hallucination-mitigation-via-contrastive-sampling-method

Anonymous Author(s)
Affiliation
Address
email

Abstract

Hallucination in large language models remains a major obstacle to reliable real-world deployment. Although preference-based learning improve factuality, sample-level preference misalignment can persist after alignment, with the model assigning excessive relative support to an incorrect continuation even when a correct continuation remains weakly represented internally. We address this by formulating hallucination mitigation as a selective contrastive post-training problem driven by the model's own internal signals. A frozen base model generates a counterfactual bad trajectory, and the adapted model contrasts gold and bad continuations from their first point of divergence, applying updates only when the incorrect branch is not yet sufficiently suppressed. This selective self-gating concentrates learning on hallucination-active and plausibly recoverable cases, avoiding unnecessary updates while requiring no expert annotation, prompt optimization, or manual data curation. Despite using only about 10% of the effective training data, the method consistently outperforms baseline training, and DPO, while also showing robust gains under out-of-distribution settings. The resulting framework provides a lightweight and scalable form of selective preference repair, improving factuality without requiring the heavy supervision and optimization overhead typical of standard preference learning pipelines.

1 Introduction

Large language models (LLMs) have achieved strong performance across a wide range of language understanding and generation tasks , yet hallucination remains a major obstacle to reliable deployment in real-world settings Devlin et al. [2019], Radford et al. [2019], Lewis et al. [2019]. Models can produce fluent and confident outputs that are unsupported, contradictory, or simply incorrect Ji et al. [2023], Lin et al. [2022], Ravi et al. [2024], which is especially problematic in question-answering scenarios where factual correctness matters more than stylistic quality. Improving reliability without introducing expensive or overly complex post-training pipelines has therefore become a central challenge for modern LLM systems.

A common approach to improving factuality is to introduce stronger supervision through alignment objectives, preference optimization, or curated training data [Lin et al., 2024a, Ouyang et al., 2022]. These methods have led to substantial progress, but sample-level preference misalignment can remain after alignment. In such cases, the problem is not always a complete absence of knowledge or reasoning ability. Instead, the model may still retain enough internal support for a correct continuation while assigning excessive relative support to a competing incorrect one during decoding. This suggests that a meaningful subset of hallucinations can be understood as failures of internal continuation preference rather than pure capability failures.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Motivated by this view, we formulate hallucination mitigation as a selective contrastive post-training problem driven by the model's internal signals. Rather than relying on external teachers or synthetic corrections, training uses two instances of the same model: a frozen base model generates a counterfactual bad trajectory, and an adapted model compares that bad continuation with the gold continuation. We identify the first token position at which the two answers diverge and perform the comparison from that point onward. This makes the training signal local to the point where generation begins to deviate and turns the model's own mistaken continuation preference into a usable correction signal.

Our method is selective in that the auxiliary confidence objective is activated only when the wrong branch is not sufficiently suppressed relative to the gold branch. The resulting objective increases the relative support of the gold branch over the bad branch only on examples that exhibit this preference misalignment. This avoids treating all training examples as equally in need of correction and concentrates updates on hallucination-active cases that appear plausibly recoverable through better use of the model's existing competence. In this sense, the method can also be viewed as a form of model-based selective sampling, since the model's own behavior helps determine which examples are informative enough to receive additional corrective updates.

This perspective leads to a lightweight and practical training pipeline. We implement the method as a partial-layer adaptation scheme that modifies only a masked subset of rows within a single transformer layer. The goal is not to broadly re-optimize the model through large-scale additional supervision, but to selectively repair misaligned continuation preferences where the model already appears capable of reaching the correct answer. Because updates are applied only on selectively activated examples, the amount of data that actually contributes to parameter updates can be substantially smaller than in standard full fine-tuning pipelines.

We therefore ask the following question: when a model already contains enough internal support to reach the correct continuation, can hallucination be reduced by selectively correcting cases in which it prefers a wrong continuation too strongly? We argue that the answer can be yes. By using base-generated counterfactual bad trajectories, comparing continuations from their first divergence onward, and activating auxiliary correction only where continuation preference is misaligned, our framework provides a simple and data-efficient alternative to broad full-data adaptation. Empirically, the proposed approach consistently outperforms full fine-tuning, baseline training, and DPO while using only about 10% of the effective training data. It requires no expert annotation, prompt optimization, or manual data curation, as all training signals are derived from the model itself, and maintains robust improvements under out-of-distribution settings despite its lightweight and computationally efficient design.

Our contributions are as follows:

- We formulate hallucination mitigation as a selective contrastive post-training problem driven by the model's internal signals, where sample-wise self-gating determines whether updates should be applied toward the gold continuation.
- We introduce a divergence-aware objective that compares gold and bad continuations from their first point of divergence and selectively strengthens the gold branch only when the incorrect continuation remains insufficiently suppressed, focusing learning on hallucination-active preference failures.
- We present a lightweight partial-layer adaptation framework that achieves consistently stronger performance than full fine-tuning, baseline training, and DPO while using only about 10% of the effective training data, requiring no expert annotation or data curation, and maintaining robust improvements under out-of-distribution settings with significantly lower computational overhead.

2 Related Work

2.1 Internal Knowledge, Misalignment, and Model-Centric Signals

Recent studies argue that large language models can hallucinate even when they already possess the relevant parametric knowledge, because they fail to reliably elicit that knowledge at generation time Zhang et al. [2024], Jiang et al. [2024], Zhang et al. [2025]. This can be understood as a form of

misalignment in which a plausible correct continuation exists internally, yet an incorrect continuation is selected with relatively stronger support. However, at the level of individual samples, it is difficult to determine from the outside whether an error arises from a genuine lack of knowledge or from a failure of internal alignment despite the presence of the relevant knowledge. In response, a recent line of hallucination mitigation research has increasingly focused on reducing errors by leveraging the model's own internal capabilities, rather than injecting additional external criteria Zhang et al. [2024], Wen et al. [2025], Liang et al. [2024].

2.2 Selective Fine-Tuning for Hallucination Repair

Uniformly fine-tuning on all available data is not always a safe strategy for mitigating hallucinations. In fact, several studies report that fine-tuning on new knowledge or on examples unfamiliar to the model can even increase hallucination tendencies Lin et al. [2024b], Gekhman et al. [2024], Dang et al. [2026], Huang et al. [2025], Wan et al. [2024]. Even factually grounded knowledge data can disrupt the model's prior beliefs or weaken its confidence in previously acquired knowledge. These findings suggest that simply increasing the absolute amount of seemingly high-quality data, from a human perspective, may itself raise the risk of hallucination. Prior work on factuality alignment has introduced approaches such as factuality-aware preference learning Tian et al. [2023] and fine-grained supervision Huang and Chen [2024], but these methods generally rely on response-level preferences, factuality labels Chaduvula et al. [2026], or sentence-level masks Gu et al. [2025]. They do not directly study a setting in which the loss is activated only when a bad branch produced by the base model is not sufficiently suppressed.

2.3 Abstention-Based Hallucination Reduction and Its Limits

To reduce hallucinations, several studies have explored training language models to abstain from answering or to respond with statements such as "I don't know" Huang et al. [2025], Lin et al. [2024b], Kim et al. [2025]. While such approaches can lower hallucination rates, they also have an important limitation: because they explicitly train abstention behavior, the model may respond conservatively even in cases where it actually possesses the relevant knowledge. In other words, these methods do not fully solve the more fine-grained problem of teaching the model to distinguish between what it truly does not know and what it does know, nor do they calibrate how strongly that knowledge should be expressed.

2.4 Learning from Good–Bad Pairs Without Direct Internal Support Comparison

Several prior methods learn from preferred–rejected response pairs Chaduvula et al. [2026], Tian et al. [2023], but they do not construct gold and bad branches from within the same model and directly compare the model's support for each. Moreover, a preferred response in such datasets does not necessarily correspond to the true answer, meaning that these methods do not directly transfer support from an incorrect answer to a correct one, but instead optimize relative comparisons among candidate responses. By contrast, our framework uses the difference between sample-specific good and bad pairs as a direct training signal.

3 Method

3.1 Problem Formulation

Given an input x (question with optional reference) and a gold answer y^*, standard supervised fine-tuning optimizes the likelihood of y^* under the model. However, hallucination often arises not from missing knowledge, but from *preference misalignment*, where the model assigns excessive relative support to an incorrect continuation despite retaining sufficient internal support for the correct one.

We address this by explicitly modeling hallucination as a *selective preference failure* between a correct continuation y^* and a competing incorrect continuation y^-. This perspective shifts the objective from learning new knowledge to *rebalancing the model's internal preference between competing continuations*.

3.2 Self-Generated Counterfactual Negatives

Instead of relying on external annotations or synthetic data, we construct negative examples using the model itself. A frozen base model θ_0 generates a counterfactual answer:

$$y^- \sim p_{\theta_0}(y \mid x), \tag{1}$$

which serves as a model-consistent incorrect continuation.

This design has two advantages. First, it avoids the need for curated preference pairs or external supervision, making the method fully self-contained. Second, because the negative examples are sampled from the model's own distribution, they naturally reflect realistic failure modes encountered at inference time.

3.3 Construction of Gold and Bad Continuations

For each training example, we start from an input prompt x constructed from the question and, when available, its accompanying reference context. The *gold continuation* y^* is taken directly from the dataset-provided ground-truth answer. We concatenate the prompt and gold answer to form the supervised target sequence:

$$\mathbf{z}^* = [x; y^*]. \tag{2}$$

The *bad continuation* y^- is produced by a frozen copy of the same base model, using exactly the same prompt x:

$$y^- = \mathrm{Decode}(p_{\theta_0}(\cdot \mid x)). \tag{3}$$

In practice, decoding is performed deterministically so that the generated bad continuation reflects the model's current default preference under a fixed inference procedure, rather than sampling noise.

This construction is important for two reasons. First, both y^* and y^- are conditioned on the same input, so the comparison isolates errors in continuation preference rather than differences in prompt formulation. Second, because y^- is generated by the model itself rather than manually written or externally curated, it reflects realistic failure modes that the model is likely to exhibit at inference time.

After generating y^-, we compare it against y^* at the token level. If the two continuations are identical or do not exhibit insufficient suppression of the incorrect branch, the example contributes little or no corrective signal under our self-gating rule. As a result, the method focuses naturally on examples where the model's own generated continuation reveals an actionable preference failure.

3.4 Divergence-Aware Local Comparison

We identify the first token position at which the gold and generated answers diverge:

$$t^* = \min\{t \mid y_t^* \neq y_t^-\}. \tag{4}$$

Rather than optimizing over the entire sequence, we restrict the training signal to tokens from t^* onward. This localizes the objective to the point where generation begins to deviate, focusing learning on hallucination-relevant segments while preserving correct prefixes.

3.5 Token-Level Objective Construction

Given an input x, we construct concatenated sequences:

$$\mathbf{z}^* = [x; y^*], \quad \mathbf{z}^- = [x; y^-]. \tag{5}$$

We compute next-token cross-entropy:

$$\ell_t = -\log p_\theta(z_t \mid z_{<t}), \tag{6}$$

for each token position.

To focus on hallucination-relevant regions, we compute losses only on the suffix:

$$\mathcal{T}^* = \{t \mid t \geq t^*\}, \tag{7}$$

$$\mathcal{L}_{gold} = \frac{1}{|\mathcal{T}^*|} \sum_{t \in \mathcal{T}^*} \ell_t^{(gold)}, \quad \mathcal{L}_{bad} = \frac{1}{|\mathcal{T}^*|} \sum_{t \in \mathcal{T}^*} \ell_t^{(bad)}. \tag{8}$$

This ensures that the model is trained only on the portion of the sequence where the gold and incorrect continuations differ.

3.6 Selective Contrastive Objective

We define a preference margin:

$$\Delta = \mathcal{L}_{bad} - \mathcal{L}_{gold}. \tag{9}$$

A large positive margin indicates correct preference, while a small or negative margin indicates insufficient suppression of the incorrect continuation.

We introduce a self-gating mechanism:

$$\mathcal{L} = \mathbb{I}[\Delta < \tau] \cdot (\mathcal{L}_{gold} - \mathcal{L}_{bad}), \tag{10}$$

where τ is a threshold.

This ensures that updates are applied only when the model fails to sufficiently prefer the correct continuation.

3.7 Sample Selection via Self-Gating

The training data is implicitly partitioned based on Δ:

- **Well-aligned:** $\Delta \gg \tau$ (no update)
- **Ambiguous:** $\Delta \approx \tau$ (optional)
- **Hallucination-active:** $\Delta < \tau$ (used for training)

Thus, the effective training set is:

$$\mathcal{D}_{active} = \{(x, y^*, y^-) \mid \Delta(x) < \tau\}. \tag{11}$$

Importantly, this set is not predefined but determined dynamically by the model itself.

3.8 Partial-Layer Selective Adaptation

We restrict training to a subset of hidden dimensions within a single transformer layer.

Let $S \subseteq \{1, \ldots, d\}$ denote selected hidden dimensions. Updates are applied as:

$$\theta_{l,S} \leftarrow \theta_{l,S} - \eta \nabla \mathcal{L}, \tag{12}$$

while all other parameters remain frozen.

This design assumes hallucination arises from localized preference distortions, enabling efficient and stable adaptation.

3.9 Training Dynamics and Interpretation

The proposed method can be interpreted as *selective preference repair*. The model updates only when it is insufficiently confident in the correct answer, leading to the following behavior:

- correctly aligned samples are ignored,
- hallucination-active samples are emphasized,
- updates focus on recoverable errors.

Because only a subset of samples satisfy $\Delta < \tau$, the effective training size is significantly reduced. In practice, only about 10% of samples contribute to updates.

Dataset	Method		New (Ours)		Comparison	
	Rate	Fails	Rate	Fails	Δ	Eval Size
(a) DPO vs. Ours						
DROP	0.4835 (4245/8780)	755	0.4327 (4126/9535)	0	**-0.0508**	9,535
FaithEval	0.2319 (282/1216)	284	0.1707 (256/1500)	0	**-0.0612**	1,500
HaluEval (dialogue)	0.8908 (8685/9750)	250	0.8384 (8383/9999)	1	**-0.0524**	10,000
HaluEval (summ.)	0.6117 (5601/9157)	843	0.5428 (5427/9999)	1	**-0.0689**	10,000
HotpotQA (fullwiki)	0.6021 (3770/6261)	1,144	0.5155 (3816/7402)	3	**-0.0866**	7,405
Average	0.5640	–	0.5000	–	**-0.0640**	–
(b) SFT vs. Ours						
DROP	0.4379 (4175/9535)	0	0.4327 (4126/9535)	0	**-0.0052**	9,535
FaithEval	0.1793 (269/1500)	0	0.1707 (256/1500)	0	**-0.0086**	1,500
HaluEval (dialogue)	0.8696 (8696/10000)	0	0.8384 (8383/9999)	1	**-0.0312**	10,000
HaluEval (summ.)	0.5611 (5611/10000)	0	0.5428 (5427/9999)	1	**-0.0183**	10,000
HotpotQA (fullwiki)	0.5201 (3850/7402)	3	0.5155 (3816/7402)	3	**-0.0046**	7,405
Average	0.5136	–	0.5000	–	**-0.0136**	–

Table 1: Hallucination rates across multiple benchmarks under three comparison settings: (a) DPO vs. Ours, (b) SFT with CE loss only vs. Ours. The Δ column reports the improvement of our method over each baseline (negative values indicate a lower hallucination rate). Data usage is approximately 10% of the full dataset. Values are rounded to the fifth decimal place.

Table 2: Perplexity comparison across evaluation datasets. Lower is better. Δ denotes the change relative to the baseline on the same dataset split.

Dataset	**Split**	**Model**	**PPL** $\downarrow$	Δ **vs. Baseline**
WikiText-2	test	Qwen2.5-7B-Instruct (baseline)	12.02	0.00
		Ours	**11.99**	-0.03
		DPO	12.07	+0.05
		SFT with CE loss only	11.99	-0.03
C4	validation	Qwen2.5-7B-Instruct (baseline)	XX.XX	0.00
		Ours	XX.XX	X.XX
		DPO	XX.XX	X.XX
		SFT with CE loss only	XX.XX	X.XX

4 Experiments

4.1 Benchmark Evaluation Across Multiple Datasets

We first evaluate the generalization performance of different training methods across five benchmark datasets, including DROP, FaithEval, HaluEval (dialogue and summarization), and HotpotQA. These datasets cover a diverse range of settings, including multi-hop reasoning, long-context understanding, and dialogue-based generation.

As shown in Table 1, DPO achieves relatively strong performance on FaithEval but exhibits substantially higher hallucination rates on more complex benchmarks such as HaluEval dialogue and HotpotQA. This suggests that uniform preference optimization struggles to generalize under distribution shift, particularly in settings requiring long-context reasoning or multi-turn consistency.

4.2 Selective Contrastive Training under Fixed Adaptation Setting

We next evaluate the proposed method under a fixed adaptation setting. All runs are trained on HaluEval-QA and evaluated on the DROP validation set. The trainable layer is fixed, and we vary the key hyperparameters controlling the selective contrastive objective: the bad-branch target b, the probability-margin target p, and the hallucination threshold h.

Method	b	p	h	Usage	Step	Halluc.	Improved	Worsened	Net
SFT-only	–	–	–	**100%**	200	0.4413	217	144	73
SFT-only	–	–	–	**100%**	400	0.4406	278	198	80
New	1.2	0.0	0.0	**0.0970%**	200	0.4397	184	96	88
New	1.2	0.0	0.0	**0.0970%**	400	0.4386	273	174	99
New	1.1	0.0	0.0	**0.0972%**	200	0.4405	191	113	78
New	1.1	0.0	0.0	**0.0972%**	400	0.4405	269	191	78
New	1.0	0.0	0.0	**0.0977%**	200	0.4414	178	106	72
New	1.0	0.0	0.0	**0.0977%**	400	0.4385	271	171	100
New	1.0	-0.1	-0.1	**0.0955%**	200	0.4418	170	102	68
New	1.0	-0.1	-0.1	**0.0955%**	400	0.4390	269	173	96
New	0.9	0.0	0.0	**0.0977%**	200	0.4414	177	105	72
New	0.9	0.0	0.0	**0.0977%**	400	0.4388	268	171	97
New	0.8	0.0	0.0	**0.0981%**	200	0.4405	179	98	81
New	0.8	0.0	0.0	**0.0981%**	400	0.4395	259	169	90
New	0.7	0.0	0.0	**0.0980%**	200	0.4416	171	101	70
New	0.7	0.0	0.0	**0.0980%**	400	0.4395	260	170	90

Table 3: Selective contrastive training under different gating configurations. b controls bad-branch suppression strength, p controls the margin constraint, and h controls hallucination activation. Usage denotes the fraction of samples used for training. Improved, worsened, and net denote the number of samples that changed relative to the untrained baseline model, Qwen-2.5-7B-Instruct. All configurations use the same trainable layer setting, with only layer 20 updated.

These hyperparameters directly control how the objective behaves at the sample level. The bad-branch target b determines the desired level of suppression for the incorrect continuation. Larger values enforce stronger penalization of the bad branch. The probability-margin target p defines an additional margin constraint between gold and bad continuations, while the hallucination threshold h determines whether a sample is considered hallucination-active and therefore used for training.

Unlike SFT-only, which applies cross-entropy updates uniformly to all samples, the proposed method activates updates only on a small subset of samples. Across all configurations, the effective usage ratio remains around 9.5%–9.8%, confirming that the method operates in a highly selective regime.

Table 3 shows that the proposed method consistently achieves stronger net improvements than SFT-only despite using only about one-tenth of the training data. At step 400, the best configuration ($b = 1.0$) achieves a net gain of 100, compared to only 80 for SFT-only.

A key observation is that the proposed method improves the correction-to-damage trade-off. While SFT-only improves slightly more samples, it also introduces significantly more regressions. In contrast, the selective method reduces the number of worsened predictions, leading to higher net gains.

The results also show that performance is robust across a range of b values, with the strongest results occurring near $b \in [1.0, 1.2]$. More restrictive gating (negative p and h) reduces the number of active samples but may slightly degrade performance by excluding recoverable cases.

Overall, these results demonstrate that selectively updating hallucination-active samples is more effective than uniformly applying cross-entropy updates, yielding better factuality with significantly lower data usage.

5 Conclusion

We proposed a selective contrastive post-training framework that reframes hallucination mitigation as a preference misalignment problem. By selectively updating the model only when incorrect continuations are not sufficiently suppressed relative to correct ones, our approach avoids unnecessary global fine-tuning and focuses learning on hallucination-critical regions.

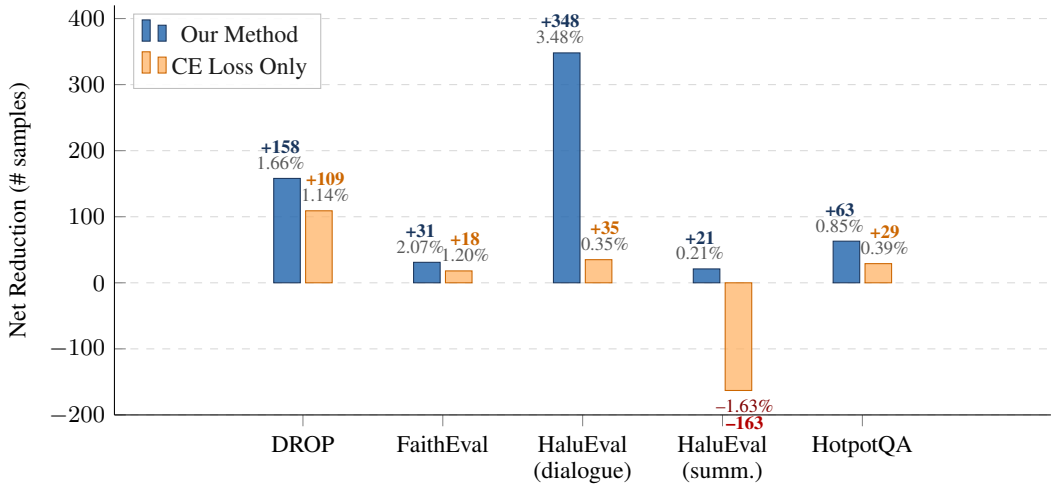

Figure 1: Net reduction in hallucinated samples compared to the untrained Qwen-2.5-7B-Instruct baseline (i.e., without any fine-tuning). Bold numbers denote how many fewer (or more, if negative) samples were hallucinated after training; gray values denote the corresponding decrease (or increase) in hallucination rate in percentage points. Our method (blue) consistently reduces hallucinations across all benchmarks, while SFT with CE loss only (orange) shows mixed results—notably increasing hallucinations by 163 samples (1.63 pp) on HaluEval summarization. DROP, FaithEval, and HotpotQA are fully out-of-distribution (cross-benchmark); HaluEval dialogue and summarization are cross-task within the same benchmark family as the training set (HaluEval QA).

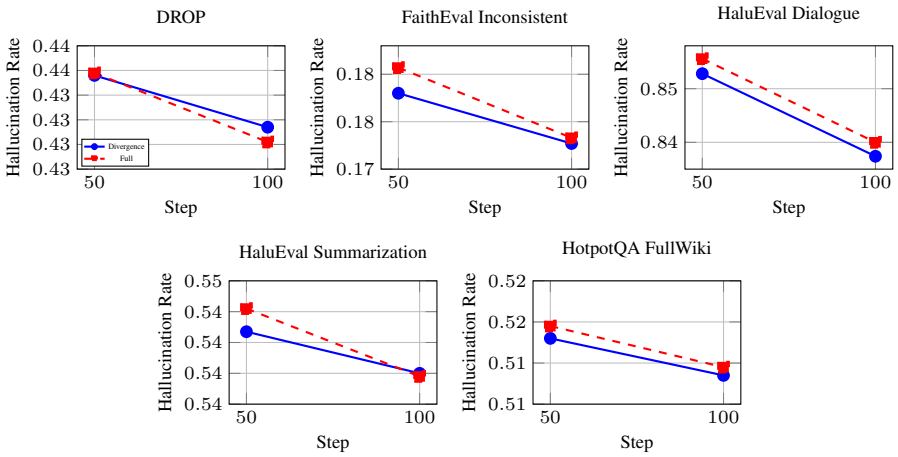

Figure 2: Dataset-wise comparison between divergence-based training and full non-divergence training.

Across multiple benchmarks, our method consistently improves factuality while using only a small fraction of the training data, highlighting its strong data efficiency. In contrast to standard pipelines such as supervised fine-tuning and preference optimization, our approach does not rely on large-scale human annotations or expensive reference model optimization.

Despite requiring only modest additional computation, the proposed framework provides a lightweight and scalable alternative for improving model reliability. These results suggest that targeted, model-driven preference repair offers a promising direction for addressing persistent hallucinations in large language models.

6 limitation

Our method has several limitations. Because training compares answer-only outputs, divergence is defined only over the final answer segment and cannot directly correct earlier reasoning errors. In addition, the method depends on the quality of the base model's bad branch and the reliability of semantic mismatch judgment, both of which can introduce weak or noisy supervision.

References

Sindhuja Chaduvula, Ahmed Y. Radwan, Azib Farooq, Yani Ioannou, and Shaina Raza. Reducing hallucinations in llms via factuality-aware preference learning, 2026.

Renfei Dang, Peng Hu, Zhejian Lai, Changjiang Gao, Min Zhang, and Shujian Huang. Understanding new-knowledge-induced factual hallucinations in llms: Analysis and interpretation, 2026. URL `https://arxiv.org/abs/2511.02626`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7765–7784, 2024.

Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. Mask-dpo: Generalizable fine-grained factuality alignment of llms, 2025.

Chao-Wei Huang and Yun-Nung Chen. Factalign: Long-form factuality alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16363–16375, 2024.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yuxuan Gu, Yangfan Ye, Liang Zhao, Weihong Zhong, Baoxin Wang, Dayong Wu, Guoping Hu, Lingpeng Kong, Tong Xiao, Ting Liu, and Bing Qin. Alleviating hallucinations from knowledge misalignment in large language models via selective abstention learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24564–24579, 2025.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. ISSN 1557-7341. doi: 10.1145/3571730. URL `http://dx.doi.org/10.1145/3571730`.

Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. On large language models' hallucination with regard to known facts. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1041–1053, 2024.

Hyuhng Joon Kim, Youna Kim, Sang-goo Lee, and Taeuk Kim. When to speak, when to abstain: Contrastive decoding with abstention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9710–9730, 2025.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. URL `https://arxiv.org/abs/1910.13461`.

Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. Learning to trust your feelings: Leveraging self-awareness in LLMs for hallucination mitigation. In *Proceedings of the 3rd Workshop on Knowledge Augmented Methods for NLP*, pages 44–58, 2024.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. Flame : Factuality-aware alignment for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 115588–115614, 2024a.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. Flame: Factuality-aware alignment for large language models. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024).*, 2024b.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, 2022.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL `https://arxiv.org/abs/2203.02155`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI, 2019.

Selvan Sunitha Ravi, Bartosz Mielczarek, Anand Kannappan, Douwe Kiela, and Rebecca Qian. Lynx: An open source hallucination evaluation model, 2024.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. Fine-tuning language models for factuality. In *Workshop on Neural Information Processing Systems (NeurIPS 2023).*, 2023.

Fanqi Wan, Xinting Huang, Leyang Cui, Xiaojun Quan, Wei Bi, and Shuming Shi. Knowledge verification to nip hallucination in the bud. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2616–2633, 2024.

Xueru Wen, Jie Lou, Xinyu Lu, Yuqiu Ji, Xinyan Guan, Yaojie Lu, Hongyu Lin, Ben He, Xianpei Han, Debing Zhang, and Le Sun. On-policy self-alignment with fine-grained knowledge feedback for hallucination mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5215–5231, 2025.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1965, 2024.

Yuji Zhang, Sha Li, Cheng Qian, Jiateng Liu, Pengfei Yu, Chi Han, Yi R. Fung, Kathleen McKeown, ChengXiang Zhai, Manling Li, and Heng Ji. The law of knowledge overshadowing: Towards understanding, predicting and preventing LLM hallucination. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23340–23358, 2025.

7 Appendix