ContrastSFT: Contrastive Logit Regularization Supervised Fine-Tuning for Mitigating Hallucinations in Large Language Models

Jingzhao Gu[1†[0009-0000-1054-8377]], Hongmin Xiao[1†[0000-0002-5645-1754]], Hangyu Li[1[0009-0007-6707-8266]] and Qiwei Wang[2[0009-0007-4309-4309]]

[1] National University of Singapore, 21 Lower Kent Ridge Rd, Singapore 119077, Singapore
[2] Harbin Institute of Technology, Shenzhen, University Town of Shenzhen, Nanshan District, Shenzhen 518055, China
† These authors contributed equally to this work.

Abstract. Large Language Models (LLMs) have achieved remarkable success in natural language generation but remain prone to hallucinations—generating content that is fluent but factually incorrect. While recent inference-time interventions like Contrastive Decoding (CD) effectively mitigate this by penalizing tokens favored by a "weak" hallucination-prone model, they introduce significant computational overhead (doubling inference latency) and fail to permanently align the model. To mitigate hallucinations in large language models (LLMs), we propose ContrastSFT, a training framework that incorporates contrastive decoding principles directly into model optimization. Specifically, we introduce Contrastive Logit Regularization (CLR), which modifies the standard Supervised Fine-Tuning (SFT) objective. Instead of uniformly increasing the likelihood of target tokens, our method penalizes token probabilities by subtracting the log-probabilities produced by a weaker baseline model. This design discourages the model from relying on superficial patterns that are often associated with hallucinations, and encourages the learning of more robust and factually grounded representations. We evaluate the proposed approach on multiple benchmarks, including natural language understanding tasks (WiCE, ParaRel) and factuality-oriented datasets (MMLU, HaluEval). Across these settings, ContrastSFT achieves consistent improvements of 5%–9% over standard SFT and existing contrastive approaches. Notably, since the contrastive behavior is integrated during training, our method does not require auxiliary models at inference time, and thus maintains the computational efficiency of standard LLMs. The implementation will be publicly released.

Keywords: large language models, hallucination mitigation, contrastive learning, supervised fine-tuning, contrastive decoding, factuality evaluation

1 INTRODUCTION

Large language models (LLMs) have recently demonstrated strong performance across a wide range of natural language processing tasks [1,2]. Their training typically follows a multi-stage pipeline, including pretraining (PT), supervised fine-tuning

(SFT), and reinforcement learning (RL) [3-9]. Despite this progress, hallucination remains an ongoing problem: models routinely output fluent but factually inaccurate or unsupported content [10].

However, one of the limitations is the SFT objective used in the state-of-the-art based on Maximum Likelihood Estimation (MLE). Although MLE induces the model to give a high probability to ground-truth tokens, it does not specifically penalize plausible but incorrect alternatives. Hence, the learned distribution will still seem to support erroneous correlations or memorized patterns rather than grounded reasoning.

In order to solve the issue without tuning training set, inference-time methods like Contrastive Decoding (CD) [11] have been offered. CD improves factuality by juxtaposing an expert model against a poor reference model, reducing tokens which are more likely under the latter. This strategy is effective, but has practical negative impacts. First, it involves using two models at once during decoding, which brings latencies and memory overhead increase to the computer system. Second, since CD only operates at inference time, it does not change the model parameters, leaving the model's intrinsic susceptibility to hallucination in which it operates little.

New alignment methods, specifically Direct Preference Optimization (DPO) [12], try to integrate contrastive signals directly into training. These approaches, however, are based on big preference data, which can be expensive to acquire and especially hard to maintain close observation of fine-grained rules of factuality.

In this work, we present ContrastSFT (Fig. 1), a training framework that embeds contrastive signals into the SFT objective without needing preference annotations. The core component, Contrastive Logit Regularization (CLR), leverages a frozen weak model during training to reshape the optimization objective. Rather than treating all non-target tokens the same, CLR down-weights tokens that are strongly preferred by the weak model, because they are frequently related to hallucination-prone patterns. This makes the model distinguish the correctness and error of the output more reliably. We evaluate our method on both natural language understanding benchmarks (ParaRel, WiCE) and factuality-based tasks (HaluEval, MMLU). Across these settings, ContrastSFT shows that it consistently outperforms standard SFT and previous contrastive methods with 5–9% improvement.

The key contribution can be summarized as:

- **ContrastSFT:** Contrastive Logit Regularization (CLR) is employed in this work as a way to bring the intuition of contrastive decoding into the training objective without the need of paired preference data.
- **Inference Efficiency:** No additional model is required at inference time, which is the same as in classical SFT.
- **Empirical Superiority:** A large number of experiments indicate that ContrastSFT enhances Factuality without sacrificing general reasoning performance and avoids the trade-offs observed in post-hoc methods.

2 Related Work

2.1 Hallucination Mitigation

Some of the existing treatment approaches to combat hallucinations generally fall into a three-category grouping, namely, data-centric, training-based, and decoding-time methods, but the categories are frequently not very clear as to which data to focus on.

Data-driven hallucination mitigation. One specific area of effort is to enhance training data quality and trustworthiness. Thus, as an example, curated large scale corpora such as The Pile [13] has been suggested to offer cleaner and more varied supervision. For instance, previous studies demonstrate that by increasing the percentage of factual or structured sources (e.g. those based on the textbook), the tendency to hallucinate will decrease during pretraining [14], [15]. Similarly, knowledge editing approaches work directly to modify or improve specific aspects of model knowledge so that improvements are applied directly to the content of the model without having to retrain the full model [16], [17]. Retrieval-augmented generation (RAG) can help to supplement these approaches [18]–[20]: it integrates external evidence into inference time, which aids in bridging knowledge gaps and provides a solid factual foundation.

Training-induced hallucination mitigation. Beyond data quality, several works investigate how training procedures themselves can be adapted to improve factuality. For instance, training objectives that explicitly prioritize factually consistent outputs have been shown to reduce hallucinations [21]. There is also growing interest in leveraging LLMs to assist data annotation, where models help identify subtle inconsistencies or errors in training corpora, leading to cleaner supervision signals [22], [23].

Another line of research targets the decoding stage, where hallucinations often manifest during generation. Methods in this category typically modify the sampling or refinement process. For example, factuality-aware sampling strategies dynamically adjust decoding parameters to bias the model toward more reliable outputs [21]. Post-editing approaches further refine generated text after an initial pass, improving factual consistency without relying on external retrieval systems [24].

2.2 Contrastive Decoding

Contrastive decoding (CD) has emerged as an effective strategy for improving factuality by comparing token probabilities across different sources. Early work demonstrated that contrasting outputs from models of different capacities can suppress hallucination-prone tokens [25]. Subsequent studies extend this idea in several directions. Some approaches refine the source of contrastive signals, for example by selecting logits from different layers within a single model [26], or by incorporating knowledge distillation to guide models toward more consistent reasoning patterns [27]. Others focus on improving the robustness of the contrastive process itself, showing that such techniques can reduce shallow copying and missing reasoning steps during generation [28]–[30], or explicitly penalize hallucination-inducing patterns during decoding [31]. Additional variants further integrate information from intermediate representations and introduce

auxiliary objectives to enhance truthfulness [10]. Despite these advances, existing CD-based methods remain fundamentally tied to inference-time intervention.

2.3 Model Editing

Model editing provides an alternative perspective by directly modifying model parameters to correct factual errors or update outdated knowledge. A range of techniques has been proposed, from prompt-based editing that leverages in-context learning [32], to parameter-efficient approaches that introduce auxiliary networks for localized updates [33]. More recent methods explore structured parameter modifications, such as injecting additional neurons [34] or identifying and adjusting key activations associated with factual knowledge [35]. These ideas have been further extended to large-scale settings, enabling simultaneous updates to thousands of knowledge associations [36]. Alongside these developments, recent studies have also examined the extent to which factual and commonsense knowledge can be localized within editable model components [37].

2.4 Preference Optimization and Alignment

Beyond SFT, modern alignment methods are often based on Reinforcement Learning from Human Feedback (RLHF) [38] and related formulations. In particular, Direct Preference Optimization (DPO) [12] incorporates contrastive signals into training by directly optimizing model outputs toward preferred responses. While effective, such approaches depend heavily on large-scale paired preference data, which can be costly to obtain. In contrast, our proposed ContrastSFT leverages a frozen weak model to generate token-level negative signals automatically, enabling contrastive learning without explicit preference annotations. This design provides a more data-efficient alternative, especially in scenarios where fine-grained factual supervision is limited.

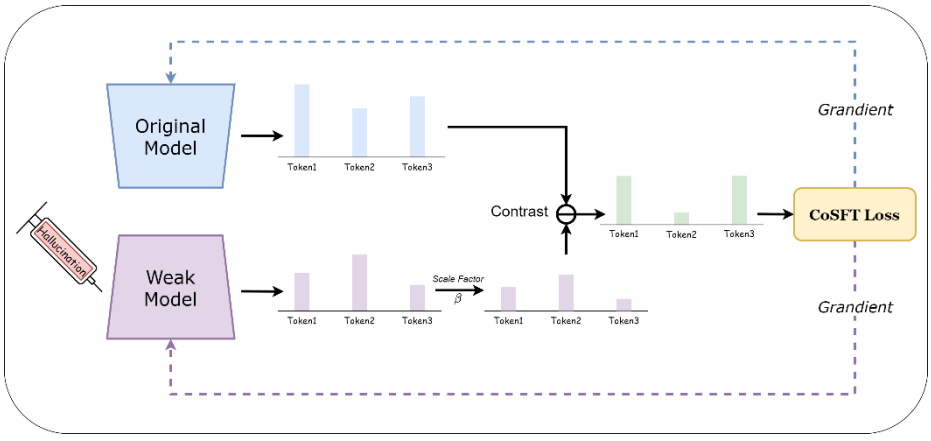

Fig. 1. Overview of ContrastSFT. During fine-tuning, a frozen weak model provides contrastive logits for CLR, suppressing hallucination-prone patterns. The weak model is removed at inference, incurring no additional latency.

3 Preliminary

In this section, we first analyze the theoretical limitations of standard Supervised Fine-Tuning (SFT) in the context of hallucination mitigation. We then formally define the task of hallucination mitigation via contrastive learning.

3.1 Motivation: The Limitation of SFT

The standard post-training paradigm for LLMs is SFT, which optimizes the model parameters θ via Maximum Likelihood Estimation (MLE). Given a prompt x and a ground-truth response y, SFT encourages the model to maximize the conditional probability $P_\theta(y|x)$.

While effective for instruction following, this objective has a critical flaw regarding factuality: *it only provides positive constraints*. The MLE objective pulls the model's distribution towards the ground truth but does not explicitly push it away from plausible but incorrect "hallucinations." Mathematically, minimizing the KL-divergence between the data distribution and the model distribution focuses on covering the modes of the data. However, if the pre-trained model has a strong prior belief in a hallucination h (e.g., a common misconception or a high-frequency ngram pattern), and h does not explicitly appear as a "negative example" in the SFT data, the probability $P_\theta(h|x)$ may remain undesirably high.

This phenomenon explains why SFT models often "hallucinate" by reverting to statistical shortcuts or parametric memory biases. To address this, we argue that the training objective must be **contrastive**: the model should not only maximize the probability of the truth y but also maximize the distance between its predictions and those of a "Weak Model" (M_{weak}) that represents these statistical shortcuts.

3.2 Task Definition

We define the problem of hallucination mitigation as a constrained conditional generation task. Let $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ be a dataset of factual instruction-response pairs, where x denotes the query and $y = (y_1, y_2, ..., y_T)$ denotes the sequence of ground-truth tokens from a vocabulary $\mathcal{V}$.

Standard SFT. The objective of standard SFT is to minimize the negative log-likelihood (NLL) over the dataset:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\,\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{t=1}^{T}\log P_\theta(y_t|x, y_{<t})\right] \tag{1}$$

where $P_\theta(y_t|x, y_{<t}) = \mathrm{softmax}(z_\theta(y_t))$, and $z_\theta \in \mathbb{R}^{|\mathcal{V}|}$ represents the logits computed by the model.

Weak Reference Model. We introduce a pre-trained Weak Model M_{weak} (parameterized by ϕ) which serves as a proxy for non-factual or heuristic-based generation. The weak model computes its own logits $z_{weak}(v|x, y_{<t})$ for any token $v \in \mathcal{V}$.

Goal. Our goal is to derive a new estimator $P_{\text{ContrastSFT}}$ that leverages the signal from z_{weak} to regularize z_θ. Specifically, we aim to optimize θ such that the model retains the general instruction-following capability of SFT while selectively suppressing tokens where P_{weak} is incorrectly confident (hallucinations), thereby improving the factuality scores on downstream benchmarks.

4 METHOD

In this section, we present **ContrastSFT**, a training framework designed to internalize the hallucination-mitigation capabilities of contrastive decoding into the model parameters. We first elaborate on the formulation of Contrastive Logit Regularization (CLR) and then provide a theoretical analysis of its gradient dynamics to explain why it effectively suppresses hallucinations.

4.1 Contrastive Logit Regularization

Let M_θ denote the target Large Language Model (Expert) parameterized by θ, and M_{weak} denote a frozen Weak Model (e.g., a smaller or non-instruction-tuned model). Given an input context x and a target token y_t at step t, standard SFT optimizes the probability $P_\theta(y_t|x)$.

We hypothesize that the raw logits z_θ of the expert model contain two components: (1) *Factual Knowledge*, which requires deep reasoning, and (2) *Surface Patterns*, which are statistical shortcuts shared with weaker models. To distill the factual component, we introduce the **Contrastive Logit** $\tilde{z}(\cdot)$, defined as:

$$\tilde{z}(v|x) = z_\theta(v|x) - \beta \cdot z_{weak}(v|x), \quad \forall v \in \mathcal{V} \tag{2}$$

where $\beta \in [0, \infty)$ is a hyperparameter governing the strength of the penalty on weak behaviors.

Unlike previous methods that apply this transformation only during inference (Contrastive Decoding), ContrastSFT incorporates it into the training objective. We define the **ContrastSFT Loss** as the negative log-likelihood of the ground truth token y_t under the regularized distribution:

$$\mathcal{L}_{\text{ContrastSFT}}(\theta) = -\sum_{t=1}^{T} \log \frac{\exp\left(z_\theta(y_t) - \beta z_{weak}(y_t)\right)}{\sum_{v \in \mathcal{V}} \exp\left(z_\theta(v) - \beta z_{weak}(v)\right)} \tag{3}$$

By minimizing Eq. 3, we force the parameters θ to adapt such that the contrastive distribution aligns with the ground truth. This effectively compels the expert model to "unlearn" the statistical biases captured by M_{weak}.

4.2 Gradient Analysis: How ContrastSFT Mitigates Hallucinations

To understand the mechanism of ContrastSFT, we analyze the gradient of the loss function with respect to the expert logits $z_\theta(v)$. The gradient is given by:

$$\frac{\partial \mathcal{L}_{\text{ContrastSFT}}}{\partial z_\theta(v)} = \tilde{P}(v|x) - \mathbb{1}(v = y_t) \tag{4}$$

where $\tilde{P}(v|x)$ is the probability derived from the contrastive logits $\tilde{z}$, and $\mathbb{1}$ is the indicator function.

Comparing this to standard SFT, where the gradient depends on the raw probability $P_\theta(v|x)$, we observe two critical regulatory effects driven by the term $-\beta z_{weak}$:

1. Suppression of "Easy" Hallucinations (Negative Mining): Consider a non-factual token v_{hal} (a hallucination) that is statistically plausible, meaning $z_{weak}(v_{hal})$ is high. In standard SFT, if the expert model $z_\theta(v_{hal})$ is also high, the gradient penalty is proportional to $P_\theta(v_{hal})$. In ContrastSFT, the term $-\beta z_{weak}(v_{hal})$ in the exponent significantly reduces the contrastive probability $\tilde{P}(v_{hal})$. This implies that the optimization process assigns lower priority to suppressing errors that are "obvious" (predictable by the weak model), allowing the gradient updates to focus on **Hard Negatives**—errors that are specific to the expert model and not explained by simple surface patterns.

2. Amplification of Factual Signal: For the ground truth token y_t, we effectively maximize $z_\theta(y_t) - \beta z_{weak}(y_t)$. If y_t is a "hard" fact (where $z_{weak}(y_t)$ is low), the penalty is small. However, if y_t is a common token (where $z_{weak}(y_t)$ is high), the model must push $z_\theta(y_t)$ to be significantly higher to overcome the $\beta z_{weak}(y_t)$ penalty. This forces the expert model to develop much higher confidence and sharper distinctiveness for factual tokens compared to a baseline model.

4.3 Inference Efficiency

A key advantage of ContrastSFT is its inference-time efficiency. Since the contrastive constraints are internalized into the parameters θ during fine-tuning, the model M_θ learns to generate factual representations intrinsically. Therefore, during the inference phase, we strictly use the original model M_θ for decoding:

$$\hat{y}_t = \text{argmax}_v \, z_\theta(v|x) \tag{5}$$

This eliminates the need to load or compute M_{weak} at runtime, maintaining the same inference latency as standard SFT, unlike CD approaches which double the computational cost.

5 Experiments

In this section, we evaluate ContrastSFT on a diverse set of benchmarks covering Natural Language Understanding (NLU) and Knowledge Factuality. We aim to answer three key research questions: **Q1 (Effectiveness):** Does ContrastSFT effectively mitigate hallucinations compared to standard SFT and inference-time intervention methods? **Q2 (Efficiency):** Does ContrastSFT retain the inference efficiency of the base model, avoiding the latency overhead of contrastive decoding? **Q3 (Robustness):** How does the regularization strength β affect the trade-off between factuality and general capability?

5.1 Experimental Setup

Datasets. We evaluate our method on five benchmark datasets: **ParaRel(QA)**, a high-quality English cloze-style paraphrase dataset comprising 328 paraphrases across 38 relations, used to assess the paraphrase consistency of pre-trained language models; **HaluEval(QA)**, a benchmark for evaluating factual accuracy and hallucination in LLMs across diverse scenarios, where we reformat its generative tasks for our experiments; **MMLU(MCQ)**, a classic multi-domain NLP benchmark, employed to reveal the factual knowledge boundaries of models (we adopt the version constructed by R-Tuning); **WiCE(MCQ)**, a fine-grained textual entailment dataset derived from Wikipedia, providing claim sentences, supporting evidence, and token-level annotations of unsupported content; and **FEVER(MCQ)**, a large-scale fact verification dataset containing 185,445 labeled claims derived from Wikipedia, each annotated as Supported, Refuted, or NotEnoughInfo, with evidence provided for the first two categories.

5.2 Metrics

Our experiments are conducted in two settings: multiple-choice (MCQ) and question answering (QA).

For MCQ, we compute Multiple-Choice Accuracy (MCA):

$$\mathrm{MCA} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left(\hat{y}_i = y_i\right) \tag{6}$$

For QA, we adopt Matching Accuracy (MA) to assess answer inclusion:

$$\mathrm{MA} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left(A_i \subset \hat{A}_i\right) \tag{7}$$

Additionally, for the WiCE and FEVER datasets, we compute the F1 score:

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{8}$$

5.3 Baselines

We evaluate ContrastSFT against the following baseline methods. More experimental details regarding baseline construction are presented in the Appendices and the next section:

- **Vanilla:** Vanilla baselines involve the direct use of both Llama3-Base and Llama3-Instruct without any additional training. These baselines can help us understand the fundamental ability of the Llama3 series models. Additionally, a Weak baseline is created by inducing hallucinations in the Llama3 models [3] to assess the impact of hallucination induction.
- **Supervised Fine-Tuning (SFT)** [1]: In this approach, we perform fine-tuning with an extensive hyperparameter search to optimize performance.

- **Contrastive Decoding (CD)** [25]: A standard contrastive decoding strategy that contrasts output distributions from models with different parameter scales to enhance response quality.
- **Induce-then-Contrast Decoding (ICD)** [31]: ICD involves creating a factually weaker LLM by inducing hallucinations in the original model and penalizing these hallucinations during decoding, thereby improving the factual accuracy of generated content.

Table 1. HaluEval Hallucinatory Dataset Profile

Data Source	Prompt Format	Dataset summarization	Actual Data Used
QA data	Instruction: The Oberoi ... in what city? Input: The Oberoi family is ... in Delhi. Hallucinated Output: The Oberoi ... in Mumbai.	10K	10K
Dialogue data	Instruction: [Human]: Could you ... Fight Club? Input: Fight Club is starring ... in Panic Room Hallucinated Output: Sure, Mr. Nobody ... movie.	10K	10K
Summarization data	Instruction: Write a summary ... given below. Input: For the first time in ... miss a beat.\n Hallucinated Output: Host ... eight-year hiatus.	10K	10K
General data	Instruction: NA Input: Identify an interesting fact about Saturn. Hallucinated Output: Saturn has ... meters thick.	10K	0.8K

5.4 Implementation details

Here, we employ the LLaMA3 series models [3] for our experiments. Specifically, we utilize three versions: Llama3.1-8B, Llama3.1-8B-Instruct, and Llama3.2-3B. These models serve as the foundation for evaluating our proposed ContrastSFT framework and the baseline methods.

Baseline Enhancement. To ensure fairness, we enhance the contrastive decoding baselines by applying SFT to the original model. These original models are trained on the same training set as ContrastSFT. The procedure not only establishes a consistent baseline but also allows us to directly assess the impact of contrastive strategies on downstream tasks.

Weak Model Construction. Following standard practices in contrastive decoding [25,26], we employ a smaller-scale language model as the weak reference to capture surface-level statistics. Specifically, we use Llama-3.2-3B-Base as M_{weak} while fine-tuning Llama-3.1-8B-Base as M_θ. The weak model remains frozen during the entire training process.

For inducing hallucinations in the weak model, we adopt the same techniques as prior studies [10,31]. This operation is crucial for contrastive decoding by ensuring the weak model generates hallucinatory outputs.

We fine-tune the Llama3 model [3] using the negative samples from the HaluEval dataset [39]. HaluEval comprises 5,000 general user queries with ChatGPT responses, along with 30,000 task-specific examples across three distinct tasks: question answering, knowledge-grounded dialogue, and text summarization. A summary of the dataset and the specific samples used in our experiments is provided in Table 1.

To make a fair comparison and ensure the effectiveness of our experiment, we use the same configuration as previous studies [10,31] to train the weak model. The detailed configuration is presented in Table 2. It is worth noting that only the hallucinatory samples from HaluEval are used to induce hallucinations, ensuring no risk of information leakage into the main experiments involving HaluEval.

Model Alignment. To align both SFT and ContrastSFT with downstream tasks, we conduct a comprehensive hyperparameter search for all models.

For the CD&SFT models illustrated in Table 4, we first train an SFT model S_1 on the train set of each dataset. The evaluation results of SFT in Table 4 are derived directly from model S_1. To construct the CD [25] baseline, we take S_1 as the original model and Llama3.2-3B as the weak model. Similarly, for the baseline ICD [31], S_1 serves as the original model, while the weak model is configured as the one described in the Weak Model Construction section.

Table 2. Fine-tuning configuration and hyper-parameters.

Configuration	Value
Model	Llama3.1-8B
Epochs	5
Devices	4 NVIDIA GeForce RTX 3090 (24G)
Total Batch Size	256
Optimizer	Adam [40]($\boldsymbol{\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}}$)
Learning rate	5.0e-4
LR scheduler type	cosine
Warm-up Ratio	0.0
LoRA Target	$q_{proj}, k_{proj}, v_{proj}, o_{proj}$
LoRA Rank	32

The model tuning experiments are performed using LoRA [41]. To ensure optimal performance for each model, we conduct a grid search over the hyperparameters, as outlined in Table 3. For hyperparameters included in the grid search, we present the search ranges; for others, we list the fixed values.

Table 3. Hyperparameter Settings.

Hyperparameter	Range or Value
Learning rate	$[\boldsymbol{1e-4, 4e-4}]$
Batch size	$[\boldsymbol{16, 64}]$
Label smoothing	$[\boldsymbol{0.05, 0.15}]$
LR scheduler type	{Cosine, Linear}
Optimizer	Adam-8bit [42]($\boldsymbol{\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1e-8}$)
LoRA target	All linear layers
Epoch	1
Warm-up ratio	0.02

Hyperparameter β. As defined in Equation 3, the hyperparameter β controls the strength of the regularization. Based on previous studies [25,31] and through preliminary experiments on the validation set, we set $\beta = 0.5$ and $\beta = 1$ as a balanced trade-off between suppressing hallucinations and maintaining text fluency for all experiments, selecting the higher score as the final result.

5.5 Main Results

NLU & Reasoning Capabilities. A key observation from Table 4 is the performance volatility of inference-time methods. Take CD as an example: despite a marginal lead (+0.37%) in the relatively straightforward ParaRel task, its F1 score plunges to 64.48% on the WiCE benchmark. This sharp decline likely stems from the "false penalty" effect (see Section 4), where the decoding-time subtraction inadvertently suppresses valid but less probable tokens. In contrast, ContrastSFT maintains a steady trajectory across varying task complexities. By embedding contrastive signals directly into the optimization objective, our method secures substantial gains—up to +9.52% on WiCE—over the SFT baseline, suggesting that fine-grained distinction is better captured during training than through post-hoc token filtering.

Factuality & General Knowledge. Results on HaluEval and MMLU further highlight the differences between training-based and inference-time approaches. On MMLU, inference-time methods such as CD and ICD lead to noticeable performance degradation, even falling below the Llama3-Instruct baseline (58.35%). This trend suggests that directly subtracting weak model logits may interfere with the model's general knowledge representations. In contrast, ContrastSFT achieves a higher score of 63.49%, indicating that the regularization can mitigate hallucinations without significantly harming general capabilities. On HaluEval, ContrastSFT performs comparably to the SFT baseline (89.10% vs. 89.25%), with only a marginal difference. At the same time, it avoids the performance drop observed in CD (86.90%) and ICD (85.95%), suggesting a more stable trade-off between factuality and overall performance.

Table 4. Comparison of different models on various datasets.

Category	Model	Datasets and Metrics (%)						
		MMLU	ParaRel	WiCE		FEVER		HaluEval
		MCA	MA	F1	MCA	F1	MCA	MA
Vanilla	Llama3-B	53.08	28.01	21.78	23.06	21.16	32.07	70.05
	Llama3-I	58.35	38.53	38.76	59.64	64.50	65.34	79.30
	Weak	40.59	36.87	29.15	41.51	40.05	45.79	34.95
CD&SFT	SFT	62.43	70.02	70.30	81.13	92.94	93.31	**89.25**
	CD†	57.33	**70.57**	64.48	80.01	89.19	89.70	86.90
	ICD†	56.59	68.93	70.65	80.71	89.92	90.32	85.95
Ours	ContrastSFT (ours)	**63.49**	70.20	**79.82**	**85.22**	**93.77**	**94.10**	89.10

† Enhanced baselines. See Section 5.4

Summary. Overall, ContrastSFT ranks first or ties for first on 4 out of 5 benchmarks. Unlike CD and ICD, which show noticeable trade-offs across tasks, our method main-

tains more consistent performance across different domains. This supports the hypothesis that incorporating contrastive signals during training can lead to more stable behavior than relying solely on post-hoc decoding strategies.

Consistency Across Domains as a Proxy for Robustness. Although the absolute gains on some datasets (e.g., +1.06% on MMLU) are relatively modest, a key advantage of ContrastSFT lies in its consistency. As shown in Table 4, it is the only method that achieves non-negative improvements (or statistically comparable results) across all five benchmarks. In contrast, CD and ICD exhibit higher variance, improving on ParaRel while degrading performance on WiCE (-5.82%) and MMLU (-5.1%). This pattern suggests that the proposed regularization contributes to more robust behavior across diverse data distributions.

6 CONCLUSIONS

In this paper, we introduced ContrastSFT, a training framework that incorporates contrastive signals into supervised fine-tuning to mitigate hallucinations in LLMs. Instead of relying solely on likelihood maximization, the proposed Contrastive Logit Regularization (CLR) leverages a weak reference model to suppress patterns that are more likely to be associated with hallucination. From a conceptual perspective, this approach extends the standard SFT objective by introducing an explicit contrastive component, encouraging the model to distinguish between reliable reasoning and spurious correlations. At the same time, by integrating the contrastive signal into training, it avoids the additional inference cost typically introduced by contrastive decoding methods.

Experimental results across both NLU and factuality benchmarks show that ContrastSFT improves factual consistency while maintaining general reasoning performance, outperforming standard SFT and several inference-time baselines in most settings. These results suggest that contrastive learning can be incorporated more directly into model training, rather than being limited to decoding-time heuristics. As a possible direction for future work, this framework could be extended to other alignment objectives, such as safety or controllability, or combined with adaptive regularization strategies to better balance factuality and diversity.

7 LIMITATIONS AND FUTURE WORK

While ContrastSFT shows strong empirical performance and maintains high inference efficiency, several limitations remain. First, the method introduces additional hyperparameters. Due to computational constraints when fine-tuning 8B-scale models, we only evaluated a discrete set of regularization strengths. A more detailed analysis of how performance varies with this parameter—particularly around potential over-regularization—would provide a clearer understanding of its behavior. Second, our experiments are currently limited to the LLaMA-3 architecture. Although the formulation of Contrastive Logit Regularization is not tied to a specific model, it would be useful to verify its effectiveness on other model families (e.g., Qwen [43]) and at larger scales such as 70B parameters. Finally, the reported results are based on single runs. Evaluating the method across multiple random seeds would help quantify variance and provide stronger statistical support for the observed improvements.

REFERENCES

1. T. Brown et al., “Language models are few-shot learners,” in Advances in Neural Information Processing Systems (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.
2. J. Yu, S. Zhou, D. Yang, S. Li, S. Wang, X. Hu, C. Xu, Z. Xu, C. Shu, and Z. Yuan, “Mquant: Unleashing the inference potential of multimodal large language models via static quantization,” in Proceedings of the 33rd ACM International Conference on Multimedia, pp. 1783–1792, 2025.
3. A. Dubey, A. Jauhri, et al., “The llama 3 herd of models,” 2024.
4. L. Ouyang et al., “Training language models to follow instructions with human feedback,” 2022.
5. R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, “Direct preference optimization: your language model is secretly a reward model,” in Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23, (Red Hook, NY, USA), Curran Associates Inc., 2024.
6. A. Yang, B. Yang, et al., “Qwen2 technical report,” 2024.
7. S. Zhou, S. Wang, Z. Yuan, M. Shi, Y. Shang, and D. Yang, “GSQtuning: Group-shared exponents integer in fully quantized training for LLMs on-device fine-tuning,” in Findings of the Association for Computational Linguistics: ACL 2025, (Vienna, Austria), pp. 22971–22988, Association for Computational Linguistics, July 2025.
8. X. Hu, Y. Cheng, D. Yang, Z. Chen, Z. Xu, Z. Yuan, S. Zhou, et al., “Ostquant: Refining large language model quantization with orthogonal and scaling transformations for better distribution fitting,” in The Thirteenth International Conference on Learning Representations.
9. X. Gu, X. Sun, C. Feng, X. Wang, and X. Chen, “Fedsit: Efficient federated fine-tuning with model splitting and importance-based tuning,” in 2025 International Joint Conference on Neural Networks (IJCNN), pp. 1–8, 2025.
10. D. Chen, F. Fang, S. Ni, F. Liang, R. Xu, M. Yang, and C. Li, “Lower layer matters: Alleviating hallucination via multi-layer fusion contrastive decoding with truthfulness refocused,” arXiv preprint arXiv:2408.08769, 2024.
11. X. L. Li, A. Holtzman, D. Fried, P. Liang, J. Eisner, T. B. Hashimoto, L. Zettlemoyer, and M. Lewis, “Contrastive decoding: Open-ended text generation as optimization,” in Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers), pp. 12286–12312, 2023.
12. R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, “Direct preference optimization: Your language model is secretly a reward model,” Advances in neural information processing systems, vol. 36, pp. 53728–53741, 2023.
13. L. Gao, S. Biderman, et al., “The pile: An 800gb dataset of diverse text for language modeling,” 2020.
14. H. Touvron, L. Martin, et al., “Llama 2: Open foundation and fine-tuned chat models,” 2023.
15. [S. Gunasekar, Y. Zhang, et al., “Textbooks are all you need,” 2023.
16. A. Sinitsin, V. Plokhotnyuk, D. Pyrkin, S. Popov, and A. Babenko, “Editable neural networks,” in International Conference on Learning Representations, 2020.
17. N. De Cao, W. Aziz, and I. Titov, “Editing factual knowledge in language models,” in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, eds.), (Online and Punta Cana, Dominican Republic), pp. 6491–6506, Association for Computational Linguistics, Nov. 2021.

18. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in Advances in Neural Information Processing Systems (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 9459–9474, Curran Associates, Inc., 2020.
19. K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in Proceedings of the 37th International Conference on Machine Learning (H. D. III and A. Singh, eds.), vol. 119 of Proceedings of Machine Learning Research, pp. 3929–3938, PMLR, 13–18 Jul 2020.
20. K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," in Findings of the Association for Computational Linguistics: EMNLP 2021 (M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, eds.), (Punta Cana, Dominican Republic), pp. 3784–3803, Association for Computational Linguistics, Nov. 2021.
21. N. Lee, W. Ping, P. Xu, M. Patwary, P. N. Fung, M. Shoeybi, and B. Catanzaro, "Factuality enhanced language models for open-ended text generation," Advances in Neural Information Processing Systems, vol. 35, pp. 34586–34599, 2022.
22. S. R. Bowman, J. Hyun, et al., "Measuring progress on scalable oversight for large language models," 2022.
23. W. Saunders, C. Yeh, J. Wu, S. Bills, L. Ouyang, J. Ward, and J. Leike, "Self-critiquing models for assisting human evaluators," 2022.
24. L. Pan, M. Saxon, W. Xu, D. Nathani, X. Wang, and W. Y. Wang, "Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies," 2023.
25. X. L. Li, A. Holtzman, D. Fried, P. Liang, J. Eisner, T. Hashimoto, L. Zettlemoyer, and M. Lewis, "Contrastive decoding: Open-ended text generation as optimization," arXiv preprint arXiv:2210.15097, 2022.
26. Y.-S. Chuang, Y. Xie, H. Luo, Y. Kim, J. R. Glass, and P. He, "Dola: Decoding by contrasting layers improves factuality in large language models," in The Twelfth International Conference on Learning Representations, 2024.
27. D. Wan, M. Liu, K. McKeown, M. Dreyer, and M. Bansal, "Faithfulness-aware decoding strategies for abstractive summarization," in Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (A. Vlachos and I. Augenstein, eds.), (Dubrovnik, Croatia), pp. 2864–2880, Association for Computational Linguistics, May 2023.
28. S. Zhong, X. Wang, H. Wang, X. Chen, Y. Xu, X. Liu, and B. Wu, "Improving graph contrastive learning with llms: A new hardness-aware negative sampling strategy," in 2025 International Joint Conference on Neural Networks (IJCNN), pp. 1–8, 2025.
29. C. Liu, G. Wang, P. Liu, X. Wei, and H. Zhu, "Hybrid multi-stage decoding for few-shot ner with entity-aware contrastive learning," in 2025 International Joint Conference on Neural Networks (IJCNN), pp. 1–7, 2025.
30. S. O'Brien and M. Lewis, "Contrastive decoding improves reasoning in large language models," 2023.
31. Y. Zhang, L. Cui, W. Bi, and S. Shi, "Alleviating hallucinations of large language models through induced hallucinations," arXiv preprint arXiv:2312.15710, 2023.
32. C. Zheng, L. Li, Q. Dong, Y. Fan, Z. Wu, J. Xu, and B. Chang, "Can we edit factual knowledge by in-context learning?," 2023.
33. E. Mitchell, C. Lin, A. Bosselut, C. Finn, and C. D. Manning, "Fast model editing at scale," 2022.

34. Z. Huang, Y. Shen, X. Zhang, J. Zhou, W. Rong, and Z. Xiong, “Transformer-patcher: One mistake worth one neuron,” 2023.
35. K. Meng, D. Bau, A. Andonian, and Y. Belinkov, “Locating and editing factual associations in gpt,” in Advances in Neural Information Processing Systems (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 17359–17372, Curran Associates, Inc., 2022.
36. K. Meng, A. S. Sharma, A. Andonian, Y. Belinkov, and D. Bau, “Mass-editing memory in a transformer,” 2023.
37. A. Gupta, D. Mondal, A. K. Sheshadri, W. Zhao, X. L. Li, S. Wiegreffe, and N. Tandon, “Editing common sense in transformers,” 2023.
38. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., “Training language models to follow instructions with human feedback,” Advances in neural information processing systems, vol. 35, pp. 27730–27744, 2022.
39. J. Li, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen, “HaluEval: A large-scale hallucination evaluation benchmark for large language models,” in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 6449–6464, Association for Computational Linguistics, Dec. 2023.
40. D. P. Kingma, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.
41. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-rank adaptation of large language models,” in International Conference on Learning Representations, 2022.
42. T. Dettmers, M. Lewis, S. Shleifer, and L. Zettlemoyer, “8-bit optimizers via block-wise quantization,” in International Conference on Learning Representations, 2022.
43. A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., “Qwen3 technical report,” arXiv preprint arXiv:2505.09388, 2025.