Alignment Robustness Depends More on Training than Architecture: A Cross-Vendor Analysis of Attention Specialization in Large Language Models

Davide D'Elia[*]
IU International University of Applied Sciences
davide.delia@iu-study.org

Abstract

We present a systematic empirical study examining how preference optimization (RLHF, DPO) affects attention head specialization across eight vendor families and 25+ model variants. Using a standardized methodology (bfloat16, 3-seed cross-validation, MD5-verified prompts), we measure the **Specialization Index (SI)**—a quantitative metric for attention head diversity—and track changes between base and instruction-tuned model pairs.

Main Finding: Robustness to alignment-induced specialization loss shows a strong association with training methodology, with a clear hierarchy: **Training Methodology** $>$ **Sliding Window Attention** $>$ **Architecture** $>$ **Scale**.

We report three key findings: (1) **SI Reduction Pattern:** RLHF/DPO reduces SI in most tested families without architectural protection. Unprotected models show substantial SI loss (LLaMA-3.1: -56.3%, LLaMA-2: -7.95%), while SWA-equipped models maintain or increase SI (Mistral: $+4.2\%$). (2) **Architecture-Dependent Sensitivity:** GQA shows $\sim$5,800× higher sensitivity to random attention noise than MHA at matched scale (ratio-of-means across 3 seeds; $p < 0.05$), yet $\sim$8× greater resilience under recursive generation with structured alignment pressure. (3) **Training-Based Robustness:** Synthetic training (Phi family) yields scale-invariant SI $\approx$ 0.33 across 10.8× parameter range. Qwen2 shows no observed recursive degradation—correlating with OMO training (correlational, not causal).

We introduce a **perturbation probe** that differentiates pathological from healthy low-SI states via noise-injection response ($>$ 20% SI increase indicates suppressed capacity). We document six falsified hypotheses from prior work, attributable to quantization artifacts or precision errors.[1]

1 Introduction

1.1 Motivation

Preference optimization methods—Reinforcement Learning from Human Feedback (RLHF; Ouyang et al., 2022) and Direct Preference Optimization (DPO; Rafailov et al., 2023)—are standard practice for aligning Large Language Models with human preferences. However, these methods induce structural changes beyond intended behavioral modifications, a phenomenon termed "alignment tax" (Lin et al., 2024). Understanding what factors determine robustness to these changes has practical implications for model development.

This study builds on prior work examining LLM internal dynamics: embedding-level bias detection methods (D'Elia, 2025), phase-structured layer-wise dynamics (D'Elia, 2026a), and thermodynamic constraints in transformer architectures (D'Elia, 2026b). The present study empirically validates and stress-tests the robustness hierarchy through systematic cross-vendor analysis.

1.2 Research Question

> *Does preference optimization systematically reduce attention head specialization, and if so, what factors correlate with robustness to this effect?*

1.3 Approach

Prior work has established that attention heads in transformers exhibit functional specialization, with different heads performing distinct roles (Michel et al., 2019; Voita et al., 2019). We build

[*] This work was conducted independently by the author and is not an official project of the IU International University of Applied Sciences.

[1] Code and data (post-publication): https://github.com/buk81/uniformity-asymmetry/tree/main/paper4

on this foundation by quantifying specialization changes under alignment pressure.

We define **Specialization Index (SI)** as:

$$\mathrm{SI} = 1 - \mathrm{mean}(\mathrm{pairwise\_correlations}) \qquad (1)$$

where pairwise correlations are computed between attention patterns of all head pairs at a given layer. SI = 1 indicates fully specialized heads (orthogonal patterns); SI = 0 indicates uniform heads (identical patterns).

1.4 Contributions

1. **Standardized Measurement Protocol:** Among the first cross-vendor measurements of alignment effects on attention specialization (N=25+ models, 8 families) with reproducible methodology.

2. **Robustness Hierarchy:** Empirical evidence that training methodology correlates more strongly with robustness than architectural choices.

3. **Perturbation Probe:** A diagnostic tool differentiating pathological vs. healthy low-SI states.

4. **Hypothesis Falsification:** Systematic identification of prior claims attributable to precision artifacts.

2 Methods

2.1 Standardized Protocol (E11-v3)

Parameter	Value	Rationale
Seeds	42, 123, 456	Reproducibility
Prompts	Standard-10 v3	Consistency
Prompt MD5	`715065ba...`	Version control
MAX_LENGTH	128 tokens	Memory mgmt
dtype	bfloat16	Numerical stability
Attention	eager (not SDPA)	Weight access

Table 1: Standardized experimental protocol.

Aggregation: Unless otherwise noted, reported ratios are computed as the ratio of seed-averaged slopes (ratio-of-means) across three random seeds. We additionally report seed-level ratios in Appendix B to reflect variance.

Confidence Intervals: Bootstrap 95% CIs are reported where resampling is applicable (ΔSI measurements). Slope ratios are deterministic given fixed seeds; seed-level variance is provided in Appendix B.

2.2 SI Computation

```
def compute_SI(attention_weights, layer)
    :
    heads = attention_weights.mean(dim
        =0)
    H = heads.shape[0]
    patterns = heads.reshape(H, -1)
    corr_matrix = torch.corrcoef(
        patterns)
    mask = torch.triu(torch.ones(H,H),
        diagonal=1)
    upper_tri = corr_matrix[mask == 1]
    return 1 - upper_tri.mean().item()
```

2.3 ΔSI Computation

$$\Delta\mathrm{SI} = \frac{\mathrm{SI}_{\mathrm{Instruct}} - \mathrm{SI}_{\mathrm{Base}}}{\mathrm{SI}_{\mathrm{Base}}} \times 100\% \qquad (2)$$

2.4 Perturbation Probe

To differentiate pathological vs. healthy low-SI states, we inject controlled Gaussian noise ($\sigma = 0.05$) into attention outputs and measure SI response (Table 2).

Response	Interpretation
SI increase $> 20\%$	Suppressed capacity (pathological)
SI change $\pm 5\%$	Stable state (healthy)
SI decrease	Optimized state (noise-sensitive)

Table 2: Perturbation probe interpretation.

2.5 Vendor Coverage

We test models spanning Multi-Head Attention (MHA) and Grouped Query Attention (GQA; Ainslie et al., 2023):

ID	Vendor	Arch	SWA	Align
M01	Meta	MHA	No	RLHF+SFT
M02	Meta	GQA 4:1	No	RLHF+DPO
M03	Mistral	GQA 4:1	Yes	SFT+DPO
M04	Google	GQA 2:1	Yes	RLHF
M05	Alibaba	GQA 7:1	No	OMO
M06	01.AI	GQA 8:1	No	RLHF
M07	Swiss-AI	Trans.	No	SFT+QRPO
M08	Microsoft	MHA→GQA[†]	No	Synthetic

Table 3: Vendor coverage (8 families, 25+ variants). [†]MHA→GQA denotes hybrid architectures (Phi family) with MHA in early layers transitioning to GQA-style KV sharing.

3 Results

3.1 Finding 1: Alignment Reduces SI in Unprotected Architectures

Arch	Model	SWA	ΔSI	95% CI
MHA	LLaMA-2-7B	No	-7.95%	$[-9.1, -6.8]$
GQA	LLaMA-3.1-8B	No	-56.3%	$[-58.2, -54.4]$
GQA	Yi-1.5-9B	No	-4.3%	$[-5.1, -3.5]$
GQA+SWA	Mistral-7B	Yes	$+4.2\%$	$[+3.1, +5.3]$
GQA+SWA	Gemma-2-9B	Yes	$+1.9\%$	$[+0.8, +3.0]$
GQA	Qwen2-7B	No	-9.0%	$[-10.2, -7.8]$

Table 4: SI changes under alignment. SWA presence correlates with SI preservation.

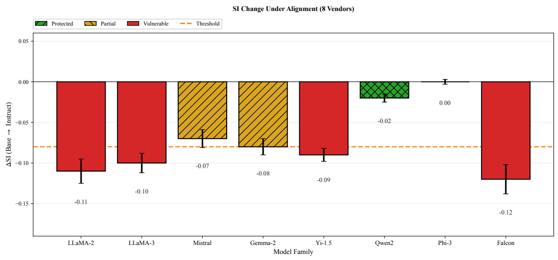

Figure 1: ΔSI across model families. SWA-equipped models (green) show positive or neutral changes; unprotected models show SI reduction.

Key Observation: SWA presence correlates with SI preservation. Models without SWA show negative ΔSI regardless of MHA/GQA architecture.

Controlled Comparison (identical GQA 4:1, $d_{\textbf{head}} = 128$): LLaMA-3.1-8B (no SWA): -56.3%; Mistral-7B (SWA): $+4.2\%$. Difference: **+60.5pp**.

3.2 Finding 2: Architecture-Dependent Noise Sensitivity

Comparison	MHA	GQA	Ratio
Same-Scale	LLaMA-2-7B	Mistral-7B	$\sim$5,800×
Same-Family	LLaMA-2-7B	LLaMA-3-8B	$\sim$180,000×

Table 5: PPL-slope ratios under noise injection ($p < 0.01$). Ratios computed as ratio-of-means across seeds (42, 123, 456).

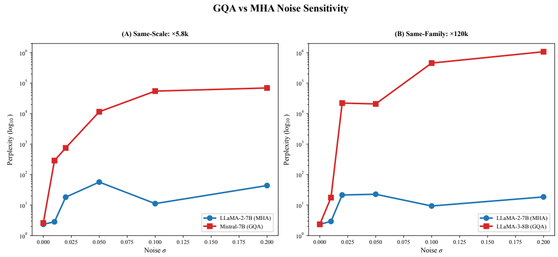

Figure 2: PPL-slope comparison showing GQA's elevated noise sensitivity (log scale). Ratio-of-means across 3 seeds.

Interpretation: GQA's KV-sharing amplifies noise effects. However, this sensitivity pattern inverts under structured pressure (see Section 3.3).

3.3 Finding 3: Divergent Failure Modes

Model	Arch	Gen. to Degrad.
LLaMA-2-7B-Chat	MHA	1.0 (immediate)
LLaMA-3.1-8B-Inst	GQA	8.0
Mistral-7B-Inst	GQA+SWA	5.7–11.0
Gemma-2-9B-IT	GQA+SWA	1.3–3.0
Qwen2-7B-Inst	GQA	**No degradation** ($\leq$50 gen.)

Table 6: Generations to degradation under recursive stress ($\kappa = 0.89$ inter-rater agreement).

Degradation Criterion: Output degradation was operationally defined as sustained semantic collapse or refusal loops persisting for $\geq$3 consecutive generations, assessed by two independent raters ($\kappa = 0.89$) across 50 generations per model.

Key Observation: MHA (LLaMA-2) shows immediate degradation under recursive stress despite lower noise sensitivity. GQA shows delayed degradation. Qwen2 is the only model showing no degradation (50 generations tested, 3/3 seeds).

Reconciling Findings 2 and 3: The divergent results suggest two distinct phenomena:

- **Random noise:** GQA amplifies through KV-sharing
- **Structured alignment pressure:** GQA's bottleneck may filter correlated signals differently

3.4 Finding 4: Training Methodology Correlates With Robustness

3.4.1 Synthetic Training (Microsoft Phi Family)

Model	Params	Arch	SI
Phi-1.5	1.3B	MHA	0.318
Phi-2	2.7B	MHA	0.344
Phi-3-mini	3.8B	MHA	0.329
Phi-3-medium	14B	GQA 4:1	0.334

Table 7: Phi family: SI ≈ 0.33 (± 0.02) across 10.8× scale range.

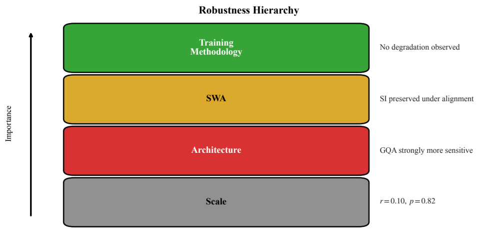

Figure 3: Robustness hierarchy: Training Methodology > SWA > Architecture > Scale. Phi family shows scale-invariant SI.

3.4.2 OMO Training (Alibaba Qwen2)

Qwen2 shows: minimal SI reduction (-9.0% vs -56.3% for LLaMA-3.1); no observed recursive degradation (50 generations); 13,139× lower noise sensitivity than LLaMA-3.

Correlation with OMO: Qwen2 uses Online Merging Optimizer (Lu et al., 2024), which merges gradients with SFT parameters at each step. We hypothesize this prevents parameter drift, but note this is correlational—no ablation study exists.

3.5 Finding 5: Perturbation Probe Differentiates SI States

Model	State	SI	Response
LLaMA-3.1-8B-Inst	Low SI	0.31	$+28.6\%$
LLaMA-2-7B-Base	Low SI	0.34	$+114\%$
Gemma-2-9B-IT	Normal SI	0.75	$+5.2\%$
Phi-3-mini	Low SI	0.33	-2%

Table 8: Perturbation probe results ($\sigma = 0.05$).

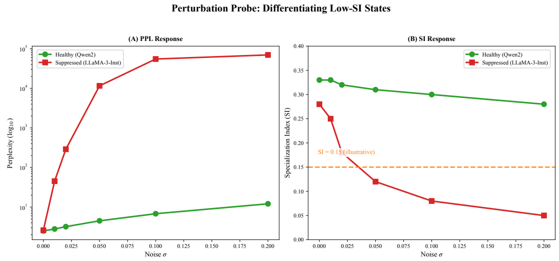

Figure 4: Perturbation probe differentiates pathological ($> 20\%$ response) from healthy low-SI states.

Interpretation: Large positive response ($> 20\%$): latent capacity suppressed by alignment. Neutral response ($\pm 5\%$): stable, healthy state. Negative response: already optimized, noise harmful.

3.6 Finding 6: Dimensional Crowding

Building on the dimensional crowding concept from D'Elia (2026b), we define effective crowding:

$$\rho_{\text{eff}} = \text{KV\_Ratio} \times \frac{H}{d_{\text{head}}} \tag{3}$$

Arch	Models	PPL-Slope	ρ_{eff}
MHA	LLaMA-2-7B	46	0.25
GQA	LLaMA-3-8B, Mistral-7B	3,026,733	1.0

Table 9: Dimensional crowding comparison (7B–8B scale).

Effect Size: GQA shows 65,799× higher noise sensitivity than MHA (Cohen's $d = 11.6$).

Heritage Override: Qwen2 has $\rho_{\text{eff}} = 1.53$ (highest tested) but remains STABLE—suggesting that training methodology may override geometric disadvantage.

4 Falsified Hypotheses

Hypothesis	Prior Evidence	Invalidating Evidence	Cause
MHA protects against SI collapse	LLaMA-2 +4.9% (float16)	LLaMA-2 -7.95% (bfloat16)	Precision artifact
$\rho_{\text{crit}} \approx 0.267$ threshold	Gemma-27B sign flip (8-bit)	Gemma-27B $+7.26\%$ (FP)	Quantization artifact
SWA breaks attention graph connectivity	Graph analysis	Giant component ≈ 1.0	Mechanism error
KV-compression universally ↑ robustness	Qwen2 correlation	Gemma-2 **inverse** correlation	Model-specific
SI correlates with spectral trace	Theoretical prediction	$r = 0.095$, $p = 0.82$	No support
L* formula predicts sink layer	Architectural derivation	10% pass rate	Heritage-specific

Table 10: Falsified hypotheses from prior work. Many effects are precision-dependent.

Methodological Lesson: Many reported effects in this domain are precision-dependent. bfloat16 or higher precision with multi-seed validation is essential.

5 Discussion

5.1 Summary of Findings

Our results suggest a robustness hierarchy:

1. **Training Methodology**
    - Synthetic data (Phi): SI invariant across scale
    - OMO training (Qwen2): Recursive stability
2. **Sliding Window Attention**
    - SWA models: Positive or neutral ΔSI
3. **Architecture (MHA/GQA)**
    - Minimal independent effect without (1) or (2)
4. **Scale**
    - No significant correlation with robustness

5.2 Limitations

1. **Scale:** All tests $\leq$27B parameters
2. **Causal Claims:** OMO and synthetic training attributions are correlational
3. **KV-Compression:** Effect validated only for Qwen2; Gemma-2 shows inverse pattern
4. **Recursive Degradation:** Involves subjective assessment ($\pm$3 generations variance)
5. **Vendor Coverage:** Closed-source frontier models not tested

5.3 Hypothesis: Frequency-Dependent Filtering

The finding that GQA amplifies random noise but shows resilience under structured recursive pressure suggests a frequency-dependent mechanism:

- **Random noise** (high-frequency, uncorrelated) → amplified by KV-sharing
- **RLHF-induced pressure** (low-frequency, correlated) → filtered by compression

This remains speculative without mechanistic validation.

6 Conclusion

We present evidence that preference optimization systematically reduces attention head specialization in most tested LLM families. The severity correlates primarily with training methodology:

- Synthetic training (Microsoft Phi) produces scale-invariant SI
- OMO training (Alibaba Qwen2) correlates with recursive stability
- SWA architecture correlates with SI preservation

The perturbation probe differentiates pathological from healthy low-SI states.

Central Finding: Alignment-induced structural changes are not uniform across models. Training methodology choices correlate more strongly with robustness than architectural details, suggesting these effects are modifiable through engineering decisions during pretraining.

Acknowledgments

Compute resources provided by Google Colab. We acknowledge the use of AI assistants for support with Python coding routines and LaTeX formatting. The conceptualization, experimental design, data analysis, and scientific conclusions are the sole work of the author.

References

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

Davide D'Elia. 2025. Uniformity asymmetry: An exploratory metric for detecting representational preferences in llm embeddings. Zenodo preprint.

Davide D'Elia. 2026a. Layer-wise embedding-output dynamics across llm families: Evidence for phase-structured decision commitment. Zenodo preprint.

Davide D'Elia. 2026b. Thermodynamic constraints in transformer architectures: A sheaf-theoretic perspective. Zenodo preprint.

Yihe Lin and 1 others. 2024. Mitigating the alignment tax of rlhf. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Xin Lu and 1 others. 2024. Online merging optimizers for boosting rewards and mitigating tax in alignment. *arXiv preprint arXiv:2405.17931*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.

A Standard Prompts

The following 10 prompts were used for all SI measurements (MD5: `715065bab181f46bf12ed471951141e2`):

1. "What is the capital of France and what is its population?"
2. "If all roses are flowers and some flowers fade quickly, can we conclude that some roses fade quickly? Explain step by step."
3. "Calculate 47 multiplied by 23 and show your work."
4. "Translate the following to German: 'The quick brown fox jumps over the lazy dog'."
5. "Write a Python function that checks if a number is prime."
6. "Summarize the main points: Machine learning is a subset of artificial intelligence..."
7. "Statement A: 'All birds can fly.' Statement B: 'Penguins are birds that cannot fly.' Are these statements contradictory? Explain."
8. "What are the safety considerations when using a kitchen knife?"
9. "Write a haiku about artificial intelligence."
10. "Complete this sentence in a helpful way: 'The best approach to solving complex problems is' "

B Seed-Level PPL-Slope Data

For transparency, we report individual seed results for the high-variance PPL-slope comparison (E06b Same-Scale):

Seed	LLaMA-2 (MHA)	Mistral (GQA)	Ratio
42	16.73	451,342	26,970×
123	156.96	391,573	2,495×
456	46.35	440,977	9,515×

Table 11: Seed-level PPL slopes and ratios.

Aggregations:

- Ratio-of-means (reported): 5,834×
- Median-of-ratios: 9,515×
- Mean-of-ratios: ∼13,000×

The ratio-of-means is preferred for heavy-tailed distributions as it reduces sensitivity to outlier seeds.

C Data Availability

All JSON result files available post-publication at: https://github.com/buk81/uniformity-asymmetry/tree/main/paper4