry position information at different granularities.

When input sequences exceed the training-time maximum length, high-frequency components may encounter positions they were never trained on, leading to out-of-distribution rotations [1, 8]. This failure mode motivates the context-extension methods described below.

2.3.3 Context-Length Extension

Three principal methods extend the effective range of RoPE beyond the training-time context length by modifying the frequency parameters:

- **Position interpolation** linearly scales all positions to fit within the original training range, uniformly reducing all frequencies [8].
- **NTK-aware scaling** adjusts the base frequency θ_{base} rather than the positions, preserving high-frequency resolution while extending low-frequency range [1].
- **YaRN** (Yet another RoPE extensioN) combines NTK-aware base frequency adjustment with per-frequency interpolation factors and an attention temperature correction, achieving extension with minimal fine-tuning [31].

These methods are not merely theoretical: several models in our experimental set use them. Qwen-3 models use YaRN to extend from 32K to 128K context (Section 5.1). Understanding how frequency modification changes the geometry of post-RoPE vectors—and consequently the attention bias structure—is part of what the rotation model (Section 4.3) captures.

2.3.4 No Positional Encoding Layers

Some architectures omit positional encoding in selected layers, relying on position information propagated through residual connections from RoPE layers below [5]. These NoPE (No Positional Encoding) layers compute attention scores from content projections alone: $s(q_t, k_i) = q_t^\top k_i / \sqrt{d}$ with no rotary transformation. NoPE layers are expected to show different bias structure than RoPE layers, since they lack the direct position-encoding mechanism. SmolLM3, used in our training dynamics study (Section 5.2), applies NoPE to every fourth layer.

2.4 Attention Topology

The attention pattern—which positions each token can attend to—varies across architectures and directly constrains the model's ability to use long contexts. In full (dense) causal attention, every position t attends to all prior positions $1, \ldots, t$. The computational cost scales quadratically with sequence length ($O(n^2 d)$ for the attention computation [35]), but all pairwise token interactions are preserved. FlashAttention reduces the memory overhead of full attention through IO-aware tiling, making full attention practical at long context lengths without changing the attention pattern [11]. All primary models in our experimental set use full causal attention.

2.4.1 Sliding-Window Attention

Sliding-window attention restricts each position to a local window of size w: position t attends only to positions $\max(1, t-w+1), \ldots, t$ [20]. This reduces the per-layer cost to $O(n \cdot w \cdot d)$ and limits per-layer receptive field to w tokens. Information beyond the window boundary can only reach position t through multi-layer propagation via the residual stream, creating a hard architectural limit on single-layer context access.

Our analysis focuses on full-attention models. The one sliding-window model in our study (Mistral-v0.2-7B) is analyzed only within its 32K window, since key pair sampling and query distribution estimation are not comparable across architectures with different attention spans.

2.4.2 Attention Sinks

Empirically, transformer language models allocate disproportionate attention weight to the first few tokens (typically the BOS token), regardless of their semantic relevance [40]. This "attention sink" pattern is pervasive across models and persists even when the initial tokens carry no useful information. Recent work shows that attention sinks emerge from the interaction between causal

masking and softmax normalization: early tokens, visible to all positions, accumulate attention as a normalization artifact [19].

Attention sinks represent a form of positional bias that consumes attention budget without contributing to content retrieval. In the context of this thesis, they illustrate how architectural constraints (causal masking) and training dynamics can produce position-dependent attention patterns that reduce the model's ability to allocate attention based on content relevance.

Chapter 3

Related Work

3.1 Overview

This chapter reviews prior work on effective context length (ECL), with emphasis on evidence quality and mechanistic interpretability. The literature can be organized into three strands: behavioral evaluation of long-context capability, mechanistic analyses of long-context failure, and context-extension or mitigation methods [8, 10, 19, 28, 31].

The definitions of claimed and effective context length introduced in Section 2.1 apply throughout this chapter.

The chapter first synthesizes behavioral evidence, then reviews mechanistic findings, then surveys mitigation families, and finally positions the research gap addressed by this thesis.

3.2 Behavioral ECL Evaluation Literature

Early long-context evaluation was dominated by retrieval-centric stress tests, especially needle-in-a-haystack (NIAH) setups [21]. In these tasks, a target fact is embedded in long distractor text and the model is asked to recover it. This evaluation style is useful for probing whether a model can access distant tokens, but it is limited as a proxy for broader long-context reasoning, aggregation, and multi-step integration.

Recent benchmarks broaden this picture in complementary ways. RULER extends synthetic evaluation beyond single retrieval tasks and reports substantial degradation as length grows even for models with near-perfect vanilla retrieval scores; in one widely cited setting, only about half of tested models maintain the benchmark threshold at 32K despite claiming at least 32K context [10].

BABILong stresses reasoning-in-a-haystack and reports that many models effectively use only a fraction of their claimed context on its QA1 setting [23]. In parallel, InfiniteBench evaluates more realistic long-input tasks and also finds substantial degradation at 100K+ context lengths [45].

Another important development is explicit confound analysis. NoLiMa shows that many NIAH-style settings contain high literal overlap between query and context, which can inflate measured long-context capability through lexical matching shortcuts [28]. On NoLiMa, performance at long lengths drops sharply for most tested models (for example, 11 of 13 tested models fall below 50% of base score at 32K), indicating that overlap control is critical when interpreting ECL claims [28]. This result is consistent with the broader view that effective context estimates are benchmark-construction dependent [15, 24].

Controlled causal studies further sharpen this evidence. FLenQA uses duplicate-padding controls to isolate length effects and reports significant reasoning degradation as input length grows [24]. Du et al. show that context length alone can hurt performance even with perfect retrieval, including settings with whitespace distractors and attention masking controls [15]. Ada-LEval's truncation-based analysis also indicates that not all nominally long-context tasks truly require full context [36].

At the realism end of the spectrum, LongBench v2 and LongBench Pro provide large natural-document evaluations [4, 9]. LongBench v2 emphasizes difficult human-annotated tasks and shows that many models still struggle on realistic long-input reasoning, with limited gains from longer retrieval context in several settings [4]. LongBench Pro additionally reports explicit claimed-versus-effective mismatches and strong truncation sensitivity for some models with very large advertised windows [9]. These datasets improve external validity, but they offer weaker causal attribution than controlled synthetic or intervention-based studies.

A practical takeaway is that different benchmark families answer different questions:

- synthetic controllable suites are strongest for causal attribution,
- overlap-controlled suites are strongest for shortcut diagnosis,
- realistic document suites are strongest for external validity.

No single benchmark family is sufficient on its own.

A second takeaway is that ECL values are not directly interchangeable across papers, because "acceptable performance" is operationalized differently. For example, RULER defines effective length using a fixed threshold reference, whereas NoLiMa uses a relative threshold tied to each model's short-context base score [10, 28]. Therefore, cross-paper comparison requires caution even when the phrase "effective context length" is shared.

Taken together, the behavioral literature establishes two robust conclusions: first, claimed context length and effective context length can diverge substantially; second, ECL estimates are highly sensitive to benchmark design, especially overlap, task composition, and length-construction choices. However, behavioral measurements alone are insufficient to identify *why* ECL fails, which motivates the mechanistic focus of this thesis.

3.3 Mechanistic Interpretability for Long-Context Failure

Mechanistic long-context analysis builds on the transformer-circuits program: model behavior is treated as computation implemented by identifiable internal subnetworks, and mechanism claims are tested by intervention rather than inferred from benchmark score trends alone [17, 29]. Within that framing, evidence is strongest when studies progress from observational signatures, to targeted perturbations, to confound-controlled evaluation.

At the observational level, recent studies identify stable positional/topological signatures. Gu et al. report attention sinks across model scales from 14M to 13B, including random-token sink rates of 70.29%–91.23%, and show sink emergence across PE families (NoPE, ALiBi, RoPE) [19]. Complementary theory shows how causal masking alone can amplify early-position influence with depth: cumulative context probability converges exponentially toward the first token under the paper's assumptions [39]. Over-squashing analysis gives a related topological account, where earlier tokens have combinatorially more information pathways to the final prediction token than mid/late tokens [6].

Interventional work then tests whether these components are causally relevant. Retrieval-head studies report that only 3–6% of heads exceed retrieval-score thresholds across eight model configurations, and masking about 5% of top retrieval heads (K=50) drops NIAH scores below 50 while random-head masking remains above 80 in reported settings [38]. On extractive QA, the same study reports F1 falling from 56.7 to 32.3 when masking 100 retrieval heads, versus 55.4 for masking 100 random heads [38]. Position-bias interventions provide complementary causal evidence outside standardized ECL benchmarks: PINE reports +3.7 to +11.7 points on the RewardBench reasoning subset across tested configurations (e.g., Llama-3-70B: 78.9 to 87.6) while providing a proof of inter-document position invariance for its transformed attention procedure [37]. A lighter intervention that scales one positional hidden-state channel reports gains up to +9.3 points on NaturalQuestions and +15.2 on KV retrieval with near-unchanged MMLU in reported settings [44]. These results are best read as transfer evidence for mechanism controllability rather than direct ECL measurement.

Mechanistic analyses of positional encoding behavior add complementary evidence. RoPE-focused circuit analysis reports frequency specialization: high-frequency bands support positional heads (for example, diagonal and previous-token patterns), while low-frequency bands carry semantic channels that are less stable at very long ranges [7]. In parallel, position-frequency analysis links

long-context failure to training exposure imbalance: for one reported corpus analysis at $L = 2048$, indices up to 1024 account for over 80% of occurrences while indices at 1536+ are below 5%; their inference-time index-shifting intervention raises average 4-needle NIAH from 67.8 to 85.7 across seven models [3].

Current evidence also has clear limits relevant to this thesis gap. Several mechanistic results are setup-constrained (for example, downstream intervention tests often on one primary model, attention-sink causal ablations scaled to 1B, or higher inference overhead for some methods such as PINE) [19, 37, 38]. Moreover, confound-controlled benchmark studies show that mechanism-informed improvements must still survive robustness checks: NoLiMa reports 11/13 models below 50% of base score at 32K, and controlled length-isolation experiments report 13.9%–85% degradation at 30K despite perfect retrieval in their tested settings [15, 28].

For this thesis, the methodological implication is explicit: long-context mechanism claims are most credible when observational signatures, targeted interventions, and confound-controlled evaluation converge on the same explanation.

3.4 Context-Extension and Mitigation Methods

Context-extension work now spans three practical families: positional-scaling methods (PI, NTK-aware scaling, YaRN, and LongRoPE), hybrid production stacks (for example, RoPE rescaling with chunked attention), and mechanism-aware inference interventions. Early scaling methods established that substantial window expansion is feasible: PI extends LLaMA to 32K with limited adaptation and improves long-range perplexity in its evaluation setup (for example, LLaMA-7B PG-19: 7.20 at 2K to 6.77 at 32K), while YaRN reports 128K perplexity and passkey gains [8, 31]. LongRoPE pushes this line further, reporting extension to 2048K with progressive search and high passkey retrieval in its synthetic setup [14]. NTK-aware scaling remains a common low-cost zero-shot baseline in follow-up evaluations, though evidence is still mostly perplexity-centric and model-specific [14, 31].

Technical reports show these methods are operationally valuable in modern model stacks. Qwen2 reports large long-context gains from YARN+DCA (for example, NeedleBench 256K: 17.13 to 85.21; LV-Eval 256K: 2.88 to 42.35 for 72B Instruct) [42]. Qwen3 continues this design direction and reports strong RULER scores at 128K in non-thinking mode (90.6 on its 235B-A22B model) [32]. In parallel, Kimi Linear reports that a NoPE + KDA hybrid can outperform its RoPE variant on RULER (84.3 vs 78.8 at 128K), indicating that extension gains need not be tied to standard RoPE extrapolation alone [22].

However, strong retrieval-style extension metrics do not reliably transfer to robust effective context on harder evaluations. RULER reports that all evaluated models claim at least 32K context, but only about half exceed the benchmark threshold at 32K [10]. BABILong reports that many

models use only about 10–20% of claimed context on QA1 and that YaRN can fail to extend effective reasoning context despite stable perplexity [23]. InfiniteBench similarly shows large degradation at 100K+ lengths, including weak open-model results in its reported YaRN-Mistral setting (19.96 average; 0.00 on Retrieve.KV) [45]. NoLiMa adds low-overlap controls and finds 11/13 models below 50% of base score at 32K, including models that look much stronger on overlap-heavy settings [28].

Mechanism-aware interventions partly close these gaps but do not eliminate them. STRING reports average 4-needle NIAH gains from 67.8 to 85.7 across seven models and +15.1/+30.9 RULER gains on Llama3.1-70B and Qwen2-72B, raising both from 64K to 100K effective length in that protocol [3]. Yet controlled studies also show that length alone can degrade task performance even with perfect retrieval, which limits how far retrieval-oriented extension improvements can be interpreted as full long-context competence [15].

Substantial engineering progress has been made in extending nominal context windows. Yet robust effective context remains mechanism- and benchmark-dependent, and extension methods alone do not explain why some models maintain performance at distance while others degrade. This motivates a framework that can characterize the internal geometric properties underlying effective context.

3.5 Research Gap and Thesis Positioning

Across the literature, behavioral benchmark work and mechanistic analysis are often developed separately: benchmark papers provide strong evidence of claimed-versus-effective context gaps, while mechanistic papers isolate specific failure pathways such as positional bias or fragile retrieval circuits [9, 10, 19, 28, 38].

As a result, no existing work provides a quantitative framework that traces behavioral long-context degradation to specific geometric properties of attention head representations and validates this connection across model families. Benchmark papers measure that models fail; mechanistic papers identify candidate failure pathways; but the link between internal geometry and behavioral ECL remains uncharacterized.

This thesis addresses this gap by developing a geometric framework that extracts positional bias structure and attention plasticity from post-RoPE query and key vectors, then correlates these mechanistic metrics with benchmark performance across 13 models from three families. The framework is observational—it measures associations between internal geometry and behavioral performance rather than establishing causality through intervention—but to our knowledge it provides the first systematic connection between per-head attention geometry and cross-model long-context capability.

3.6 Summary

Behavioral studies provide robust evidence that effective context can be far below claimed context and that estimates are sensitive to benchmark design. Mechanistic studies identify candidate causes, including position-related effects and retrieval-path fragility. Context-extension methods improve usable length in many settings but do not provide a mechanistic quantification framework for ECL [8–10, 19, 28, 31, 38].

The next chapter develops the geometric framework that connects these behavioral and mechanistic perspectives: it extracts positional bias and attention plasticity from internal representations and provides metrics that can be correlated with behavioral long-context performance.

Chapter 4

Methodology

4.1 Q/K Capture Protocol

The methodology operates on post-RoPE query and key vectors extracted from transformer attention heads during inference. This section defines the capture protocol; Sections 4.2–4.4 define the three analyses applied to the captured vectors.

4.1.1 Captured Representation

For each attention head, we collect the query vector q_t and key vector k_t at each sampled token position t. These are post-RoPE, post-projection vectors—the exact inputs to the dot-product attention score $s(q,k) = q^\top k$ [35]. Post-RoPE vectors carry both content information (from the token embedding and learned projections) and position information (from the rotary encoding) [34].

Analyzing pre-RoPE vectors would characterize the learned representations but miss the positional encoding that shapes attention patterns. Analyzing attention weights directly would lose the Q/K decomposition needed for the geometric analyses in Sections 4.3 and 4.4.

4.1.2 Input Data and Position Sampling

Vectors are captured during inference on 500 examples from LongBench-Pro containing samples with 128K+ token context [9]. Using realistic, long, diverse text—rather than synthetic probes or random tokens—ensures the captured vectors reflect the model's behavior on naturalistic inputs representative of its training distribution.

Uniform position bucketing divides the context window into fixed-width buckets of $B_{\min}$ tokens.

Within each bucket, tokens are sampled with probability $p_{\text{keep}} = 1/B_{\min}$, yielding in expectation one sample per bucket per sequence. This ensures uniform coverage across the full context window, avoiding the front-loading that would occur under uniform token sampling, since most documents concentrate tokens in earlier positions. The bucket structure also provides a natural granularity for the per-position analyses in Section 4.4.

4.1.3 Head Sampling

For each model, 300 query heads are sampled uniformly at random across all eligible layers. In grouped-query attention (GQA) models [2], multiple query heads share a single key head. The corresponding key head for each sampled query head is determined by the GQA mapping:

$$h_k = \left\lfloor h_q \,/\, (n_q / n_{kv}) \right\rfloor \tag{4.1}$$

where h_q is the query head index and n_q, n_{kv} are the number of query and key-value heads per layer. Each (query head, key head) pair is analyzed independently—the shared key is broadcast to each query in the group, so each pair produces its own geometric structure.

Only full-attention heads are captured; sliding-window layers (if present) are excluded. Specific per-model and per-family configuration details are described in Section 5.4.

4.2 PCA Decomposition

The first analysis applies principal component analysis (PCA) to the captured Q/K vectors. PCA finds the directions of maximum variance without assumptions about what those directions represent, providing a model-agnostic structural fingerprint of each attention head.

4.2.1 Procedure

For each attention head, we concatenate all captured query vectors and key vectors into a single point cloud. Each vector $x_i \in \mathbb{R}^d$ is annotated with its type (Q or K) and its token position t_i. This combined pool lets PCA discover directions that separate Q from K as well as directions that encode position—both contribute to variance in the attention head's representation space.

For each principal component PC_j with eigenvector v_j, we compute the Pearson correlation between the projection $x_i^\top v_j$ and the token position t_i, separately for query vectors ($r_q^{(j)}$) and key vectors ($r_k^{(j)}$). We sign-canonicalize so the dominant correlation (the one with larger absolute value) is positive.

4.2.2 Head Taxonomy

Based on the correlations $(r_q^{(0)}, r_k^{(0)})$ on the first principal component, we classify heads into categories using thresholds of 0.3 (low) and 0.7 (high):

- **Position-dominated**: both $|r_q^{(0)}|$ and $|r_k^{(0)}|$ exceed 0.7 (position is the primary variance driver for both Q and K).
- **Q-positional**: $|r_q^{(0)}|$ exceeds 0.7 but $|r_k^{(0)}|$ does not (the Q cluster encodes position more than K on PC0).
- **Content-focused**: both below 0.3 (position is not the dominant variance direction).
- **Mixed**: all remaining heads.

This taxonomy is descriptive—it summarizes what PCA reveals about variance structure. It is not a causal classification: a head classified as "position-dominated" may still perform content-dependent retrieval if the content signal, though lower-variance, is sufficient to determine key preference in the dot product.

4.2.3 Limitations and the Q/K Confound

PCA maximizes total variance, which on the first component typically captures two confounded sources: the Q/K cluster offset (queries and keys occupy different regions of embedding space) and the position gradient (vectors at different positions spread along a consistent direction).

This confound inflates $r_q^{(0)}$ relative to $r_k^{(0)}$. Because the Q cluster centroid happens to align with the position gradient direction, the query projection onto PC0 shows a stronger position correlation than the key projection—even though keys are the vectors whose positional encoding most directly enters the attention score via RoPE [34].

PCA can discover that position structure exists and is pervasive, can reveal that Q/K separability and positional encoding co-occur on the same principal component, and can provide cross-model structural comparisons. However, PCA cannot isolate position from Q/K identity, cannot provide a parametric bias term, and cannot determine whether position dominates content in the attention score. These limitations motivate the targeted rotation analysis in the following section.

4.3 Planar Rotation Model

To resolve the confound identified in Section 4.2.3, we construct a 2D plane in each head's embedding space with axes that have guaranteed semantic meaning: one captures all linear position covariance,

the other captures maximum Q/K separation. The mechanistically important axis is the first (yielding a parametric bias term); the second serves a bookkeeping role, removing the Q/K centroid offset so that the remaining $(d-2)$-dimensional complement cleanly contains content signal and symmetric RoPE structure.

4.3.1 Axis a: Drift Direction

The drift direction a is the unit vector in $\mathbb{R}^d$ that maximizes linear covariance between its projection and token position:

$$a = \arg\max_{\|u\|=1} \mathrm{Cov}(u^\top x,\, t) \tag{4.2}$$

where x ranges over the combined pool of Q and K vectors and t is the corresponding token position. The closed-form solution is:

$$a = \frac{\mathrm{Cov}(x, t)}{\|\,\mathrm{Cov}(x, t)\,\|} \tag{4.3}$$

where $\mathrm{Cov}(x, t) \in \mathbb{R}^d$ is the vector of covariances between each embedding dimension and position.

By construction, *all* directions orthogonal to a have exactly zero linear position covariance with the combined Q+K pool. This makes a the unique direction where position "lives" linearly in the combined representation.

Because Q and K may encode position at different rates, using the combined pool captures the shared positional direction—the one that enters the dot product $q^\top k$. Separate Q-only or K-only drift directions would miss this shared structure.

Axis b: Separation Direction. The separation direction b is the unit vector orthogonal to a that maximizes the Q/K centroid separation:

$$b = \arg\max_{\substack{\|u\|=1 \\ u \perp a}} (\mu_Q - \mu_K)^\top u \tag{4.4}$$

where μ_Q and μ_K are the centroids of query and key vectors respectively. The solution projects the centroid difference onto the complement of a and normalizes:

$$b = \frac{(\mu_Q - \mu_K) - [(\mu_Q - \mu_K)^\top a]\, a}{\|(\mu_Q - \mu_K) - [(\mu_Q - \mu_K)^\top a]\, a\|} \tag{4.5}$$

4.3.2 The $\{a, b\}$ Plane

Together, a and b span a 2D plane in $\mathbb{R}^d$. Projecting all Q and K vectors onto this plane reveals two geometric roles:

- **Along a (horizontal)**: tokens at different positions spread out, creating a position gradient. The *key drift slope* α_K measures the linear regression coefficient of key projections onto a against token position. The *query centroid* μ_Q^a is the mean query projection onto a. This axis carries the mechanistically important asymmetric positional bias.

- **Along b (vertical)**: the Q and K clusters separate, with centroids at different heights. Since $b \perp a$, this axis has zero position covariance and does not contribute to positional bias. Its contribution to the attention score $q_b \cdot k_b$ is approximately constant across keys at different positions and therefore cancels in softmax. Axis b serves primarily to account for the Q/K centroid offset that confounds PCA (Section 4.2.3) and to enable 2D visualization of the head geometry.

4.3.3 Bias Strength

The positional contribution of axis a to the attention score is quantified by the *bias strength*:

$$\text{bias\_strength} = \mu_Q^a \times \alpha_K \tag{4.6}$$

This scalar measures how much the Q cluster's position on the drift axis amplifies the K position gradient. When bias_strength > 0, nearer keys (smaller position values) receive higher attention scores—a recency bias [26].

The derivation follows from the dot-product decomposition on axis a. The projection of the attention score onto the drift direction is $q_a \cdot k_a$, where $q_a = q^\top a$ and $k_a = k^\top a$. The mean key projection at position t_k is approximately $\bar{k}_a(t_k) \approx \alpha_K \cdot t_k + \text{intercept}$, and the mean query projection is approximately μ_Q^a. Therefore, the mean positional contribution to the score grows linearly with key position at rate $\mu_Q^a \times \alpha_K$.

Bias strength is one scalar per head. It captures a specific geometric mechanism—the interaction between the Q centroid position on the drift axis and the K position gradient along that axis—that produces a systematic preference for keys at certain absolute positions regardless of content.

Complement Space. The $(d-2)$-dimensional subspace orthogonal to $\{a, b\}$ contains two types of structure:

- **RoPE rotation planes**: pairs of dimensions where position encodes as a rotation angle [34]. In these planes, Q and K follow circular arcs parameterized by position. The symmetric rotation structure contributes *relative*-position information to $q^\top k$ but not the asymmetric absolute-position drift captured by axis a.

- **Content dimensions**: dimensions carrying semantic signal with zero positional encoding.

Axis a is where *asymmetric* positional bias lives—the mechanism that favors keys at certain absolute positions regardless of content. Axis b accounts for the Q/K centroid offset but does not contribute to key selection (Section 4.3.2). The complement is where content signal and *symmetric* relative-position structure reside. This separation is exact by construction.

4.3.4 Relationship to PCA

Both PCA and the rotation model are orthogonal transformations of the same Q/K vectors that preserve the dot product $q^\top k$ exactly. They optimize different objectives: PCA maximizes explained variance, while the rotation targets position covariance (axis a) and Q/K separation (axis b).

A key empirical reversal illustrates the difference: on PCA's first component, $|r_q^{(0)}| > |r_k^{(0)}|$ (queries appear more positional), whereas on axis a, $|r_k^{(a)}| > |r_q^{(a)}|$ (keys encode position more strongly). The reversal occurs because PC0 conflates the Q/K centroid offset with position encoding—the large Q-K cluster separation inflates the apparent query-position correlation. Axis a isolates position by construction, revealing that keys are the primary carriers of the position gradient, consistent with the role of RoPE in encoding key positions for attention scoring [34].

4.4 Attention Plasticity

The rotation model quantifies the bias mechanism—how much position enters the attention score and through what geometric pathway. But bias magnitude alone does not determine whether position dominates content. A head with large bias strength may still attend flexibly if the content signal is strong enough to overcome the positional preference. Attention plasticity measures the functional consequence: does position actually constrain which key wins?

Attention can be viewed as a content-based reranking mechanism: the dot-product scores $q^\top k_i$ induce a ranking over keys, and softmax converts this ranking into a weight distribution. Effective long-context processing requires that this ranking reflect content relevance rather than positional proximity. Any total ordering is fully determined by its pairwise comparisons—if position corrupts pairwise key orderings, the global ranking is necessarily corrupted. Pairwise comparison is therefore the natural unit for evaluating ranking quality.

Fix an attention head and a prefix $x_{1:n}$. The keys k_i for $i \le n$ are deterministic functions of the prefix—they are fixed once the context is given. Now consider a query at position $t > n$. The query q_t depends on future tokens beyond the prefix, drawn from the workload distribution $\mathscr{D}$. From the perspective of the prefix, q_t is a random variable.

This asymmetry is fundamental: keys carry the context to retrieve from; queries carry the question being asked. Plasticity measures whether the model's ranking of keys depends on the

question (content-driven) or is predetermined by position.

4.4.1 Definitions

Definition 4.1 (Pairwise preference)**.** For a fixed prefix $x_{1:n}$, a query position $t > n$, and a pair of key indices (i, j) with $i < j \leq n$, the *pairwise preference probability* is:

$$p_{n,t}(x_{1:n}, i, j) = \Pr_{q_t}\left[q_t^\top k_i > q_t^\top k_j\right] \tag{4.7}$$

where the probability is taken over the random query q_t.

When $p \approx 0.5$, the content of the query determines which key wins—different queries retrieve different information from the same context. When $p \approx 0$ or $p \approx 1$, the outcome is determined by the key positions, independent of the query.

Definition 4.2 (Pairwise plasticity)**.** For a fixed prefix and key pair, the *pairwise plasticity* is the scaled Bernoulli variance of the preference indicator:

$$\mathrm{PP}_{n,t}(x_{1:n}, i, j) = 4 \cdot p \cdot (1 - p) \tag{4.8}$$

where $p = p_{n,t}(x_{1:n}, i, j)$.

Pairwise plasticity achieves its maximum of 1 when $p = 0.5$ (the key ordering is maximally query-dependent) and approaches 0 when $p \to 0$ or $p \to 1$ (the ordering is rigid regardless of the query).

Definition 4.3 (Attention plasticity)**.** The *attention plasticity* at query position t is:

$$\mathrm{AP}_t = \mathbb{E}\left[\mathrm{PP}_{N,t}(X_{1:N}, I, J)\right] \tag{4.9}$$

where the expectation is over: (i) a random prefix length N uniform in $\{1, \ldots, t-1\}$, (ii) a random prefix $X_{1:N}$ from $\mathscr{D}$, and (iii) a random key pair (I, J) uniform over admissible pairs.

$\mathrm{AP}_t \in [0, 1]$ depends only on the query position t for a fixed attention head and workload. The function $t \mapsto \mathrm{AP}_t$ is the *plasticity profile* of the head.

4.4.2 Positional-Semantic Decomposition

To compute AP_t in closed form, we decompose query vectors using a Householder reflection that aligns the positional drift direction with the first coordinate axis.

Let $\beta \in \mathbb{R}^d$ be the vector of linear regression slopes from regressing query vectors onto their positions: $\beta_j = \mathrm{Cov}(q_j, t)/\mathrm{Var}(t)$. The Householder matrix

$$H = I - \frac{2uu^\top}{\|u\|^2}, \qquad u = \hat{\beta} - e_1, \qquad \hat{\beta} = \frac{\beta}{\|\beta\|} \tag{4.10}$$

maps $\hat{\beta}$ to the first standard basis vector e_1. Since H is orthogonal ($H^\top H = I$), applying it to both query and key vectors preserves all dot products: $q^\top k = (Hq)^\top (Hk)$.

In the rotated basis, the query decomposes as:

- **Coordinate 1 (positional)**: $q_1^{\mathrm{rot}} = \alpha_{\mathrm{pos}} + \beta_{\mathrm{pos}} \cdot t + \varepsilon$, where ε captures residual positional variation with variance σ^2_{pos}. All linear position covariance is concentrated in this coordinate.
- **Coordinates** $2, \ldots, d$ **(semantic)**: position-decorrelated components carrying content information. Their distribution depends on the query position bucket b: mean μ_b and covariance Σ_b.

This decomposition is distinct from the rotation in Section 4.3. The planar rotation constructs a $\{a, b\}$ plane from the combined Q+K pool for geometric characterization. The Householder reflection here uses query-only drift for the theoretical framework. Both are orthogonal transformations preserving $q^\top k$; they serve different purposes.

4.4.3 Gaussian Closed Form

Under the positional-semantic decomposition, the score difference $D = q_t^\top (k_i - k_j)$ for a key pair with difference vector $\delta = k_i^{\mathrm{rot}} - k_j^{\mathrm{rot}}$ in the rotated basis is approximately Gaussian. In practice, positions are bucketed and each bucket is represented by its midpoint τ_q (Section 4.4.5); the continuous-position variable t in the decomposition above is replaced by τ_q. The parameters are:

$$\mu = \delta_1 \cdot (\alpha_{\mathrm{pos}} + \beta_{\mathrm{pos}} \cdot \tau_q) + \delta_{2:d}^\top \mu_b \tag{4.11}$$

$$v = \delta_1^2 \cdot \sigma^2_{\mathrm{pos}} + \delta_{2:d}^\top \,\mathrm{diag}(\sigma_b^2)\, \delta_{2:d} \tag{4.12}$$

where τ_q is the query bucket position, α_{pos} and β_{pos} are the intercept and slope of the positional coordinate, σ^2_{pos} is the residual positional variance, and μ_b, σ_b^2 are the bucket-specific semantic mean and diagonal variance.

The key preference probability and pairwise plasticity then follow from the Gaussian CDF:

$$p = \Phi\left(\frac{\mu}{\sqrt{v}}\right) \tag{4.13}$$

$$\text{PP} = 4 \cdot \Phi\left(\frac{\mu}{\sqrt{v}}\right) \cdot \left(1 - \Phi\left(\frac{\mu}{\sqrt{v}}\right)\right) \tag{4.14}$$

where Φ is the standard normal CDF. The ratio $\mu/\sqrt{v}$ is the signal-to-noise ratio of the score difference. When keys differ in position ($\delta_1 \neq 0$) and the query position τ_q grows, the positional term in μ grows linearly (via $\beta_{\text{pos}} \cdot \tau_q$) while v remains bounded. This drives p away from 0.5 and plasticity toward zero.

4.4.4 Plasticity Decay

The Gaussian closed form reveals a fundamental asymmetry: the positional contribution to the mean μ grows linearly with query position τ_q, while the variance v is position-independent. As the query moves further into the context, the positional signal increasingly dominates the content signal.

Theorem 4.4 (Plasticity decay bound)**.** *Under the positional-semantic model with non-zero positional drift ($\beta_{pos} \neq 0$):*

$$AP_t \leq AP_\infty + C \cdot \exp(-c \cdot t^2) \tag{4.15}$$

for constants $C, c > 0$ depending on the model parameters. If all key pairs have distinct positional coordinates, $AP_t \to 0$ as $t \to \infty$.

In words: attention plasticity inevitably decays with query position whenever the query distribution exhibits linear positional drift—which all examined models do. The rate of decay depends on the model's specific bias strength and content signal strength, making the *shape* of the decay profile diagnostic of the model's long-context capability. The formal proof, which relies on sub-Gaussian concentration of the score difference, is deferred to Appendix A.

4.4.5 Bucketing and Aggregation

Per-query-position plasticity. The primary computation is per-query-bucket: for each query position bucket j with representative position τ_j, pool all eligible keys from earlier buckets, sample key pairs, compute the Gaussian closed form (Equations 4.11–4.14), and average. This produces AP as a function of query position—the plasticity profile.

Table 4.1: Summary of the three-analysis framework. Each analysis operates on the same captured Q/K vectors and answers a distinct question about the head's embedding geometry.

Analysis	Computes	Question	Key metric
PCA	Variance decomposition, position correlation	How is Q/K variance distributed?	r_q, r_k on PC0; head taxonomy
Rotation	Targeted axis projections, drift slopes	What is the parametric form of positional bias?	bias_strength $= \mu_Q^a \times \alpha_K$
Plasticity	Pr(random query flips key ordering)	Does positional bias dominate content?	$\mathrm{AP}_{\mathrm{drop}}$

2D heatmaps. Each sampled key pair also produces a 2D coordinate: the inter-key distance $|t_{k_i} - t_{k_j}|$ and the query-to-key-midpoint distance $\tau_q - \frac{1}{2}(t_{k_i} + t_{k_j})$. Binning plasticity by these two distances reveals structure invisible in the 1D position profile: keys close together but far from the query yield high plasticity (content determines the winner), while keys far apart with one near the query yield low plasticity (position dominates).

Aggregate metrics. We report four summary statistics per head, aggregated to model level:

- $\mathrm{AP}_{\mathrm{overall}}$: bucket-size-weighted mean across query position buckets.
- $\mathrm{AP}_{\mathrm{first\,20\%}}$: mean plasticity for query positions in the first 20% of context.
- $\mathrm{AP}_{\mathrm{last\,20\%}}$: mean plasticity for query positions in the last 20% of context.
- $\mathrm{AP}_{\mathrm{drop}} = \mathrm{AP}_{\mathrm{first\,20\%}} - \mathrm{AP}_{\mathrm{last\,20\%}}$: the plasticity degradation across context.

$\mathrm{AP}_{\mathrm{drop}}$ is the key diagnostic metric. A model with low $\mathrm{AP}_{\mathrm{drop}}$ maintains flexible attention at distant positions; a model with high $\mathrm{AP}_{\mathrm{drop}}$ loses the ability to perform content-driven retrieval as context length grows.

4.5 Connecting the Three Analyses

The three analyses are not successive levels of abstraction. They are three complementary questions about the same Q/K geometry, each providing information the others cannot. Table 4.1 summarizes the framework.

Two aspects of the inter-analysis relationships deserve emphasis.

Two related orthogonal transformations. The rotation analysis (Section 4.3) and the plasticity framework (Section 4.4) both use orthogonal transformations that preserve $q^\top k$, but they differ in construction and purpose. The rotation constructs axes from the combined Q+K pool—axis a for shared positional drift, axis b for Q/K separation—and provides geometric characterization of the bias mechanism. The plasticity framework uses a query-only Householder reflection—coordinate 1 for query positional drift—to enable the formal decay theorem and Gaussian closed form. On the drift axis, both transformations isolate position covariance, and the resulting projections are closely related. The rotation additionally constructs axis b (Q/K separation), which the plasticity framework does not need.

Logical progression. The analyses address progressively deeper questions. PCA discovers that position structure exists and is pervasive. The rotation isolates and parameterizes the bias mechanism, yielding bias strength as a scalar summary. Plasticity tests whether the bias functionally constrains attention, accounting for the content signal that may compensate.

When plasticity is low, the rotation analysis tells us *why* (large bias strength). When bias strength is large, plasticity tells us *whether it matters* (whether content can compensate). This complementarity motivates reporting all three analyses: a head may have large bias strength yet high plasticity (content dominates despite the bias), or small bias strength yet low plasticity (content signal is too weak to drive flexible retrieval even with minimal positional preference). The results in Chapter 6 demonstrate both patterns across the model families in our study.

Chapter 5

Experiments

5.1 Cross-Model Study

The cross-model study analyzes 13 transformer language models spanning three primary families and two predecessor models. The primary families—Ministral-3, Qwen-3, and Llama 3.2—provide within-family scaling comparisons (0.6B–14B parameters) and across-family architectural comparisons. Two predecessor models provide cross-generation comparisons and per-length RULER scores for direct plasticity-to-benchmark correlation.

5.1.1 Primary Model Families

Ministral-3. The Ministral-3 family comprises three models at 3B, 8B, and 14B parameters [25]. All three use full attention with RoPE positional encoding and a claimed context length of 256K tokens—the largest among our primary models. The family uses a uniform GQA configuration (32 query heads, 8 key-value heads) across all scales, varying only in depth and hidden dimension (Table 5.1).

Qwen-3. The Qwen-3 family spans five scales from 0.6B to 14B parameters [32], providing the widest parameter range in our study. All models use full attention with RoPE, extended to 128K context via YaRN scaling (factor 4.0, base context 32K) [31]. The GQA configuration varies across scales: the smallest models (0.6B, 1.7B) use 16 query heads, the mid-range models (4B, 8B) use 32, and the 14B model uses 40—all with 8 key-value heads. Qwen-3 additionally applies RMS normalization to query and key vectors before the rotary embedding [32].

Llama 3.2. The Llama 3.2 family includes three models at 1B, 3B, and 11B parameters [16]. All use full attention with RoPE and a claimed context of 128K tokens. The 1B model is architecturally distinctive, with a head dimension of 64 rather than the 128 shared by all other models in the study. The 11B model is a vision-language model whose text self-attention layers use standard RoPE, while its cross-attention layers (for vision-text fusion) have no positional encoding—a NoPE (no position encoding) pattern. Our analysis captures only the text self-attention path.

5.1.2 Predecessor Models

Two predecessor models complement the primary families by providing published RULER per-length scores at six context bins (4K–128K), enabling direct plasticity-to-benchmark correlation at length granularity.

Llama-3.1-8B. The previous-generation Llama with full 128K attention [16]. Its RULER scores show a gradual decline from 95.5 at 4K to 77.0 at 128K [10], providing a reference trajectory for models with graceful degradation.

Mistral-v0.2-7B. The predecessor of Ministral-3 with 32K sliding-window attention [27]. Its RULER scores show strong performance within the window (93.6 at 4K) but rapid decline beyond it (49.0 at 64K, 13.8 at 128K) [10]. Because the sliding-window mechanism limits the attention span, our analysis of this model is restricted to within-window positions; key pair sampling and query distributions are not directly comparable to the full-attention models.

5.1.3 Selection Criteria

Three criteria governed model selection. First, all models have open weights, ensuring reproducibility. Second, all primary models support 128K or longer claimed context, placing them in the regime where effective context length diverges from claimed length (Chapter 3). Third, the primary families overlap with LongBench-Pro's evaluated model set (7 of 11 primary models), while the predecessor models overlap with RULER's published per-length scores (Section 5.3).

All models are analyzed in their base (non-instruct) form to measure the geometry established during pre-training and context extension, prior to any instruction-tuning modifications. Benchmark scores are taken from instruct variants of the same architectures, as benchmarks are typically evaluated on instruct-tuned models. This introduces a confound: instruction tuning could alter the attention geometry we measure. We discuss this limitation in Section 7.5.

Table 5.1: Model configurations for the cross-model study. All models use grouped-query attention with 8 key-value heads. Context length denotes the claimed maximum; effective context may differ (Chapter 3). Configurations are from HuggingFace model repositories.

Model	Params	Layers	Q heads	d_h	Context	Attention	Pos. enc.
Ministral-3							
3B	3B	26	32	128	256K	Full	RoPE
8B	8B	34	32	128	256K	Full	RoPE
14B	14B	40	32	128	256K	Full	RoPE
Qwen-3							
0.6B	0.6B	28	16	128	128K[†]	Full	RoPE + YaRN
1.7B	1.7B	28	16	128	128K[†]	Full	RoPE + YaRN
4B	4B	36	32	128	128K[†]	Full	RoPE + YaRN
8B	8B	36	32	128	128K[†]	Full	RoPE + YaRN
14B	14B	40	40	128	128K[†]	Full	RoPE + YaRN
Llama 3.2							
1B	1B	16	32	64	128K	Full	RoPE
3B	3B	28	24	128	128K	Full	RoPE
11B-Vision	11B	40	32	128	128K	Full[‡]	RoPE + NoPE
Predecessor models							
Llama-3.1-8B	8B	32	32	128	128K	Full	RoPE
Mistral-v0.2-7B	7B	32	32	128	32K	Sliding	RoPE

[†] Via YaRN scaling (factor 4.0, base 32K) [31]. [‡] Vision-language model; text self-attention layers use full attention with RoPE, cross-attention layers have no positional encoding.

5.1.4 Model Configurations

Table 5.1 summarizes the architectural configurations. All models use grouped-query attention [2] with 8 key-value heads. Head dimension is uniformly 128 across models, with the exception of Llama-3.2-1B (64). The three families differ primarily in their positional encoding strategy: Ministral-3 uses standard RoPE, Qwen-3 extends RoPE via YaRN, and Llama 3.2 uses standard RoPE with the 11B variant additionally featuring NoPE cross-attention layers.

5.2 Training Dynamics Study

While the cross-model study provides a snapshot of Q/K geometry across different architectures and scales, it cannot reveal how this geometry develops during training. The training dynamics study traces a single model—SmolLM3-3B—across 10 checkpoints spanning pre-training through long-context extension [5].

5.2.1 Why SmolLM3

Open intermediate checkpoints are uncommon: most labs release only final weights. SmolLM3-3B provides several properties that make it uniquely suited for a training dynamics study:

- **Open checkpoints.** Ten checkpoints are publicly available across three pre-training stages and two long-context extension phases, stored as separate branches in the HuggingFace repository.
- **Clear phase transitions.** Training progresses from a 4K context window through 32K to 64K, with corresponding RoPE base frequency (θ) increases from 50K to 2M to 5M. Each transition provides a natural experiment for observing how context extension reshapes the Q/K geometry.
- **NoPE layer pattern.** Every fourth layer (9 of 36) uses no positional encoding, while the remaining 27 layers apply RoPE. This architectural feature enables direct comparison of plasticity in position-aware versus position-agnostic layers within the same model.
- **Comparable scale.** At 3B parameters, SmolLM3 is scale-matched to Ministral-3-3B and Llama-3.2-3B from the cross-model study (though architecturally distinct), enabling cross-study comparison at matched parameter count.

5.2.2 Checkpoint Selection

Table 5.2 lists the 10 selected checkpoints. The first six span pre-training stages 1–3, all with a 4K context window and $\theta = 50{,}000$. Stage 1 is sampled at four points (early, ~1/3, ~2/3, and end) to capture the main pre-training trajectory. Stages 2 and 3 are represented by their final checkpoints, capturing the effects of data mix changes and annealing respectively.

The remaining four checkpoints cover long-context extension in two phases: 4K→32K and 32K→64K. Each phase is sampled at onset (4K additional steps) and convergence (20K additional steps), capturing both the initial disruption and eventual adaptation of the Q/K geometry under context scaling.

5.2.3 Architecture and Capture Differences

SmolLM3-3B uses 36 layers with grouped-query attention (16 query heads, 4 key-value heads) and a head dimension of 128 [5]. The NoPE pattern assigns no positional encoding to layers 3, 7, 11, 15, 19, 23, 27, 31, and 35 (0-indexed); the remaining 27 layers apply RoPE. All layers use full attention—no sliding window is present in any checkpoint configuration.

Table 5.2: SmolLM3-3B training checkpoints. Pre-training stages 1–3 share a 4K context window; long-context (LC) extension progressively increases context length and RoPE base frequency θ. Step counts for LC phases indicate additional steps within that phase.

#	Phase	Steps	Context	RoPE θ
1	Pre-train stage 1 (early)	40K	4K	50K
2	Pre-train stage 1 (~1/3)	1.2M	4K	50K
3	Pre-train stage 1 (~2/3)	2.4M	4K	50K
4	Pre-train stage 1 (end)	3.44M	4K	50K
5	Pre-train stage 2	4.2M	4K	50K
6	Pre-train stage 3 (annealing)	4.72M	4K	50K
7	LC 4K→32K (onset)	+4K	32K	2M
8	LC 4K→32K (converged)	+20K	32K	2M
9	LC 32K→64K (onset)	+4K	64K	5M
10	LC 32K→64K (final)	+20K	64K	5M

The SmolLM3 checkpoints require several configuration adjustments relative to the cross-model captures. The tokenizer maximum length matches each checkpoint's context window (4K, 32K, or 64K rather than 128K), and the position-sampling bucket size scales proportionally (256, 2,048, or 4,096 rather than 8,192), maintaining a consistent 16-bucket granularity across all stages. The tokenizer is loaded from the main SmolLM3-3B repository rather than the checkpoint branches, which do not include tokenizer files.

5.3 Benchmarks

Two complementary benchmarks validate the mechanistic metrics against behavioral performance. LongBench-Pro provides realistic task evaluation across 7 of our primary models; RULER provides synthetic per-length evaluation for 2 predecessor models.

5.3.1 LongBench-Pro (Primary)

LongBench-Pro (LBP) is a large-scale long-context evaluation benchmark comprising 1,500 samples across 25 tasks in a bilingual (English/Chinese) format [9]. Inputs range from 8K to 256K tokens, and all tasks use a multiple-choice format for consistent automated scoring. Samples are binned by input length (8K, 16K, 32K, 64K, 128K, 256K), providing length-stratified evaluation.

Seven of our 11 primary models have exact matches in the LBP evaluation: Ministral-3 at all three scales (3B, 8B, 14B), Qwen-3 at three scales (4B, 8B, 14B), and Llama-3.2-3B. We use non-thinking mode scores, which correspond to standard autoregressive generation without chain-of-thought prompting—matching the inference mode of our Q/K capture on base models.

LBP is our primary benchmark because its tasks require content-dependent reasoning rather than simple retrieval. Since our plasticity metric measures the competition between content-driven and position-driven attention patterns, a benchmark that rewards content comprehension provides the most appropriate validation target.

5.3.2 RULER (Secondary)

RULER is a synthetic benchmark comprising 13 tasks—including NIAH variants, multi-key/value/query retrieval, variable tracking, aggregation, and question answering—evaluated at context lengths from 4K to 128K [10]. Unlike LBP's realistic document tasks, RULER's synthetic construction provides precise control over what information must be retrieved from which positions.

Two of our models have published RULER per-length scores: Llama-3.1-8B and Mistral-v0.2-7B [10]. Llama-3.1-8B shows a gradual decline (95.5 at 4K to 77.0 at 128K), providing per-length granularity for plasticity-to-benchmark correlation. Mistral-v0.2-7B scores are available but its sliding-window architecture limits plasticity analysis to within-window positions (Section 5.1), so it serves primarily as a within-window reference for the Ministral-3 family comparison.

5.3.3 Complementary Roles

The two benchmarks exercise different aspects of long-context capability. LBP tests realistic content comprehension at an aggregate level across 7 models, addressing the question: does plasticity predict performance on tasks that require understanding long documents? RULER tests controlled retrieval at per-length granularity across 2 models, addressing the question: does plasticity track the length-dependent degradation curve?

If plasticity correlates with performance on both realistic and synthetic benchmarks, this provides convergent evidence that the metric captures a genuine mechanistic property relevant to effective context length.

5.4 Implementation

5.4.1 Capture Protocol

All Q/K captures were performed on a single NVIDIA B200 GPU (192 GB VRAM, bfloat16 precision) rented through Vast.ai. The capture protocol is uniform across all cross-model configurations. Input text is drawn from a 500-example subset of LongBench-Pro containing samples with 128K+ token context [9]. Models process inputs in bfloat16 precision with a maximum sequence length of 131,072

tokens (128K). Token positions are sampled using uniform bucket sampling with a minimum bucket size of 8,192, yielding 16 position buckets across the full context window.

For each model, 300 query heads are sampled uniformly at random (seed 0) from the set of all query heads across eligible layers. If a model has fewer than 300 total query heads, all are captured. Key heads are derived from selected query heads via the GQA mapping: $h_k = \lfloor h_q / (n_q / n_{kv}) \rfloor$, where n_q and n_{kv} are the number of query and key-value heads per layer. Only full-attention layers are captured; sliding-window layers (if present) are excluded.

All vectors are captured post-RoPE—after the rotary positional embedding has been applied to the query and key projections. For NoPE layers (SmolLM3 and Llama-3.2-11B cross-attention layers), the captured vectors are the raw linear projections, since no positional encoding is applied.

5.4.2 Family-Specific Notes

Three model families require specific handling during capture:

- **Qwen-3.** The YaRN RoPE scaling configuration is applied via config override (factor 4.0, original context 32K) [31] to enable 128K inference. Query and key vectors are RMS-normalized per head before the rotary embedding is applied [32].
- **Llama-3.2-11B-Vision.** As a vision-language model, Q/K vectors are captured from the text self-attention path only, excluding vision encoder and cross-attention components.
- **SmolLM3.** NoPE layers (every 4th layer) skip the rotary embedding step entirely; the post-RoPE capture point for these layers yields raw projected vectors identical to pre-RoPE.

Three analyses—PCA structure, rotation, and plasticity—are applied to the same set of captured Q/K vectors. Each produces per-head metrics that are aggregated into layer-level and model-level summaries. The methodology for each analysis is described in Chapter 4.

Chapter 6

Results

6.1 Position Bias Geometry

This section addresses RQ1: *how does position information manifest in the geometry of attention heads?* PCA reveals that position is the dominant source of Q/K variance. The rotation model then corrects a confound in PCA's axes and isolates the parametric form of positional bias.

6.1.1 PCA Reveals Positional Dominance

Across all 11 primary models, the first principal component of the combined Q+K point cloud captures ~34% of total variance (PC1 captures ~8%; subsequent components decline further). Of total Q+K variance, 23–32% in queries and 9–20% in keys is linear in token position (varying by family), indicating that positional encoding is the single largest structural feature in these representations.

On PC0, queries show stronger position correlation than keys ($|r_q^{(0)}| \approx 0.80$ vs $|r_k^{(0)}| \approx 0.49$). This apparent asymmetry—suggesting queries encode position more than keys—will be corrected in the next subsection.

The head taxonomy based on PC0 correlations classifies 26.6% of heads as position-dominated, 14.2% as Q-positional, 3.8% as content-focused, and 55.4% as mixed. Most heads carry some positional structure; purely content-focused heads are rare. Families cluster distinctly in (r_q, r_k) space (Figure A.1 in Appendix A): Llama heads concentrate in the top-right (both correlations high), small Qwen models appear Q-positional, and Ministral heads occupy a moderate region. This clustering suggests that positional bias geometry is family-determined rather than scale-determined.

PCA's dominant component conflates two sources of structure: positional encoding and Q/K identity separation. The high r_q on PC0 partly reflects the Q cluster centroid aligning with the

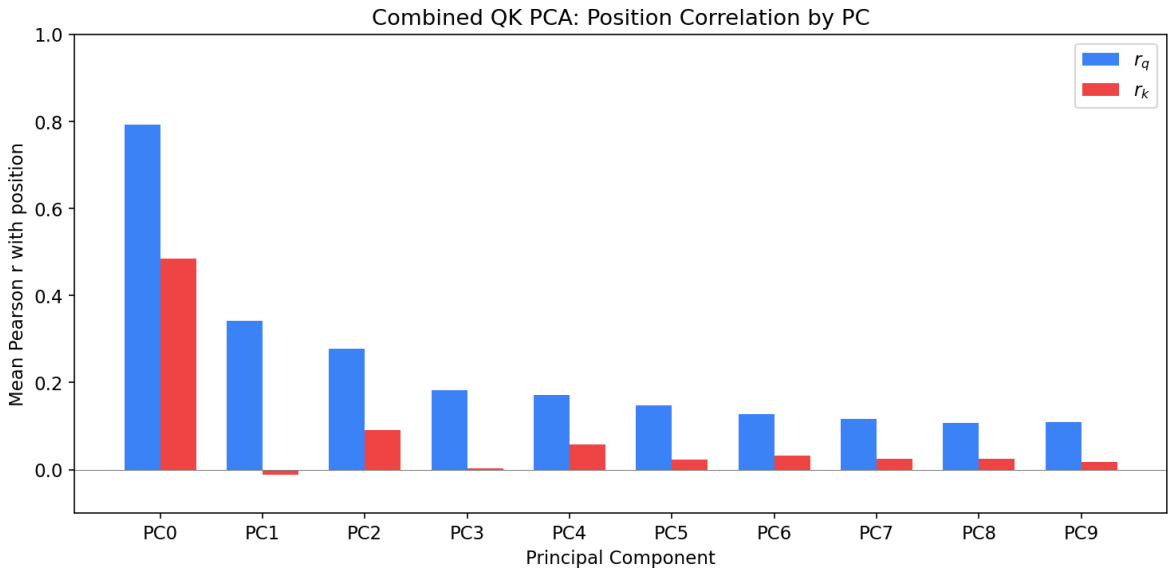


Figure 6.1: Position correlation on principal components across 11 models. PC0 carries the dominant positional signal (23–32% of query variance and 9–20% of key variance is linear in position), while subsequent components contribute minimal position information.

position gradient direction, not a genuine asymmetry in position encoding strength (Section 4.2.3). To disentangle these, we apply the rotation model with axes targeted at specific geometric quantities.

6.1.2 Rotation Isolates the Bias Mechanism

The drift axis a is constructed to maximize position covariance (Section 4.3), so high position correlation on this axis is by design. The non-trivial finding is how the signal distributes between Q and K: the PCA asymmetry reverses, with $|r_k^{(a)}| \approx 0.87 > |r_q^{(a)}| \approx 0.74$. Keys encode position more strongly than queries on the mechanistically relevant axis. This is a substantive correction: it changes which vector carries the position signal. Since RoPE applies identical rotations to both Q and K [34], the asymmetry is not prescribed by the encoding scheme but is a learned property of the model weights.

The $\{a, b\}$ plane captures 30–43% of variance across models (comparable to PCA's top two components) but with guaranteed semantic meaning: axis a carries all linear position covariance, axis b accounts for the Q/K centroid offset.

Positional bias takes a precise parametric form: bias_strength = $\mu_Q^a \times \alpha_K$ (Equation 4.6). This scalar measures how the query cluster's position on the drift axis amplifies the key position gradient. Across all 3,239 analyzed heads, 99.0% show positive bias strength—a near-universal recency bias favoring keys at nearer positions across the examined models. Only 31 heads exhibit primacy bias

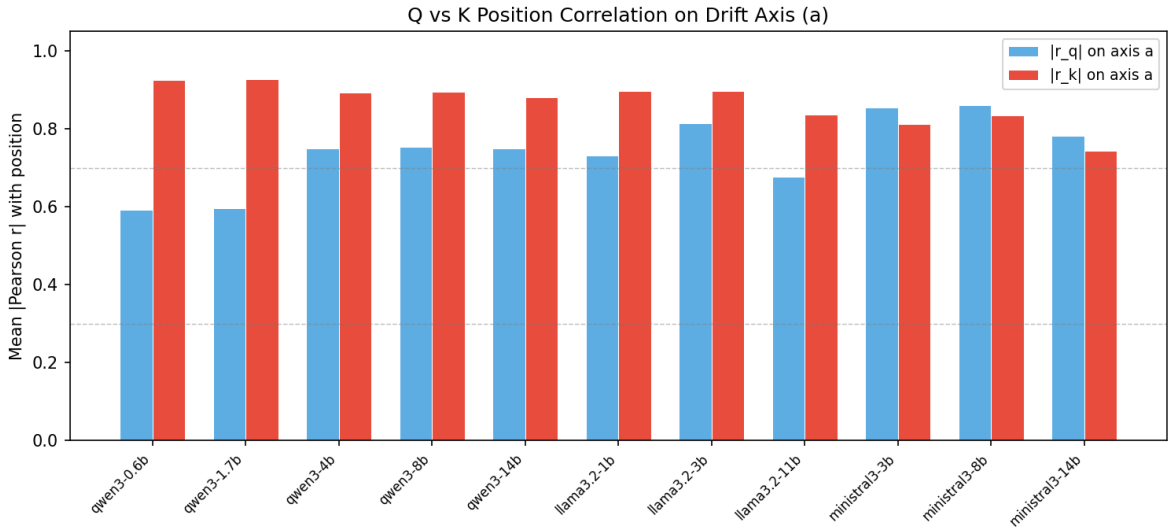


Figure 6.2: Position correlation on the drift axis a across 11 models. Keys (red) consistently show stronger position correlation than queries (blue)—the reverse of the PCA finding in Figure 6.1. The reversal occurs because PCA's first component conflates Q/K identity with positional encoding.

(Figure A.2 in Appendix A). Mean bias strength is 3.0×10^{-4} for all three Ministral-3 scales, 0.8–1.0×10^{-3} for Qwen-3, and 6.0–7.0×10^{-4} for Llama-3.2—tight within families, confirming that the mechanism is architecture-determined.

A Simpson's paradox appears in the relationship between bias strength and Q/K separation strength: the overall correlation is +0.48, but within the Llama family it reverses to −0.65 (Figure A.3 in Appendix A). Families have genuinely different bias-separation trade-offs that are masked by pooling across architectures.

Rotation quantifies the bias mechanism—how much positional preference enters each head's attention score. But a large bias does not necessarily constrain attention: if the content signal is strong, the query can still select semantically relevant keys regardless of position. The next section tests whether positional bias functionally dominates content.

6.2 Plasticity Profiles

This section addresses RQ2: *does positional bias functionally constrain which keys the model attends to?* Attention plasticity (Section 4.4) measures whether the query content determines key selection or whether position dominates. All models show plasticity decline with context position, but the *rate* of decline separates model families.

Table 6.1: Quintile plasticity profiles for 11 primary models. Each column shows mean attention plasticity for query positions in that quintile of the context window. All models decline from left to right; the rate of decline varies by family.

Model	0–20%	20–40%	40–60%	60–80%	80–100%	AP_{drop}
Ministral-3						
3B	0.664	0.633	0.615	0.606	0.592	0.072
8B	0.654	0.624	0.602	0.588	0.571	0.083
14B	0.655	0.620	0.607	0.602	0.587	0.068
Qwen-3						
0.6B	0.695	0.628	0.592	0.567	0.526	0.169
1.7B	0.702	0.640	0.603	0.579	0.540	0.162
4B	0.699	0.628	0.581	0.540	0.512	0.187
8B	0.694	0.630	0.585	0.545	0.517	0.177
14B	0.680	0.615	0.576	0.548	0.519	0.161
Llama-3.2						
1B	0.703	0.653	0.631	0.527	0.487	0.216
3B	0.686	0.621	0.561	0.489	0.456	0.230
11B	0.658	0.587	0.534	0.557	0.489	0.169

6.2.1 Plasticity Declines with Position

Table 6.1 reports the quintile plasticity profile for all 11 primary models. Every model shows declining plasticity across context, but the profiles differ qualitatively.

Three patterns emerge. Ministral-3 models show gradual, near-linear decline with the flattest profiles of any family ($AP_{drop} \approx 0.07$–0.08). Qwen-3 models decline more steeply ($AP_{drop} \approx 0.16$–0.19), with acceleration in the second half of context; larger models are slightly flatter. Llama-3.2 models show the steepest decline overall ($AP_{drop} \approx 0.17$–0.23), with Llama-3.2-3B dropping from 0.686 to 0.456—the largest degradation of any model.

Llama-3.2-11B exhibits an anomalous non-monotone profile: plasticity dips at 40–60%, partially recovers at 60–80%, then collapses at 80–100%. Llama-3.1-8B (a predecessor model) shows the same pattern, consistent with a Llama-family trait at 32+ layers rather than an artifact of the vision architecture.

Aggregate plasticity ($AP_{overall}$) does *not* predict performance across families. Ministral-3B has the highest aggregate plasticity (0.622) among models with LongBench-Pro scores, yet scores only 30.18. The aggregate confounds attention mechanics with base model capability. The informative metrics are positional: AP_{drop} and $AP_{last\,20\%}$.

6.2.2 2D Geometry: Plasticity Depends on Two Distances

The 1D position profiles in Table 6.1 collapse essential structure. The 2D bucket heatmaps (Figure 6.3) reveal that plasticity is a function of *both* inter-key distance (how far apart the two competing keys are) and key-to-query distance (how far the keys are from the query).

The geometry of the competition follows a consistent pattern: keys close together but far from the query yield high plasticity (content determines the winner), while keys far apart with one near the query yield low plasticity (position dominates). The transition between these regimes is where families diverge.

Ministral-3 is strikingly uniform: the warm region (> 0.55) fills nearly the entire triangle, with the dark corner confined to extreme inter-key distances (> 90K). The 3B, 8B, and 14B heatmaps are nearly identical in structure. Qwen-3 shows a clear diagonal gradient from warm top-left to dark bottom-right, with the dark region starting at ~50K inter-key distance. Remarkably, the pattern is nearly identical across scales from 0.6B to 14B—the architecture determines the geometry; scale adjusts the level. Llama-3.2-3B shows the most extreme contrast of any model, with the dark zone starting earlier (~40K) and reaching below 0.2. Llama-3.2-11B is uniquely non-smooth, with dark patches at specific distance combinations and a bright recovery zone around 57K–65K query-to-key distance.

6.2.3 Head Heterogeneity

Per-head plasticity profiles (Figure 6.4) reveal different architectural strategies for distributing content-based and position-based attention across heads.

Ministral-3-3B produces a tight bundle with low inter-head variance: nearly all heads decline gently and uniformly. The architecture produces homogeneous attention behavior where every head contributes similarly to content-based retrieval. Qwen-3-4B shows a bimodal distribution: content-specialized heads near 0.8–1.0 coexist with position-locked heads near 0.2–0.3. The model achieves content retrieval through heterogeneous specialization. Llama-3.2-3B demonstrates the steepest decline. Position bias eventually dominates every head, even initially plastic ones.

6.3 Training Dynamics

The cross-model results provide a static picture of how different architectures shape attention geometry. To understand how this geometry develops, we trace SmolLM3-3B across 10 training checkpoints spanning pre-training and long-context extension (Section 5.2). This section integrates all three analyses along the temporal axis.

Table 6.2: Joint rotation and plasticity metrics across SmolLM3-3B training checkpoints. Bias strength doubles during pre-training then collapses 10× during long-context extension, while position correlation ($|r_k^{(a)}|$) remains high throughout.

Phase	Checkpoint	bias_str	$\|\alpha_K\|$	$\|r_k^{(a)}\|$	$\text{AP}_{\text{overall}}$	$\text{AP}_{\text{first 20\%}}$	AP_{drop}
Pre-train	stg1 40K	0.0092	3.19e-3	0.861	0.585	0.609	0.040
Pre-train	stg1 1.2M	0.0160	3.23e-3	0.886	0.542	0.579	0.062
Pre-train	stg1 2.4M	0.0165	3.32e-3	0.893	0.545	0.582	0.061
Pre-train	stg1 3.44M	0.0166	3.34e-3	0.894	0.545	0.579	0.057
Pre-train	stg2 4.2M	0.0168	3.33e-3	0.893	0.548	0.585	0.062
Anneal	stg3 4.72M	0.0194	4.08e-3	0.918	0.516	0.552	0.058
LC 4K→32K	step 4K	0.0032	5.01e-4	0.911	0.507	0.585	0.153
LC 4K→32K	step 20K	0.0032	5.07e-4	0.913	0.507	0.582	0.147
LC 32K→64K	step 4K	0.0018	2.68e-4	0.906	0.502	0.590	0.168
LC 32K→64K	step 20K	0.0018	2.69e-4	0.905	0.503	0.588	0.161

6.3.1 Position Structure Emerges Early

PCA reveals that positional structure develops rapidly during early pre-training. The query position correlation $|r_q^{(0)}|$ on PC0 jumps from 0.67 at step 40K to 0.82 by step 1.2M. Key correlation increases more gradually. Head taxonomy evolves in parallel: position-dominated heads grow from 34% to 55% within stage 1, while content-focused heads virtually disappear (Figures A.4 and A.5 in Appendix A).

By step 1.2M, the structural fingerprint—which heads are positional, which are mixed—is largely set. Subsequent pre-training and the annealing phase refine but do not restructure the head roles.

6.3.2 Bias Grows During Pre-Training, Collapses During LC Extension

The rotation analysis tracks the parametric evolution of positional bias across training (Table 6.2).

During pre-training, bias strength doubles from 0.009 to 0.019, driven by growth in the key drift slope α_K. The model progressively learns to encode position in key vectors.

At long-context extension onset (stage 3 → LC 4K→32K), bias strength drops 6× within the first 4K training steps. It drops a further 1.8× during the 32K→64K phase, for a total ~10× reduction. The collapse mechanism is specific: α_K drops 93% (the key position gradient flattens), while $|r_k^{(a)}|$ holds above 0.90 (keys still encode position with high fidelity, but the encoding becomes more uniform across positions). Simultaneously, Q/K separation strength increases 14%—the representations become more geometrically distinct, not less.

6.3.3 Bias Collapse Is Necessary but Not Sufficient

The plasticity columns of Table 6.2 reveal a dissociation between bias reduction and long-context flexibility.

Short-context plasticity ($AP_{\text{first }20\%}$) recovers from 0.552 (end of annealing) to 0.588 during LC extension—back to early stage 1 levels. Bias reduction works locally: reducing the positional term in the score difference restores content-driven key selection at nearby positions.

However, long-context plasticity ($AP_{\text{last }20\%}$) drops to 0.427—far below the 0.569 the model achieved at 3.5K–4K positions during pre-training at comparable bias levels. Despite a 10× bias collapse, AP_{drop} *triples* from ~0.06 to ~0.16. The plasticity gradient steepens as the context window grows.

The excess AP_{drop} after bias collapse ($\sim 0.16 - 0.06 = 0.10$) reflects content signal decay at distance: after positional bias is nearly zeroed, content variance in the complement subspace loses strength at distant positions, so even minimal residual bias dominates. Per-position plasticity values during the LC extension phases (Table A.1 in Appendix A) confirm that short-context plasticity holds steady while distant positions show progressively lower values as the context window expands. The implications of this finding are discussed in Chapter 7.

6.4 Benchmark Validation

This section addresses RQ3: *do plasticity profiles correspond to behavioral long-context performance?* We correlate plasticity metrics with LongBench-Pro aggregate scores across 7 matched models and with per-length data where available.

6.4.1 Plasticity Drop Predicts LongBench-Pro Ordering

Table 6.3 presents the 7 models with both mechanistic metrics and LongBench-Pro scores. The key finding is that AP_{drop}—the plasticity degradation from early to late context—separates model families in the same order as benchmark performance.

Across families, AP_{drop} separates Ministral (~0.07) from Qwen (~0.17) from Llama (0.23), matching the LBP ordering. Within the Qwen family, the relationship is monotonic: 14B ($AP_{\text{drop}} = 0.161$, LBP 37.1) outperforms 8B (0.177, 33.4) outperforms 4B (0.187, 31.3)—larger models degrade less and score higher.

$AP_{\text{last }20\%}$ also separates models: Llama-3.2-3B at 0.456 (last) maps to LBP 15.71 (last), while Ministral models maintain 0.57–0.59 at distant positions, corresponding to LBP scores of 30–40.

Table 6.3: Mechanistic metrics and LongBench-Pro scores for 7 matched models, sorted by LBP score. AP_{drop} separates families: Ministral (~0.07), Qwen (~0.17), Llama (0.23), matching the LBP ordering. Non-thinking mode scores used throughout.

Model	Family	LBP	$AP_{overall}$	$AP_{first\,20\%}$	$AP_{last\,20\%}$	AP_{drop}
Ministral-3-14B	Ministral	40.14	0.615	0.655	0.587	0.068
Ministral-3-8B	Ministral	37.80	0.608	0.654	0.571	0.083
Qwen-3-14B	Qwen	37.11	0.590	0.680	0.519	0.161
Qwen-3-8B	Qwen	33.41	0.597	0.694	0.517	0.177
Qwen-3-4B	Qwen	31.26	0.594	0.699	0.512	0.187
Ministral-3-3B	Ministral	30.18	0.622	0.664	0.592	0.072
Llama-3.2-3B	Llama	15.71	0.565	0.686	0.456	0.230

Table 6.4: Ministral-3-14B per-length LongBench-Pro scores and matched plasticity values. Both decline with length, though plasticity decline (12.9%) is shallower than LBP decline (18.4%), reflecting length-independent task difficulty.

	8K	16K	32K	64K	128K	256K
LBP score	51.88	48.52	48.75	45.70	42.36	37.59
Plasticity	0.684	0.660	0.639	0.615	0.596	—

6.4.2 Base Capability vs. Context Preservation

Ministral-3-3B is the diagnostic outlier in Figure 6.7: it has the lowest AP_{drop} (0.072, comparable to the 14B variant) and the highest aggregate plasticity (0.622) among all 7 models, yet scores only 30.18 on LBP. Its flat plasticity profile indicates strong context preservation, but the low benchmark score reflects limited base capability at 3B scale.

This dissociation clarifies what plasticity measures and what it does not. Benchmark performance depends on both base model capability (knowledge, reasoning ability) and context preservation (maintaining attention flexibility at distance). Plasticity metrics capture the second factor. AP_{drop} isolates context preservation by measuring the *slope* of plasticity decline, which is robust to the absolute level of model capability. Within a family (controlled base architecture), AP_{drop} and LBP move together. Across families, the base capability confound must be accounted for.

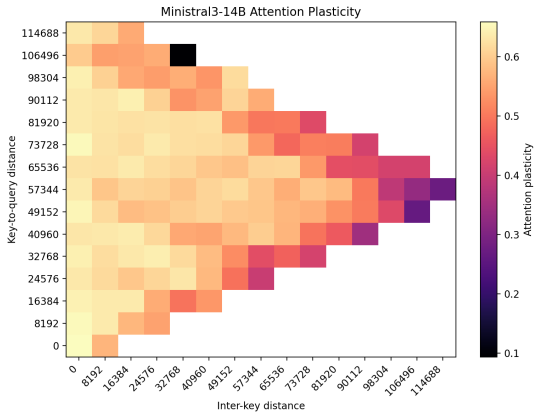
6.4.3 Per-Length Correspondence

For Ministral-3-14B—the one model with published per-length LBP scores [9]—plasticity at each length bin tracks benchmark performance (Table 6.4).

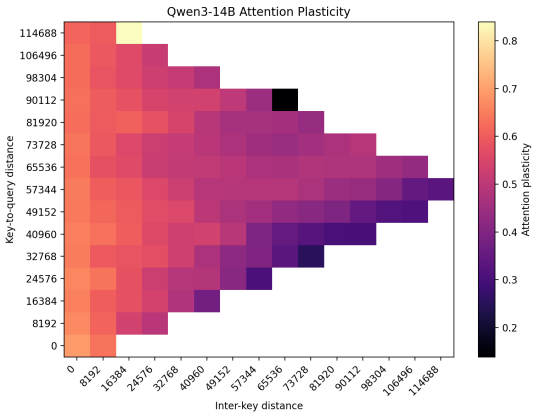
Both metrics decline with length, but plasticity decline (12.9% from 8K to 128K) is shallower than LBP decline (18.4%). The gap reflects length-independent task difficulty: some benchmark

tasks are harder at longer lengths for reasons unrelated to attention flexibility (e.g., more candidate answers, higher reasoning complexity).

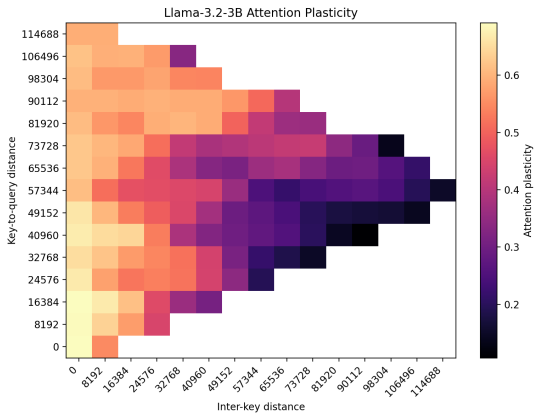
SmolLM3-3B at the end of long-context training ($AP_{drop} \approx 0.16$) matches Qwen-3 models (0.16–0.19)—both are standard LC-trained models of comparable scale. Ministral-3 achieves $AP_{drop} \approx 0.07$ at comparable bias levels over a longer context window, suggesting that its training recipe or architecture may address factors beyond positional bias reduction.



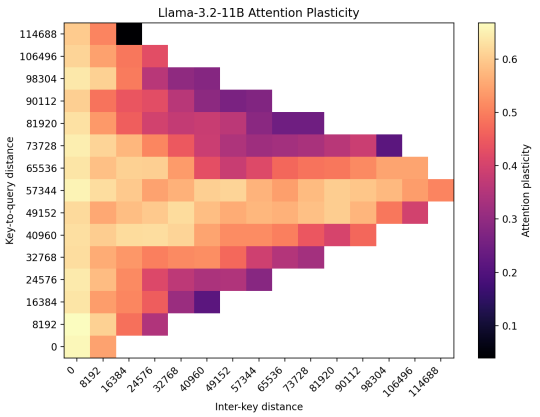
(a) Ministral-3-14B



(b) Qwen-3-14B

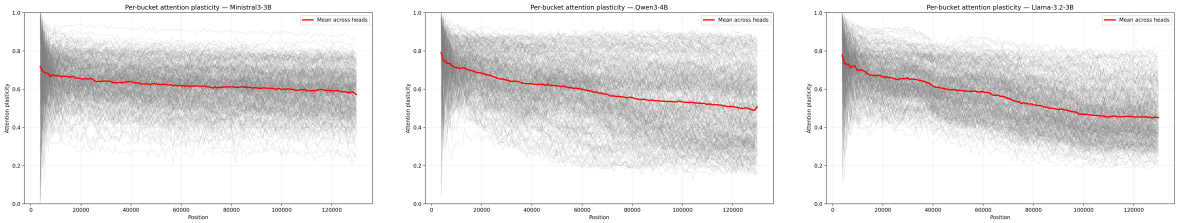


(c) Llama-3.2-3B



(d) Llama-3.2-11B

Figure 6.3: 2D plasticity heatmaps for four representative models. Horizontal axis: inter-key distance $|t_{k_i} - t_{k_j}|$. Vertical axis: key-to-query distance. Warm colors indicate high plasticity (content-driven key selection); dark colors indicate low plasticity (position-dominated). Ministral maintains high plasticity nearly everywhere; Qwen shows a diagonal gradient; Llama-3.2-3B shows the strongest contrast; Llama-3.2-11B shows non-smooth patches suggesting NoPE layer interference.



(a) Ministral-3-3B (b) Qwen-3-4B (c) Llama-3.2-3B

Figure 6.4: Per-head plasticity profiles (gray) with model mean (red). Three qualitatively different organizations: Ministral produces a tight, homogeneous bundle; Qwen shows bimodal specialization with content-focused heads near 1.0 and position-locked heads near 0.2; Llama shows a wide spread that converges at distant positions—position eventually dominates every head.

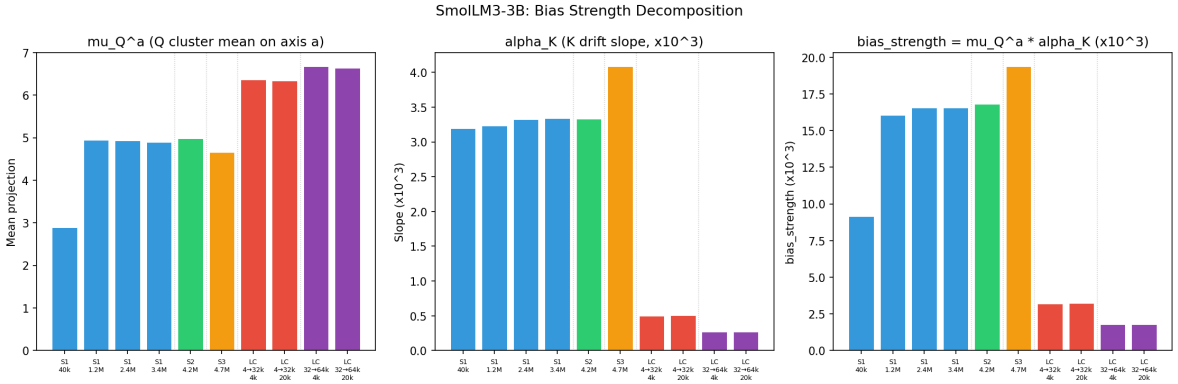


Figure 6.5: Decomposition of bias strength during SmolLM3-3B training. μ_Q^a (query centroid projection) increases during LC extension while α_K (key drift slope) collapses, producing the net bias reduction. The collapse is driven entirely by key slope flattening.

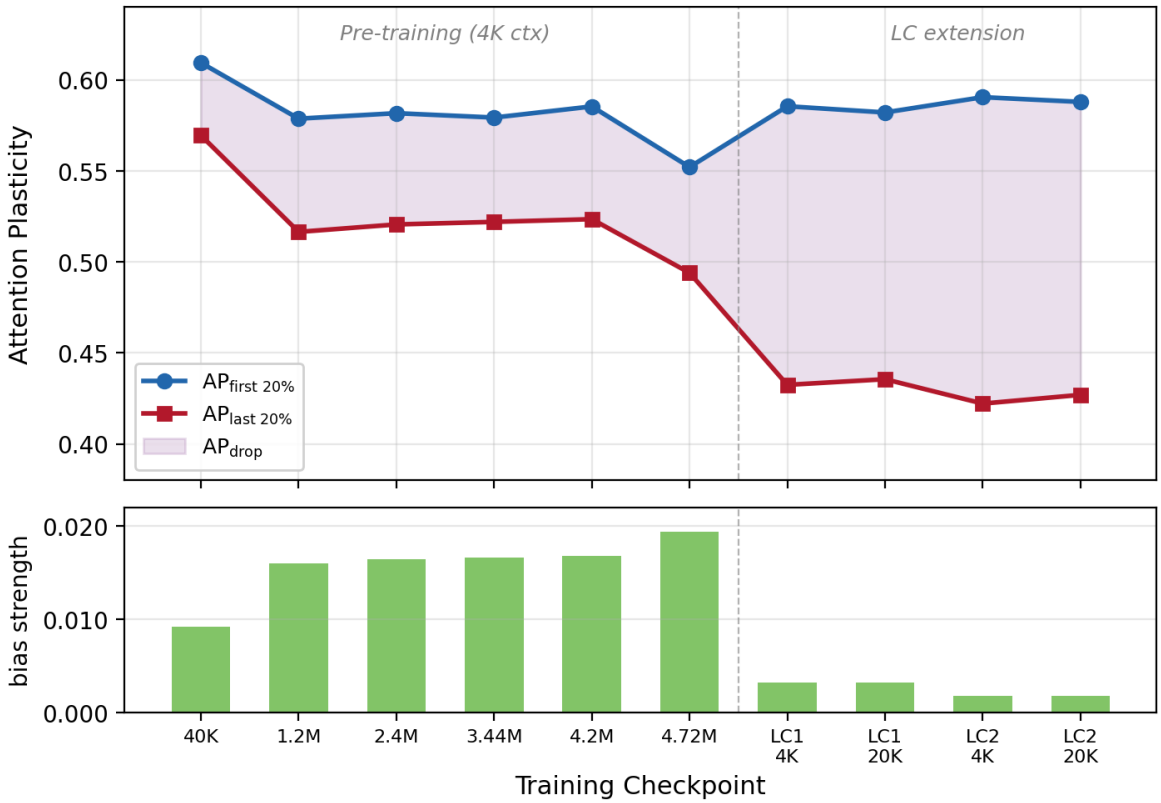


Figure 6.6: SmolLM3-3B plasticity trajectory across training. Short-context plasticity ($AP_{\text{first }20\%}$) recovers during LC extension, but long-context plasticity ($AP_{\text{last }20\%}$) continues to decline. The gap (AP_{drop}) triples despite a 10× reduction in bias strength.

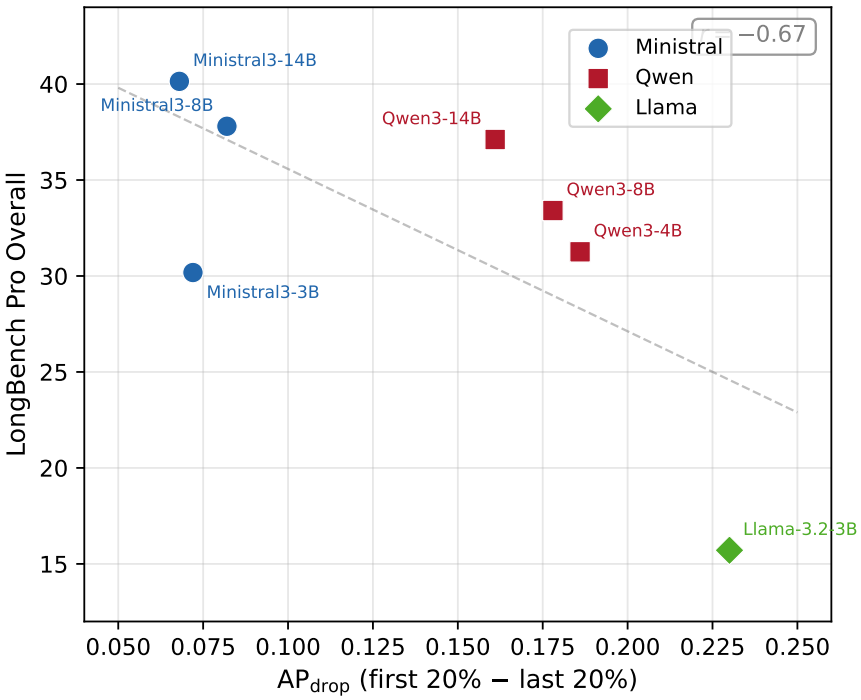


Figure 6.7: AP_{drop} vs. LongBench-Pro overall score for 7 matched models. Lower plasticity drop (flatter profiles) associates with higher benchmark scores. Ministral-3-3B (circled) sits off the trend: low AP_{drop} but low LBP, suggesting its bottleneck is base capability rather than context preservation.

Chapter 7

Discussion

7.1 Bias Reduction Is Necessary but Not Sufficient

The training dynamics of SmolLM3-3B (Section 6.3) reveal the central interpretive finding of this thesis: positional bias reduction and effective context utilization are related but not equivalent. This section develops the implications.

7.1.1 The Dissociation

During long-context extension, bias strength collapses 10× through α_K flattening (Table 6.2). This collapse is mechanistically specific: the key position gradient flattens while position correlation $|r_k^{(a)}|$ remains above 0.90. Keys still encode position with high fidelity, but the encoding becomes more uniform—positions are distinguished without being preferentially weighted.

The collapse recovers short-context plasticity ($AP_{\text{first }20\%}$: 0.552 → 0.588), returning it to early stage 1 levels. Locally, bias reduction works: removing the positional term from the score difference restores content-driven key selection at nearby positions.

However, at 48K–64K positions, plasticity is 0.43—far below the 0.57 the model achieved at 3.5K–4K during pre-training at comparable bias levels. Despite a 10× bias collapse, AP_{drop} *triples* from ~0.06 to ~0.16 (Figure 6.6). A factor beyond positional bias—plausibly content signal decay—constrains long-context flexibility.

7.1.2 Content Signal Decay

The excess AP_{drop} after bias collapse ($\sim 0.16 - 0.06 = 0.10$) reflects the content component of the score difference losing variance at distant positions. After the positional term is nearly zeroed, even minimal residual bias dominates if the content signal is weak.

Two candidate mechanisms may drive this content signal decay:

- **RoPE rotation accumulation.** The rotation planes in the complement subspace progressively decorrelate distant keys from the query, reducing the content-signal variance in the score difference. This would be a structural property of RoPE, not a training artifact—and would affect all RoPE-based models regardless of how thoroughly bias is reduced during training.
- **Attention sink competition.** Early-position tokens accumulate disproportionate attention weight through the attention sink phenomenon [19, 40]. This concentration reduces the effective attention budget available for long-context content, even when the per-head bias mechanism is weakened.

We do not resolve which mechanism dominates—this remains an open question (Section 7.6).

7.1.3 Cross-Model Confirmation

SmolLM3 after long-context extension ($AP_{drop} \approx 0.16$) matches Qwen-3 models (0.16–0.19)—both undergo standard LC training at comparable scales. Ministral-3 achieves $AP_{drop} \approx 0.07$ at comparable bias levels over a $4\times$ longer context window (256K vs. 64K). Whatever Ministral-3's training recipe accomplishes, its low AP_{drop} is consistent with maintaining content signal strength at distance—not just reducing bias.

This suggests that current LC extension methods, which focus on making position encodings generalize to longer sequences (RoPE scaling, NTK-aware interpolation [1], YaRN [31]), primarily address the bias component of ECL degradation. The content-decay component may require complementary strategies.

7.2 Why the Profile Matters More Than the Scalar

Aggregate plasticity ($AP_{overall}$) fails as a cross-family predictor of long-context performance. The positional degradation pattern—AP_{drop}—succeeds. This distinction has practical implications for model evaluation.

The Ministral-3B diagnostic. Ministral-3-3B is the diagnostic outlier (Figure 6.7): it has the highest aggregate plasticity (0.622) among all 7 models with LongBench-Pro scores, yet scores only 30.18. Its plasticity profile is nearly flat ($AP_{drop} = 0.072$, comparable to the 14B variant), indicating strong context preservation. The low benchmark score reflects limited base capability at 3B scale—the model's knowledge and reasoning ability, not its attention mechanics, is the bottleneck.

Decomposing benchmark performance. Long-context benchmark performance depends on two factors: *base model capability* (knowledge, reasoning, instruction following) and *context preservation* (maintaining attention flexibility at distance). Aggregate plasticity approximates the second factor, but since context preservation varies less than base capability across model scales, the aggregate is dominated by the noisier capability term in cross-family comparisons.

AP_{drop} isolates context preservation more cleanly. By measuring the *slope* of plasticity decline rather than the absolute level, it is robust to differences in base capability. Within a family (controlled architecture and training recipe), both AP_{drop} and LBP move together—larger Qwen-3 models degrade less and score higher. Across families, AP_{drop} separates families by context preservation strategy while remaining agnostic to base capability.

Practical implication. When evaluating a model for long-context deployment, aggregate benchmark scores conflate context preservation with base capability. A model with low aggregate LBP but flat AP_{drop} is *capability-limited*—it preserves context well but lacks the knowledge or reasoning ability to exploit it. A model with high aggregate LBP but steep AP_{drop} is *context-limited*—it performs well on average but may fail unpredictably on inputs requiring information from distant positions. Per-position plasticity profiles provide a more targeted diagnostic than aggregate scores for distinguishing these failure modes.

7.3 Unifying the Three Analyses

PCA, rotation, and plasticity are not three independent studies—they form a coherent pipeline where each step resolves a question left open by the previous one.

7.3.1 The Progression

PCA discovers that 9–32% of Q/K variance is linear in token position and that heads cluster by family in (r_q, r_k) space. But its axes confound position with Q/K identity (Section 4.2.3), and it cannot determine whether the observed positional variance functionally constrains attention.

The rotation model resolves the first problem by constructing axes with guaranteed semantic

meaning: axis a carries all linear position covariance, axis b accounts for the Q/K centroid offset. This corrects PCA's apparent $r_q > r_k$ asymmetry (Figure 6.2) and provides a parametric decomposition: bias_strength = $\mu_Q^a \times \alpha_K$ (Equation 4.6). But bias magnitude does not equal bias impact—a head with large bias could still attend flexibly if the content signal is strong.

Plasticity resolves the second problem by measuring the competition directly: given a random query and two keys, does the positional term or the content term determine which key receives higher attention? The plasticity metric converts geometric structure into functional consequence.

7.3.2 The Geometric Connection

The 2D plasticity heatmaps (Figure 6.3) integrate both factors from the rotation model. Inter-key distance relates to δ_1, the positional coordinate difference between keys on the drift axis—larger inter-key distance means larger positional score difference, amplified by α_K. Query-to-key-midpoint distance relates to the query's own positional coordinate, which enters the score through μ_Q^a. At larger query positions, the positional term in the score mean grows, increasing its dominance over content variance.

The heatmap is the joint effect of both rotation parameters, mediated by content variance in the complement subspace. The family-specific patterns (Ministral uniform, Qwen diagonal gradient, Llama steep contrast) reflect different configurations of $(\mu_Q^a, \alpha_K, \sigma^2_{\text{content}})$.

7.3.3 What Each Analysis Uniquely Contributes

Even within the unified framework, each analysis offers something the others cannot:

- **PCA**: model-agnostic structural fingerprint, head taxonomy, variance budget. The head taxonomy tracks qualitative changes during training (Figure A.4 in Appendix A) that are invisible to the other analyses.
- **Rotation**: parametric decomposition with training-dynamics specificity. The bias decomposition (Figure 6.5) identifies *which component* collapses during LC extension (α_K, not μ_Q^a)—a mechanistic distinction PCA and plasticity cannot make.
- **Plasticity**: functional relevance and benchmark prediction. Only plasticity connects to behavioral performance (Table 6.3) and captures the 2D competition geometry that reveals how inter-key distance and query-to-key distance jointly determine position dominance.

7.4 Implications for Long-Context Training

The findings suggest specific directions for improving long-context training beyond current position-extension methods.

Evaluate with positional profiles, not aggregates. Perplexity on long documents is a standard evaluation metric during long-context training. But perplexity averages over all positions, masking long-context degradation. The SmolLM3 trajectory (Table 6.2) shows that AP_{drop} triples during LC extension even as the model successfully processes longer inputs. Per-position plasticity profiles—or any position-resolved metric—would detect this steepening gradient that aggregate metrics miss.

Address content signal decay explicitly. Current LC extension methods (position interpolation [8], NTK-aware RoPE scaling [1], YaRN [31], LongRoPE [14]) focus on making position encodings generalize to longer sequences. Our results suggest this addresses the bias component but not the content-decay component. Potential approaches to the latter include:

- Training objectives that explicitly reward long-context retrieval, rather than relying on next-token prediction which averages over all positions.
- Architectural modifications that preserve content signal fidelity across positions, such as differential attention [43] or content-gated mechanisms.
- Data strategies that expose the model to long-range dependencies during pre-training, not only during LC fine-tuning—analogous to the upsampled long-document data of [41].

The Ministral-3 target. Ministral-3 achieves 2× flatter plasticity profiles than standard LC-trained models at comparable bias levels. While the specific training recipe is proprietary, the metric provides a concrete, mechanistically grounded target: $AP_{drop} < 0.10$ across 128K+ context. This gives long-context training an explicit optimization objective beyond "extend the context window."

7.5 Limitations

Observational design. All analyses are observational: we measure correlations between mechanistic metrics and behavioral performance, not causal effects. Claiming that low plasticity *causes* ECL failure would require interventions—ablating low-plasticity heads and measuring retrieval accuracy degradation. We make associative claims only.

Single capture dataset. All analyses are computed over a fixed set of 500 LongBench-Pro examples at 128K+ length. The stability of plasticity profiles across document types (code, dialogue, structured data) is untested. The profiles may partly reflect dataset-specific statistics rather than purely architectural properties.

Three model families. Eleven models from three families (Ministral-3, Qwen-3, Llama-3.2) is a limited sample. The findings may not generalize to architecturally different models such as mixture-of-experts models [12, 13], models with linear attention [18, 30], or models at substantially larger scales. The training dynamics story relies on a single model's trajectory (SmolLM3-3B).

Pairwise ranking, not softmax weights. Plasticity evaluates attention as a reranking mechanism: it measures whether query content determines the pairwise ordering of keys. Since any total ordering is determined by its pairwise comparisons, corrupted pairwise orderings imply a corrupted global ranking. However, plasticity does not directly measure softmax weight concentration. A head could produce correct pairwise orderings yet still spread attention too thinly across many keys, or conversely produce incorrect orderings that are masked by softmax saturation. The metric captures ranking quality, not weight allocation.

Linear drift and Gaussian assumptions. The plasticity decay theorem (Theorem 4.4) and the Gaussian closed form (Equations 4.11–4.14) assume that positional drift is linear and that score differences are approximately Gaussian. Both are empirically supported—linear fits achieve high R^2 on the drift axis, and Q-Q plots of score differences show approximate normality—but they are approximations. Non-linear positional structure (e.g., from RoPE's rotation planes in the complement subspace) is not captured by the linear model and may contribute to plasticity decay through mechanisms the theorem does not describe. The Gaussian assumption may also break down for heads with strongly non-Gaussian content distributions, though the central limit theorem provides some robustness when aggregating over many key pairs.

Per-head independence. We analyze each attention head independently. Per-head plasticity does not capture cross-layer interactions: a low-plasticity early head feeding into a high-plasticity later head may not limit model-level ECL. Circuit-level analysis [17, 29] would be needed to understand how individual head profiles compose into model-level behavior.

Base vs. instruct model confound. Mechanistic metrics are computed on base (non-instruct) model weights, while benchmark scores come from instruct-tuned variants of the same architectures. Instruction tuning could alter the attention geometry we measure, in which case the base-model metrics may not fully correspond to the instruct-model behavior on benchmarks. The correlation

between plasticity drop and benchmark ordering is therefore an association between base-model geometry and instruct-model performance, not a direct measurement of the mechanism underlying the benchmark scores.

GQA key sharing. In GQA models, multiple query heads share the same key head. We analyze each (query head, shared key head) pair independently but do not study whether query heads within a GQA group develop coordinated or divergent plasticity profiles. Coordination within GQA groups could amplify or mitigate the effects we observe at the individual head level.

7.6 Future Work

Interventional validation. The natural next step is to ablate or modify low-plasticity heads and measure the effect on downstream retrieval accuracy. If high-plasticity heads are necessary for long-context performance, ablating them should degrade performance selectively at distant positions. Conversely, clamping bias_strength to zero in specific heads should improve long-context plasticity if the bias mechanism is causal.

Characterizing content signal decay. The residual AP_{drop} after bias collapse is attributed to content signal decay, but the precise mechanism is uncharacterized. Candidate explanations include RoPE rotation accumulation in the complement subspace, attention sink competition from early-position tokens, and representational drift where content features themselves become less informative at distant positions. Distinguishing these requires controlled experiments varying RoPE parameters, context length, and initial-token presence.

Broader model coverage. The framework should be extended to architecturally diverse models: mixture-of-experts models [12, 13] where expert routing may interact with plasticity; hybrid architectures that interleave attention with state-space layers [18], where plasticity applies to the attention layers and the SSM layers provide a natural contrast; and larger scales (70B+) to test whether the plasticity profile changes qualitatively.

Per-length benchmark validation. Obtaining per-length LongBench-Pro scores for the remaining 6 matched models would enable per-position correlation analysis across families, not just for Ministral-3-14B. RULER analysis [10] with per-length scores would test whether plasticity predicts the catastrophic performance cliff that some models exhibit beyond their effective context length.

Temporal extension. Applying the analysis to other models with open intermediate checkpoints would test whether the three-phase pattern (bias growth, bias collapse, content decay persistence) is universal or specific to SmolLM3's training recipe.

Chapter 8

Conclusion

Large language models increasingly claim context windows of 128K tokens and beyond, yet behavioral evaluations consistently show performance degradation well before these limits. This thesis developed a geometric framework to understand why.

We introduced three complementary analyses—PCA decomposition, a planar rotation model, and attention plasticity—all operating on post-RoPE query and key vectors captured from a single forward pass. PCA reveals that 9–32% of Q/K variance is linear in token position (varying by family and vector type), with heads clustering by model family. The rotation model isolates positional bias into a scalar quantity (bias_strength = $\mu_Q^a \times \alpha_K$) and corrects a confound in PCA that makes queries appear more positional than keys. Attention plasticity converts this geometric structure into a functional measure: the probability that query content, rather than position, determines key selection.

Across 13 models from three families, plasticity drop—the degradation from early to late context positions—separates model families in the same order as LongBench-Pro benchmark scores: Ministral-3 ($AP_{drop} \approx 0.07$) outperforms Qwen-3 (~0.17) outperforms Llama-3.2 (~0.23). Within the Qwen-3 family, the relationship is monotonic with scale; in the Ministral family, base capability rather than context preservation is the bottleneck at 3B (Section 6.4). These results provide evidence that plasticity profiles serve as a mechanistically grounded diagnostic of effective context length.

Tracking SmolLM3-3B across 10 training checkpoints revealed that long-context extension collapses positional bias 10× through key drift slope flattening, recovering short-context plasticity. Yet AP_{drop} triples during the same process. Bias reduction is necessary but not sufficient: after the positional term is nearly zeroed, content signal decay at distant positions becomes the binding constraint. This finding suggests that current position-extension methods address the bias component of ECL degradation but not the content-decay component, and that maintaining content signal fidelity across positions may be an important complementary dimension of long-context training.

Bibliography

[1] bloc97 (Bowen Peng). *NTK-Aware Scaled RoPE allows LLaMA models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation.* Reddit post, r/LocalLLaMA. Later formalized in YaRN (Peng et al., ICLR 2024). June 2023. URL: https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/.

[2] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. "GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints". In: *Conference on Empirical Methods in Natural Language Processing*. 2023. arXiv: 2305.13245 [cs.CL]. URL: https://arxiv.org/abs/2305.13245.

[3] Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. "Why Does the Effective Context Length of LLMs Fall Short?" In: *International Conference on Learning Representations*. Singapore, Apr. 2025. arXiv: 2410.18745 [cs.CL]. URL: https://openreview.net/forum?id=eoln5WgrPx.

[4] Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. "LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks". In: *Annual Meeting of the Association for Computational Linguistics*. 2025, pp. 3639–3664.

[5] Elie Bakouch, Carlos Miguel Patiño, Anton Lozhkov, Edward Beeching, Aymeric Roucher, Nouamane Tazi, Aksel Joonas Reedi, Guilherme Penedo, Hynek Kydlicek, Clémentine Fourrier, Nathan Habib, Kashif Rasul, Quentin Gallouédec, Hugo Larcher, Mathieu Morlon, Joshua Lochner, Vaibhav Srivastav, Xuan-Son Nguyen, Colin Raffel, Lewis Tunstall, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. *SmolLM3: smol, multilingual, long-context reasoner*. Hugging Face Blog. Technical blog post accompanying the SmolLM3-3B model release. July 2025. URL: https://huggingface.co/blog/smollm3.

[6] Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, João G.M. Araújo, Alex Vitvitskyi, Razvan Pascanu, and Petar Veličković. "Transformers Need Glasses! Information Over-squashing in Language Tasks". In: *Advances in Neural Information Processing Systems*. Vancouver, Canada: Curran Associates, Inc., Dec. 2024. DOI: 10.52202/079017-3114. arXiv: 2406.04267 [cs.CL]. URL: https://proceedings.neurips.

cc/paper_files/paper/2024/hash/b1d35561c4a4a0e0b6012b2af531e149-Abstract-Conference.html.

[7] Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. "Round and Round We Go! What Makes Rotary Positional Encodings Useful?" In: *International Conference on Learning Representations*. Singapore, Apr. 2025. arXiv: 2410.06205 [cs.CL]. URL: https://openreview.net/forum?id=GtvuNrk58a.

[8] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. *Extending Context Window of Large Language Models via Position Interpolation*. 2023. arXiv: 2306.15595 [cs.CL].

[9] Ziyang Chen, Xing Wu, Junlong Jia, Chaochen Gao, Qi Fu, Debing Zhang, and Songlin Hu. *LongBench Pro: A More Realistic and Comprehensive Bilingual Long-Context Evaluation Benchmark*. Jan. 2026. arXiv: 2601.02872 [cs.CL]. URL: https://arxiv.org/abs/2601.02872.

[10] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. "RULER: What's the Real Context Size of Your Long-Context Language Models?" In: *Conference on Language Modeling*. Philadelphia, PA, USA, Oct. 2024. arXiv: 2404.06654 [cs.CL]. URL: https://openreview.net/forum?id=kIoBbc76Sy.

[11] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. New Orleans, LA, USA: Curran Associates, Inc., Dec. 2022, pp. 16344–16359. DOI: 10.5555/3600270.3601459. arXiv: 2205.14135 [cs.LG]. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/67d57c32e20fd0a7a302cb81d36e40d5-Paper-Conference.pdf.

[12] DeepSeek-AI. "DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model". In: *arXiv preprint arXiv:2405.04434* (2024). URL: https://arxiv.org/abs/2405.04434.

[13] DeepSeek-AI. "DeepSeek-V3 Technical Report". In: *arXiv Technical Report* (Dec. 2024). arXiv: 2412.19437 [cs.CL]. URL: https://arxiv.org/abs/2412.19437.

[14] Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. "LongRoPE: Extending LLM Context Window Beyond 2 Million Tokens". In: *International Conference on Machine Learning*. 2024.

[15] Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A. Huerta, and Hao Peng. "Context Length Alone Hurts LLM Performance Despite Perfect Retrieval". In: *Findings of the Association for Computational Linguistics: EMNLP*. 2025.

[16] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. *The Llama 3 Herd of Models*. 2024. arXiv: 2407.21783 [cs.AI].

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

[18] Albert Gu and Tri Dao. "Mamba: Linear-Time Sequence Modeling with Selective State Spaces". In: *International Conference on Learning Representations*. 2024. arXiv: 2312.00752 [cs.LG].

[19] Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. "When Attention Sink Emerges in Language Models: An Empirical View". In: *International Conference on Learning Representations*. Singapore, Apr. 2025. arXiv: 2410.10781 [cs.CL]. URL: https://openreview.net/forum?id=78Nn4QJTEN.

[20] 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. arXiv: 2310.06825 [cs.CL].

[21] Greg Kamradt. *Needle In A Haystack – Pressure Testing LLMs*. GitHub repository. Nov. 2023. URL: https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

[22] Kimi Team. "Kimi Linear: An Expressive, Efficient Attention Architecture". In: *arXiv Technical Report* (2025). arXiv: 2510.26692 [cs.CL].

[23] Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. "BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack". In: *Advances in Neural Information Processing Systems*. Vol. 37. Dec. 2024. arXiv: 2406.10149 [cs.CL]. URL: https://proceedings.neurips.cc/paper_files/paper/2024/hash/c0d62e70dbc659cc9bd44cbcf1cb652f-Abstract-Datasets_and_Benchmarks_Track.html.

[24] Mosh Levy, Alon Jacoby, and Yoav Goldberg. "Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large Language Models". In: *Annual Meeting of the Association for Computational Linguistics*. 2024. arXiv: 2402.14848 [cs.CL].

[25] Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, et al. "Ministral 3". In: *arXiv preprint arXiv:2601.08584* (2026). URL: https://arxiv.org/abs/2601.08584.

[26] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. "Lost in the Middle: How Language Models Use Long Contexts". In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 157–173.

[27] Mistral AI. *Mistral 7B v0.2*. Hugging Face model release. Base model weights released March 23, 2024. 32K context window, RoPE theta = 1e6, no sliding window attention. No separate technical report; see arXiv:2310.06825 for the original Mistral 7B. Mar. 2024. URL: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2.

[28] Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. "NoLiMa: Long-Context Evaluation Beyond Literal Matching". In: *International Conference on Machine Learning*. 2025.

[29] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. *In-context Learning and Induction Heads*. Transformer Circuits Thread. Mar. 2022. arXiv: 2209.11895 [cs.LG]. URL: https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

[30] Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Xingjian Du, Teddy Ferdinan, Haowen Hou, Przemysław Kazienko, Kranthi Kiran GV, Jan Kocoń, Bartłomiej Koptyra, Satyapriya Krishna, Ronald McClelland Jr., Jiaju Lin, Niklas Muennighoff, Fares Obeid, Atsushi Saito, Guangyu Song, Haoqin Tu, Cahya Wirawan, Stanisław Woźniak, Ruichong Zhang, Bingchen Zhao, Qihang Zhao, Peng Zhou, Jian Zhu, and Rui-Jie Zhu. "Eagle and Finch: RWKV with Matrix-Valued States and Dynamic Recurrence". In: *Conference on Language Modeling*. Philadelphia, PA, USA, Oct. 2024. arXiv: 2404.05892 [cs.CL]. URL: https://openreview.net/forum?id=soz1SEiPeq.

[31] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. "YaRN: Efficient Context Window Extension of Large Language Models". In: *International Conference on Learning Representations*. 2024.

[32] Qwen Team. "Qwen3 Technical Report". In: *arXiv Technical Report* (May 2025). arXiv: 2505.09388 [cs.CL]. URL: https://arxiv.org/abs/2505.09388.

[33] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. "Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context". In: *arXiv Technical Report* (2024). arXiv: 2403.05530 [cs.CL].

[34] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. "RoFormer: Enhanced Transformer with Rotary Position Embedding". In: *Neurocomputing* 568 (2024), p. 127063.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, pp. 5998–6008.

[36] Chonghua Wang, Haodong Duan, Songyang Zhang, Dahua Lin, and Kai Chen. "Ada-LEval: Evaluating Long-Context LLMs with Length-Adaptable Benchmarks". In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2024, pp. 3712–3724.

[37] Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. "Eliminating Position Bias of Language Models: A Mechanistic Approach". In: *International Conference on Learning Representations*. Singapore, Apr. 2025. arXiv: 2407.01100 [cs.CL]. URL: https://openreview.net/forum?id=fvkElsJOsN.

[38] Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. "Retrieval Head Mechanistically Explains Long-Context Factuality". In: *International Conference on Learning Representations*. Singapore, Apr. 2025. arXiv: 2404.15574 [cs.CL]. URL: https://openreview.net/forum?id=EytBpUGB1Z.

[39] Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. "On the Emergence of Position Bias in Transformers". In: *International Conference on Machine Learning*. 2025.

[40] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. "Efficient Streaming Language Models with Attention Sinks". In: *International Conference on Learning Representations*. 2024.

[41] Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. "Effective Long-Context Scaling of Foundation Models". In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Mexico City, Mexico, 2024, pp. 4643–4663. URL: https://aclanthology.org/2024.naacl-long.260/.

[42] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. *Qwen2 Technical Report*. 2024. arXiv: 2407.10671 [cs.CL].

[43] Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. "Differential Transformer". In: *International Conference on Learning Representations*. Apr. 2025. arXiv: 2410.05258 [cs.CL]. URL: https://openreview.net/forum?id=OvoCm1gGhN.

[44] Yijiong Yu, Huiqiang Jiang, Xufang Luo, Qianhui Wu, Chin-Yew Lin, Dongsheng Li, Yuqing Yang, Yongfeng Huang, and Lili Qiu. "Mitigate Position Bias in LLMs via Scaling a Single Hidden States Channel". In: *Findings of the Association for Computational Linguistics: ACL*. 2025.

[45] Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. "∞Bench: Extending Long Context Evaluation Beyond 100K Tokens". In: *Annual Meeting of the Association for Computational Linguistics*. 2024, pp. 15262–15277.

Appendix A

Supplementary Material

A.1 Proof of Plasticity Decay (Theorem 4.4)

This section provides the formal proof of Theorem 4.4, which states that attention plasticity decays with query position under the positional-semantic model. The proof proceeds in three steps: we state the model assumptions, establish a tail bound on single-pair plasticity, and aggregate over key pairs.

A.1.1 Model Assumptions

We work under the positional-semantic decomposition of Section 4.4.2. After applying the Householder reflection H (Equation 4.10), the query at position t decomposes as:

$$q_1^{\text{rot}} = \alpha_{\text{pos}} + \beta_{\text{pos}} \cdot t + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma_{\text{pos}}^2) \tag{A.1}$$

$$q_{2:d}^{\text{rot}} \sim \mathcal{N}(\mu_b, \Sigma_b) \tag{A.2}$$

where the positional residual ε is independent of the semantic components $q_{2:d}^{\text{rot}}$, and μ_b, Σ_b are the mean and covariance of the semantic components for query bucket b. The positional drift rate $\beta_{\text{pos}} \neq 0$ (empirically verified for all examined models).

A.1.2 Score Difference Distribution

Lemma A.1 (Gaussian score difference)**.** *Under the positional-semantic model, for a fixed key pair (k_i, k_j) with difference vector $\delta = k_i^{rot} - k_j^{rot}$ in the rotated basis, the score difference $D = q_t^\top (k_i - k_j)$*

conditioned on query position t is Gaussian:

$$D \mid t \sim \mathcal{N}\big(m(t),\, v\big) \tag{A.3}$$

with mean $m(t) = \delta_1(\alpha_{pos}+\beta_{pos}\cdot t)+\delta_{2:d}^\top \mu_b$ and variance $v = \delta_1^2\sigma_{pos}^2+\delta_{2:d}^\top \operatorname{diag}(\sigma_b^2)\,\delta_{2:d}$. The variance v is independent of t.

Proof. Since H is orthogonal, $q^\top(k_i - k_j) = (Hq)^\top(H(k_i - k_j))$. Expanding in the rotated basis:

$$D = \delta_1 \cdot q_1^{\text{rot}} + \delta_{2:d}^\top \cdot q_{2:d}^{\text{rot}} \tag{A.4}$$

Both terms are linear functions of Gaussian random variables. The first term has mean $\delta_1(\alpha_{\text{pos}} + \beta_{\text{pos}}\, t)$ and variance $\delta_1^2\sigma_{\text{pos}}^2$. The second has mean $\delta_{2:d}^\top \mu_b$ and variance $\delta_{2:d}^\top \Sigma_b\, \delta_{2:d}$. By independence of ε and $q_{2:d}^{\text{rot}}$, D is Gaussian with the stated parameters. Since σ_{pos}^2 and Σ_b are position-independent, v does not depend on t. □

A.1.3 Single-Pair Decay

Lemma A.2 (Single-pair plasticity decay)**.** *For a key pair with $\delta_1 \neq 0$ and $\beta_{pos} \neq 0$, the pairwise plasticity satisfies:*

$$PP(t) \leq C_0 \cdot \exp\left(-\frac{\gamma^2 t^2}{8}\right) \tag{A.5}$$

for all $t \geq T_0$, where $\gamma = \delta_1\beta_{pos}/\sqrt{v} \neq 0$, and C_0, T_0 are constants depending on γ and $z_0 = m(0)/\sqrt{v}$.

Proof. Define $z(t) = m(t)/\sqrt{v}$. By Lemma A.1, $z(t) = \gamma\, t + z_0$ where $\gamma = \delta_1\beta_{\text{pos}}/\sqrt{v}$ and $z_0 = (\delta_1\alpha_{\text{pos}} + \delta_{2:d}^\top\mu_b)/\sqrt{v}$. The pairwise plasticity is:

$$\text{PP}(t) = 4\,\Phi(z(t))\,(1 - \Phi(z(t))) \tag{A.6}$$

We use the standard Gaussian tail bound: for $z > 0$, $1-\Phi(z) \leq \exp(-z^2/2)/(\sqrt{2\pi}\, z)$. Since $\Phi(z) \leq 1$, we obtain:

$$\text{PP}(t) \leq 4\,(1 - \Phi(|z(t)|)) \leq \frac{4}{\sqrt{2\pi}\,|z(t)|}\,\exp\left(-\frac{z(t)^2}{2}\right) \tag{A.7}$$

for $|z(t)| \geq 1$. Since $\gamma \neq 0$, for $t \geq T_0 := 2|z_0|/|\gamma|$ we have $|z(t)| \geq |\gamma|t - |z_0| \geq |\gamma|t/2$, giving:

$$z(t)^2 = (\gamma\, t + z_0)^2 \geq (|\gamma|t/2)^2 = \gamma^2 t^2/4 \tag{A.8}$$

Substituting into (A.7) and absorbing the polynomial prefactor $1/|z(t)|$ into the constant C_0 yields the bound (A.5). □

A.1.4 Aggregate Decay

Proof of Theorem 4.4. Partition the set of admissible key pairs into $\mathcal{S}_0 = \{(i,j) : \delta_1 = 0\}$ (keys at the same bucket position) and $\mathcal{S}_1 = \{(i,j) : \delta_1 \neq 0\}$ (keys at distinct positions). Let $\pi_0 = \Pr[\text{pair} \in \mathcal{S}_0]$ under the uniform key-pair sampling.

For pairs in $\mathcal{S}_0$: $m(t) = \delta_{2:d}^\top \mu_b$ is independent of t, so $\mathrm{PP}(t)$ is constant. Their contribution to AP_t converges to:

$$\mathrm{AP}_\infty = \pi_0 \cdot \mathbb{E}[\mathrm{PP} \mid \delta_1 = 0] \tag{A.9}$$

For pairs in $\mathcal{S}_1$: by Lemma A.2, each pair's plasticity decays as $\mathrm{PP}(t) \le C_0^{(i,j)} \exp(-\gamma_{ij}^2 t^2/8)$. Since $\delta_1 \neq 0$ implies $|\delta_1| \ge \delta_{\min} > 0$ (bucket positions are discrete with minimum spacing), we have $\gamma_{ij}^2 \ge \delta_{\min}^2 \beta_{\mathrm{pos}}^2 / v_{\max}$, where $v_{\max}$ is the maximum variance across key pairs. Define $c = \delta_{\min}^2 \beta_{\mathrm{pos}}^2/(8 v_{\max})$. Then for all pairs in $\mathcal{S}_1$ and $t \ge T_0$:

$$\mathbb{E}[\mathrm{PP}(t) \mid \delta_1 \neq 0] \le \bar{C} \cdot \exp(-c \cdot t^2) \tag{A.10}$$

where $\bar{C} = \mathbb{E}[C_0^{(i,j)} \mid \delta_1 \neq 0]$. Combining both cases:

$$\mathrm{AP}_t = \pi_0 \cdot \mathbb{E}[\mathrm{PP} \mid \delta_1 = 0] + (1-\pi_0) \cdot \mathbb{E}[\mathrm{PP}(t) \mid \delta_1 \neq 0] \le \mathrm{AP}_\infty + C \cdot \exp(-c \cdot t^2) \tag{A.11}$$

with $C = (1-\pi_0)\bar{C}$. If all key pairs have distinct positional coordinates ($\pi_0 = 0$), then $\mathrm{AP}_\infty = 0$ and $\mathrm{AP}_t \to 0$. □

A.2 Supplementary Figures: Position Bias Geometry

A.3 Supplementary Figures: Training Dynamics

Table A.1: Per-position plasticity values during SmolLM3-3B long-context extension phases. Short-context plasticity (0–1K, 1K–2K) holds steady while distant positions show progressively lower values as the context window expands.

Checkpoint	0–1K	1K–2K	2K–4K	4K–8K	8K–16K	16K–32K	32K–48K	48K–64K
LC 4K→32K 4K	0.651	0.629	0.598	0.556	0.525	0.464	—	—
LC 4K→32K 20K	0.657	0.626	0.594	0.553	0.524	0.467	—	—
LC 32K→64K 4K	—	0.676	0.641	0.600	0.560	0.526	0.479	0.426
LC 32K→64K 20K	—	0.679	0.639	0.598	0.557	0.526	0.480	0.430

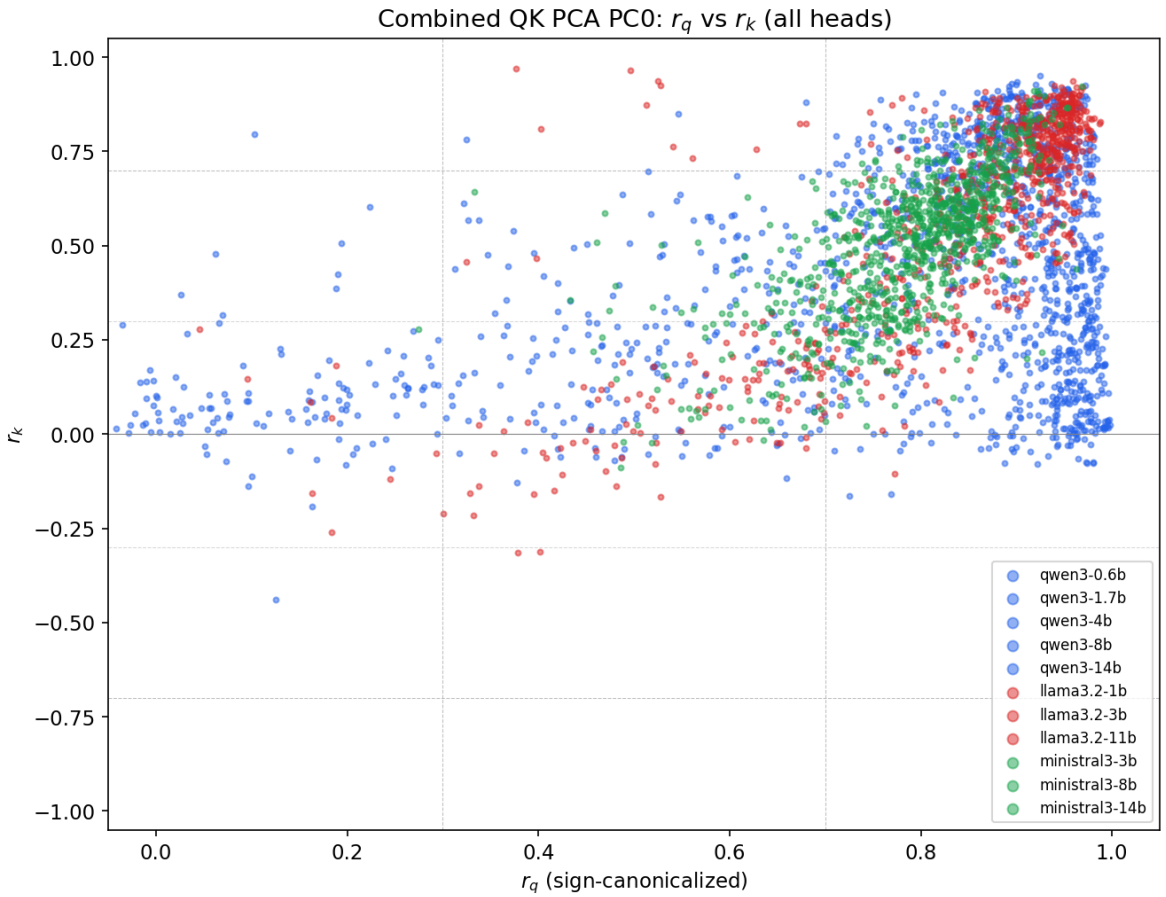


Figure A.1: Head taxonomy scatter plot in (r_q, r_k) space on PC0 across all 11 primary models. Each point is one attention head; colors indicate model family. Llama heads concentrate in the top-right (both correlations high), Qwen models appear Q-positional (high r_q, moderate r_k), and Ministral heads occupy a moderate region. Dashed lines at $|r| = 0.3$ and $|r| = 0.7$ indicate taxonomy thresholds.

A.4 LongBench-Pro Task Structure

Table A.2 lists the 25 secondary tasks in LongBench-Pro, grouped under 11 primary task categories [9]. The benchmark follows a balanced combinatorial design: 25 tasks × 2 languages (EN, ZH) × 6 length bins × 5 samples per cell = 1,500 total samples.

Table A.3 shows the per-length-bin sample distribution. All bins contain exactly 250 samples by design, ensuring that aggregate scores weight each context length equally. Length assignment uses the Qwen tokenizer with ±20% tolerance around target lengths [9].

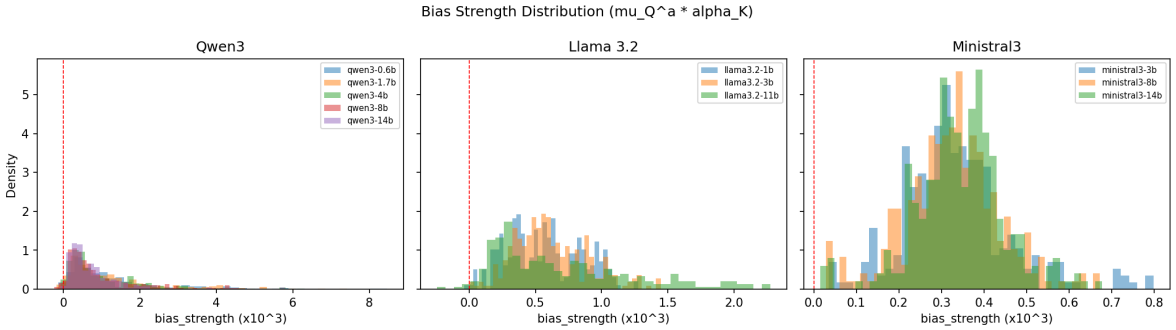


Figure A.2: Distribution of bias strength ($\mu_Q^a \times \alpha_K$) across all 3,239 analyzed heads, grouped by model family. 99.0% of heads show positive bias strength (recency bias); only 31 heads exhibit primacy bias. Distributions are tight within families, confirming that the bias mechanism is architecture-determined.

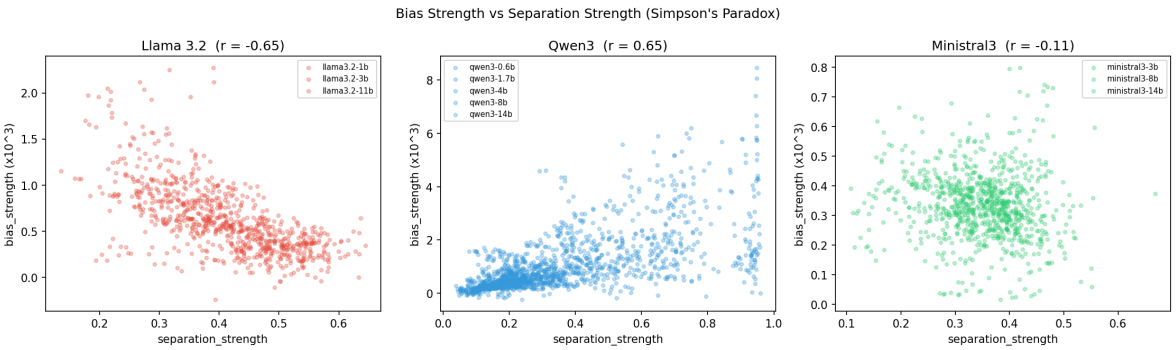


Figure A.3: Bias strength vs. Q/K separation strength per head, colored by model family. The overall correlation is +0.48, but within the Llama family it reverses to −0.65—a Simpson's paradox. Families have genuinely different bias-separation trade-offs that are masked by pooling across architectures.

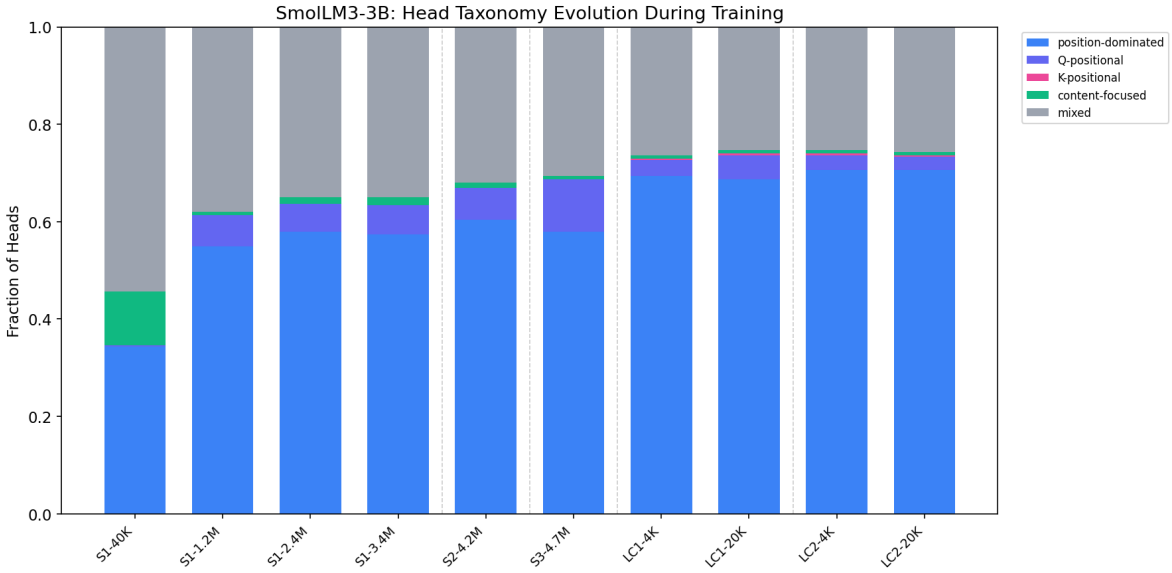


Figure A.4: Head taxonomy evolution across SmolLM3-3B training checkpoints. Position-dominated heads grow from 34% to 55% within stage 1, while content-focused heads virtually disappear. The structural fingerprint is largely set by step 1.2M; subsequent training refines but does not restructure head roles.

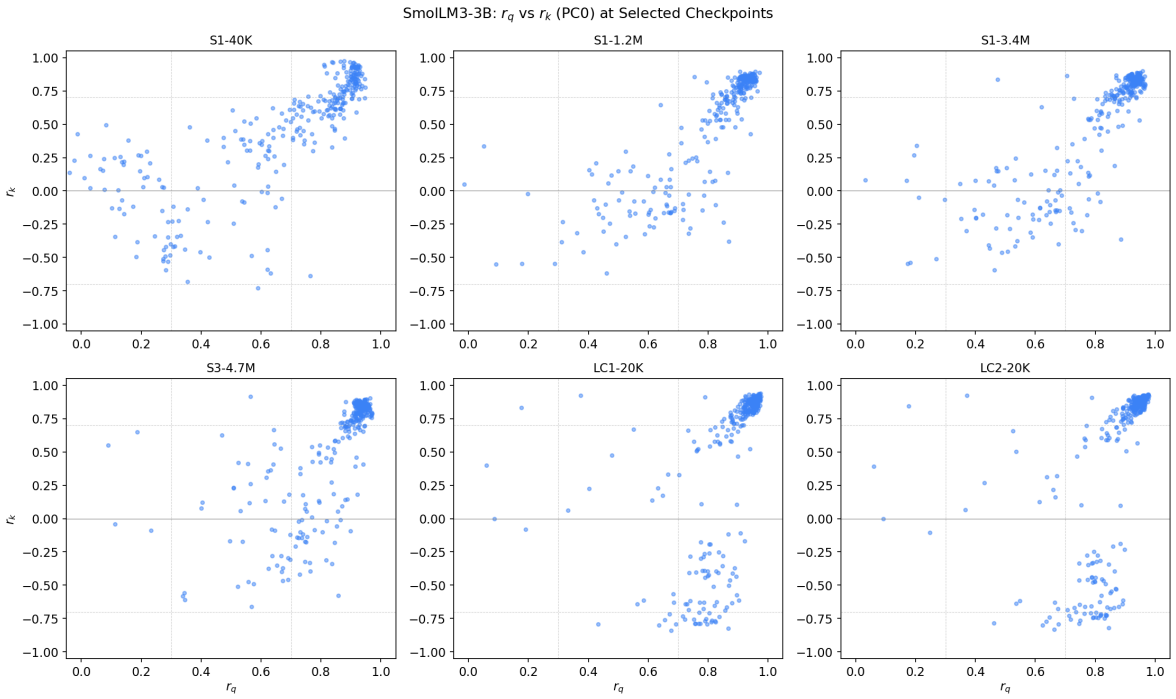


Figure A.5: (r_q, r_k) scatter plots across 6 representative SmolLM3-3B training checkpoints. The head population progresses from a diffuse cloud at step 40K to a tight position-dominated cluster by the end of pre-training. Long-context extension does not alter the taxonomy structure established during pre-training.

Table A.2: LongBench-Pro secondary tasks grouped by primary category. Each secondary task has 60 samples (2 languages × 6 length bins × 5 samples). Context requirement indicates whether the task requires global integration across the full document (F) or localized retrieval from a specific region (P).

ID	Primary category	Secondary task	Ctx
T1.1	Retrieval & Ranking	Global cohesive retrieval	F
T1.2		Key-snippet retrieval	P
T2.1	Sequencing & Reconstruction	Global timeline	F
T2.2		Local causal-chain sorting	P
T3.1	Evidence-Grounded QA	Multi-doc integration QA	F
T3.2		Single-hop fact QA	P
T4.1	Summarization & Synthesis	Global constrained summary	F
T4.2		Query-focused summary	P
T5.1	Attribution & Citation	Full-sentence alignment	F
T5.2		Key-statement alignment	P
T6.1	Aggregation & Clustering	Large-scale clustering	F
T6.2		Targeted subset	P
T6.3		Global frequency	F
T7.1	Consistency & Compliance	Global conflicts	F
T7.2		Targeted rule violation	P
T7.3		Anomaly sweep	F
T8.1	Structured & Numeric	Multi-source verification	F
T8.2	Reasoning	Targeted aggregation	P
T8.3		Procedural state tracking	F
T9.1	Version & Code Diff	Dependency-aware impact	F
T9.2		Localized interface changes	P
T10.1	Rule Induction & ICL	Large-scale rule induction	F
T10.2		Targeted rule induction	P
T11.1	Dialogue Memory &	Long-range entity/commitment	F
T11.2	Long-Horizon Tracking	Short-range reference/state	P

Table A.3: LongBench-Pro sample counts per length bin. The balanced design ensures equal representation across context lengths.

	8K	16K	32K	64K	128K	256K
Samples per bin	250	250	250	250	250	250
English	125	125	125	125	125	125
Chinese	125	125	125	125	125	125