COMPASS: Context-Modulated PID Attention Steering System for Hallucination Mitigation

Kenji Sahay, Snigdha Pandya, Rohan Nagale, Anna Lin, Shikhar Shiromani, Parham Sharaf, Kevin Zhu, Sunishchal Dev

Algoverse · Georgia Institute of Technology · UC Berkeley

kenjisahay@gmail.com, snigdhapandya@gmail.com

Overview

Large language models often produce fluent but factually incorrect statements due to misallocation of attention between contextual inputs and parametric knowledge. We introduce **COMPASS**, a lightweight framework that dynamically steers attention to retrieved context during generation. Using the **Context Reliance Score (CRS)**, COMPASS identifies underutilizing attention heads, and a **PID controller** adjusts them in real time. Across HotpotQA, XSum, HaluEval, and RAGTruth, COMPASS reduces hallucinations by **2.8–5.8%** absolute while maintaining single-stream decoding.

1. Motivation & Contributions

Problem: LLMs produce *contextual hallucinations*—outputs conflicting with input context despite relevant evidence being present.

Limitations of Existing Methods:

- Post-hoc filtering requires multi-pass decoding
- Contrastive decoding lacks interpretability
- Re-ranking methods add significant latency

Our Contributions:

1. **COMPASS:** Decoding-time attention adjustment via pre-softmax, context-key-only bias
2. **CRS:** Online per-head context-sensitivity signal
3. **Classifier-Guided Scaling:** Modulate only when risk is elevated
4. **Single-Stream Efficiency:** No retraining or multi-pass decoding

2. Method: Context Reliance Score

We quantify each head's context reliance as attention mass on context keys:

$$p_{\text{ctx}}(t, \ell, h) = \sum_{i \in K_C} A_t(\ell, h)[i] \tag{1}$$

With logit transform for stability:

$$\text{CRS}(t, \ell, h) = \log \frac{\tilde{p}_{\text{ctx}}}{1 - \tilde{p}_{\text{ctx}}} \tag{2}$$

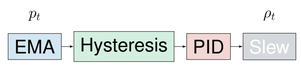
Features: Windowed statistics (mean, std, delta) over $W \in \{4, 8, 16\}$ tokens.

3. Hallucination Classifier

XGBoost classifier maps CRS features to risk $p_t \in [0, 1]$:

Model	Dataset	AUROC
Qwen-2.5-7B	HotpotQA	0.839
Qwen-2.5-7B	XSum	0.953
Qwen-2.5-7B	HaluEval	0.886
LLaMA-2-7B	RAGTruth	0.858
Mistral-7B	RAGTruth	0.912

4. PID Controller



Parameters: β=0.8 (EMA), h=0.01 (hysteresis), K_P=0.8, K_I=0.2, $\rho_{\max}$=1.0

The controller produces nonnegative log-gain ρ_t with anti-windup and slew limiting.

5. Attention Bias Mechanism

When risk is elevated, modify attention logits:

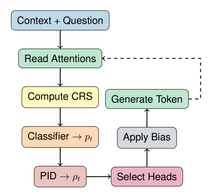
$$\tilde{Z}_t(\ell, h)[i] = Z_t(\ell, h)[i] + \rho_t \cdot a_\ell(h) \tag{3}$$

for context keys $i \in C$ only.

Properties:

- Only context keys biased
- Only last-query row modified
- Equivalent to multiplicative boost: $\exp(\rho_t \cdot a_\ell(h))$

6. System Architecture



Single-stream loop (every k tokens): Read attentions → CRS features → Risk prediction → PID update → Head selection → Apply bias

7. Experimental Setup

Models: LLaMA-2-7B/13B, Mistral-7B, Qwen-2.5-7B
Datasets: HotpotQA, XSum, HaluEval, RAGTruth
Labeling: Gemini 2.5-Flash (93% human agreement)
Configuration: Top-K=16 heads/layer, layers 16–31, λ=0.3 prior blend
Baselines: Unmodified, Lookback Lens, Contrastive Decoding, Random-head scaling

8. Results

Model	Dataset	MR↓	SD↓	CO↑
Qwen-2.5-7B	HotpotQA	4.2%	-14.2%	+0.06
Qwen-2.5-7B	XSum	2.8%	-11.4%	+0.04
Qwen-2.5-7B	RAGTruth	3.1%	-16.7%	+0.08
Qwen-2.5-7B	HaluEval	5.8%	-13.8%	+0.05
LLaMA-2-7B	RAGTruth	4.2%	-18.3%	+0.09
LLaMA-2-13B	RAGTruth	**5.8%**	**-22.4%**	**+0.12**
Mistral-7B	RAGTruth	4.9%	-20.1%	+0.11

MR: Mitigation Rate (absolute ↓) **SD:** Span Density **CO:** Context Overlap

Key Findings:

- Consistent 2.8–5.8% hallucination reduction
- Larger models benefit more from modulation
- Improved grounding (CO↑) without sacrificing fluency
- Single-pass decoding maintained

9. Method Comparison

Method	Real-time	Single-pass	Interpretable	No Retrain
Contrastive Dec.	✓	–	–	✓
Lookback Lens	✓	–	✓	✓
DAGCD	✓	–	–	✓
COMPASS	✓	✓	✓	✓

10. Limitations

- Short-horizon signals may miss gradual risk in long contexts
- Per-step gating lacks global discourse awareness
- Sensitive to PID hyperparameters
- Modest overhead on smaller GPUs

11. Conclusion

COMPASS demonstrates that **closed-loop feedback control** on attention logits can steer outputs toward contextually supported tokens:

$$\tilde{Z}_t(\ell, h)[i] = Z_t(\ell, h)[i] + \rho_t \cdot a_\ell(h), \quad i \in C$$

Takeaways: Lightweight, interpretable mitigation achieving 2.8–5.8% reduction without retraining. Control-theoretic methods show promise for LLM alignment.

References

[1] Chuang et al. "Lookback Lens" EMNLP'24 [2] Huang et al. "DAGCD" ACL'25 [3] Shi et al. "Context-Aware Decoding" arXiv'23 [4] Vaswani et al. "Attention Is All You Need" NeurIPS'17 [5] Voita et al. "Multi-Head Attention" ACL'19 [6] Åström & Murray "Feedback Systems" 2008

Contact: snigdhapandya@gmail.com, kenjisahay@gmail.com