Alleviating Hallucinations from Knowledge Misalignment in Large Language Models via Selective Abstention Learning

Lei Huang[1], Xiaocheng Feng[1,2*], Weitao Ma[1], Yuchun Fan[3], Xiachong Feng[4], Yuxuan Gu[1], Yangfan Ye[1], Liang Zhao[1], Weihong Zhong[1], Baoxin Wang[5], Dayong Wu[5], Guoping Hu[5], Lingpeng Kong[4], Tong Xiao[3], Ting Liu[1], Bing Qin[1,2]
[1] Harbin Institute of Technology, China [2] Peng Cheng Laboratory, China
[3] Northeastern University, China [4] The University of Hong Kong, China [5] iFLYTEK Research, China
`{lhuang, xcfeng, qinb}@ir.hit.edu.cn` `{bxwang2, dywu2, gphu}@iflytek.com`

Abstract

Large language models (LLMs) are known to suffer from severe hallucination issues. One of the main causes lies in the knowledge misalignment between the pre-training stage and the supervised fine-tuning stage. The unfamiliar knowledge encountered during fine-tuning may encourage LLMs to generate facts that are not grounded in parametric knowledge. To address this, we propose SEAL[1], a novel training objective with an abstention mechanism, in which the model learns to selectively reject tokens that misalign with the desired knowledge distribution via a special `[REJ]` token. This allows the model to have the alternative of acknowledging the insufficiency of knowledge rather than blindly assigning high probability to all ground-truth answers. We further propose a regularized decoding objective that penalizes uncertain predictions during inference by using the `[REJ]` probability learned during training. Extensive experiments on six short-form and long-form QA datasets with three LLMs of different sizes demonstrate that our method effectively alleviates hallucinations caused by knowledge misalignment. Further analysis highlights the adaptations of our method in answer refusal scenarios and its ability to effectively maintain the model's instruction-following capabilities.

1 Introduction

Large language models (LLMs) (OpenAI, 2023; AI@Meta, 2024) have shown remarkable capabilities in capturing factual knowledge from the large-scale pre-training corpus. However, they still exhibit a notable tendency to generate factually incorrect content, known as hallucinations (Huang et al., 2025b), which presents significant challenges in their real-world applications.

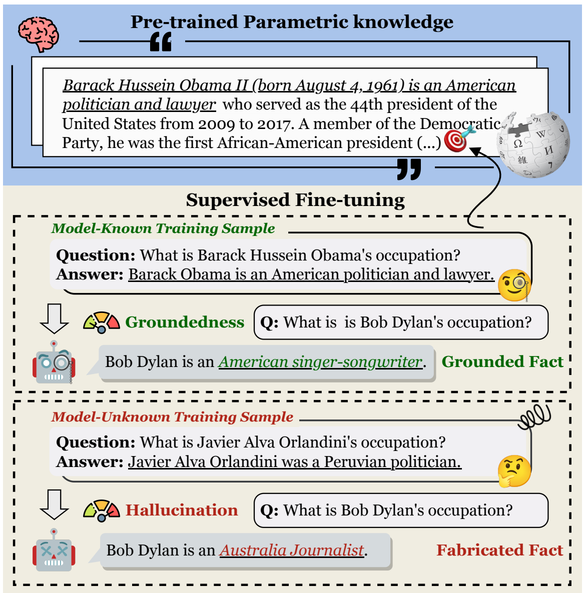

Pre-trained Parametric knowledge
Barack Hussein Obama II (born August 4, 1961) is an American politician and lawyer who served as the 44th president of the United States from 2009 to 2017. A member of the Democratic Party, he was the first African-American president (...)
Supervised Fine-tuning
Model-Known Training Sample
Question: What is Barack Hussein Obama's occupation?
Answer: Barack Obama is an American politician and lawyer.
Groundedness
Q: What is is Bob Dylan's occupation?
Bob Dylan is an American singer-songwriter.
Grounded Fact
Model-Unknown Training Sample
Question: What is Javier Alva Orlandini's occupation?
Answer: Javier Alva Orlandini was a Peruvian politician.
Hallucination
Q: What is Bob Dylan's occupation?
Bob Dylan is an Australia Journalist.
Fabricated Fact

Figure 1: An illustration of the impact of new factual knowledge encountered in fine-tuning on hallucinations. During fine-tuning, *model-known* samples (*e.g.*,*"Barack Obama"*) teach the model to ground the generation to its parametric knowledge, whereas *model-unknown* samples (*e.g.*,*"Javier Alva Orlandini"*) encourage the model to fabricate facts that are not grounded in its parametric knowledge, leading to hallucinations.

Recent research has indicated that a significant cause of hallucination lies in the *knowledge misalignment* between the pre-training stage and the supervised fine-tuning (SFT) stage (Schulman, 2023; Kang et al., 2024; Gekhman et al., 2024). During the post-training stage, fine-tuning serves as an essential step to fully activate the knowledge captured during the pre-training stage (Zhou et al., 2023a). However, this process may encounter new factual knowledge misaligned with the knowledge embedded in LLMs. As shown in Figure 1, the new factual knowledge can inadvertently encourages the model to fabricate facts not grounded in its parametric knowledge (Liu et al., 2024b), thus giving rise to hallucination (Huang et al., 2025a,b).

Current efforts to mitigate *knowledge misalignment* primarily focus on two aspects: one line of

[*] Corresponding Author
[1] Acronym for **SE**lective **A**bstention **L**earning

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 24564–24579
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

research involves pre-filtering out *model-unknown* samples, fine-tuning models solely on samples within their knowledge boundaries (Ghosal et al., 2024). Despite effectiveness, these *unknown* samples are typically *model-specific*, making it infeasible to annotate accurately. Another approach (Tian et al., 2024; Lin et al., 2024a) aims to utilize the pre-trained model itself to generate training samples, avoiding the introduction of new knowledge. However, the generated supervised data lacks reliable validation, often resulting in poor quality and may even introduce additional hallucinations.

In this work, we propose SEAL, a novel training objective with an *abstention mechanism*, enabling the model to selectively reject tokens that misalign with the desired knowledge distribution (§3.2). Specifically, we introduce a special token `[REJ]`, and each time the model fails to predict the ground-truth token, a portion of the target probability is shifted to the `[REJ]` token based on the predicted logits. This allows the model to have the alternative of acknowledging the insufficiency of knowledge, rather than blindly assigning high probability to all ground-truth answers. During this process, the `[REJ]` token captures the uncertainty arising from knowledge discrepancies in model predictions. Building upon this, we further propose abstention-aware decoding (§3.3), incorporating the uncertainty reflected by the `[REJ]` token into the search-based decoding strategy. By penalizing uncertain predictions at each decoding step, this strategy guides the model to navigate towards more confident and factual trajectories.

To validate the effectiveness of SEAL, we conduct extensive experiments on three representative LLMs across different sizes, covering six short-form and long-form factual question-answering (QA) datasets. The results show that SEAL effectively alleviates hallucinations induced by knowledge misalignment. Compared with the vanilla MLE objective, our method improves the factuality of LLMs by 8.59% and 10.80% in short-form and long-form QA, respectively, while maintaining their ability to follow instruction. Further analysis highlights that `[REJ]` shows effective calibration and can be expanded to answer refusal scenarios.

2 Related Work

Factuality Hallucination Mitigation. Factuality hallucination in LLMs (Huang et al., 2025b) refers to the generated content that deviates from established world knowledge. The factors leading to such hallucinations are diverse, spanning almost the entire lifecycle of LLMs, from pre-training (Allen-Zhu and Li, 2024) to supervised fine-tuning (Schulman, 2023), alignment (Lin et al., 2024a) and the decoding stage (Li et al., 2023a). Numerous studies have explored ways to improve factuality at various stages, such as continual pre-training (Chang et al., 2024), uncertainty calibration (Cohen et al., 2024), factuality alignment (Tian et al., 2024), and contrastive decoding (Chuang et al., 2024; Huang et al., 2024a,b). In this work, we primarily focus on mitigating hallucinations induced by the unfamiliar factual knowledge encountered during the fine-tuning stage, which has recently garnered significant attention (Kang et al., 2024; Gekhman et al., 2024).

Improving Factuality during Fine-tuning. Recent research (Kang et al., 2024) has revealed that the distribution of unfamiliar knowledge actually controls how LLMs hallucinate. This has inspired a line of studies that focus on avoiding introducing unknown knowledge during fine-tuning, either filtering out unknown samples (Ghosal et al., 2024) or leveraging the base model itself to generate supervised fine-tuning examples (Lin et al., 2024a). More recently, Liu et al. (2024b) proposed disentangling skills and knowledge learning during the fine-tuning process, encouraging the groundedness of LLMs through synthetic data. Unlike these methods, we take a more fundamental perspective by equipping the model with an abstention mechanism during training, effectively alleviating the issue of blind imitation in traditional MLE objectives.

3 Methodology

In this section, we start with the problem formulation, followed by our method, which aims to alleviate hallucinations caused by knowledge misalignment. An overview is presented in Figure 2.

3.1 Problem Formulation

Given a pre-trained base model, denoted as $\mathcal{M}_\theta$, and a fine-tuning dataset $\mathcal{D}$, our objective is to fine-tune $\mathcal{M}_\theta$ on factual question-answering (QA) tasks. The dataset $\mathcal{D}$ comprises a set of QA pairs, represented as $\mathcal{D} = \left\{(q_i, a_i)\right\}_{i=1}^{N}$, where each q_i is a knowledge-seeking question, and a_i is the corresponding ground-truth answer. Each QA pair is transformed into a structured instruction-response pair $(\mathbf{x}, \mathbf{y})$ using pre-defined prompt templates de-

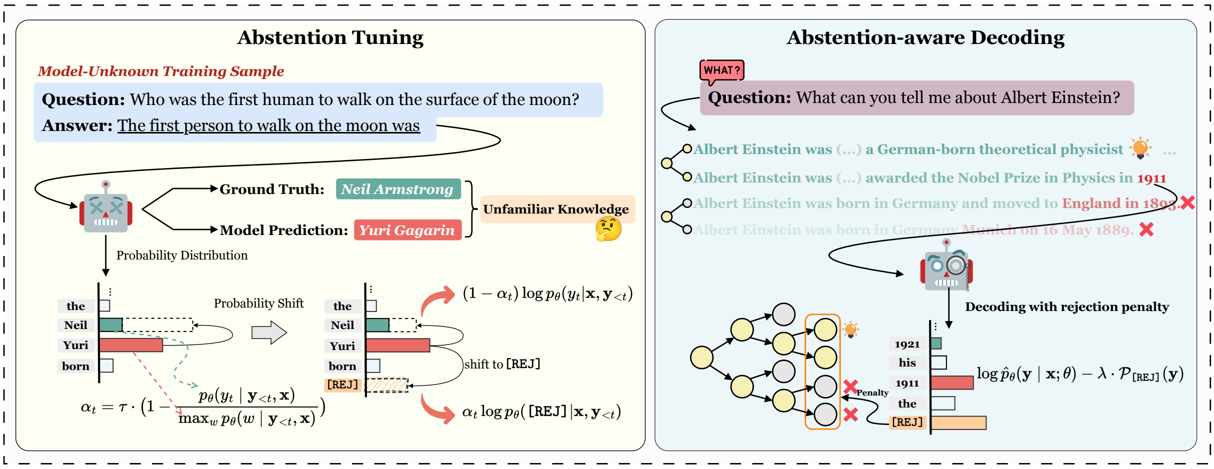

Abstention Tuning
Model-Unknown Training Sample
Question: Who was the first human to walk on the surface of the moon?
Answer: The first person to walk on the moon was
Ground Truth: Neil Armstrong
Model Prediction: Yuri Gagarin
Unfamiliar Knowledge
Probability Distribution
the
Neil
Yuri
born
Probability Shift
the
Neil
Yuri
born
[REJ]
shift to [REJ]
$(1-\alpha_t)\log p_\theta(y_t|\mathbf{x},\mathbf{y}_{<t})$
$\alpha_t \log p_\theta(\texttt{[REJ]}|\mathbf{x},\mathbf{y}_{<t})$
$\alpha_t = \tau \cdot (1 - \frac{p_\theta(y_t \mid \mathbf{y}_{<t}, \mathbf{x})}{\max_w p_\theta(w \mid \mathbf{y}_{<t}, \mathbf{x})})$
Abstention-aware Decoding
WHAT?
Question: What can you tell me about Albert Einstein?
Albert Einstein was (...) a German-born theoretical physicist ...
Albert Einstein was (...) awarded the Nobel Prize in Physics in 1911
Albert Einstein was born in Germany and moved to England in 1893.
Albert Einstein was born in Germany Munich on 16 May 1889.
Decoding with rejection penalty
1921
his
1911
the
[REJ]
Penalty
$\log \hat{p}_\theta(\mathbf{y} \mid \mathbf{x};\theta) - \lambda \cdot \mathcal{P}_{\texttt{[REJ]}}(\mathbf{y})$

Figure 2: An overview of SEAL: (1) abstention tuning (§3.2) enables the model to recognize its knowledge limitations by dynamically allocating part of the probability to the [REJ] token, thereby avoiding overfitting to misaligned knowledge distributions; (2) abstention-aware decoding (§3.3) utilizes the [REJ] probability learned during fine-tuning to penalize uncertain predictions, guiding the generation towards more truthful outputs.

noted by $f(\cdot)$. Our experimental setup encompasses both short-form and long-form QA datasets, for which we have designed distinct sets of prompt templates; details are provided in Appendix A.

For each instruction $\mathbf{x}$, standard supervised fine-tuning seeks to maximize the likelihood of the ground-truth answer $\mathbf{y}$ using the maximum likelihood estimation (MLE) objective, formalized in Equation 1. This is also mathematically equivalent to the cross-entropy loss objective, where the target distribution is modeled as a one-hot vector.

$$\begin{aligned} \mathcal{L}(\theta) &= -\log p_\theta(\mathbf{y}|\mathbf{x}) \\ &= -\sum_{t=1}^{|\mathbf{y}|} \log p_\theta(y_t|\mathbf{x}, \mathbf{y}_{<t}), \end{aligned} \tag{1}$$

where y_t is the token from the ground-truth answer, selected from a pre-defined vocabulary $\mathcal{V}$.

3.2 Abstention Tuning

The vanilla MLE objective aims to maximize the likelihood of *all ground-truth answers*. However, due to discrepancies between the factual knowledge in fine-tuning samples $\mathcal{D}$ and the parametric knowledge embedded within the LLM $\mathcal{M}_\theta$, there exist some samples that exceed the knowledge scope of the base model. Forcibly fitting these *unknown samples* can inadvertently encourage the model to fabricate facts not grounded in its pre-existing knowledge, resulting in hallucinations.

Drawing inspiration from the token selection strategies (Cohen et al., 2024; Lin et al., 2024b) in the pre-training stage, we propose a training objective with a dynamic *abstention mechanism* for supervised fine-tuning. This mechanism enables the model to selectively reject tokens that misalign with the desired knowledge distribution. Specifically, we add a special token, [REJ], into the vocabulary $\mathcal{V}$. Whenever the model fails to predict the ground-truth token, we dynamically adjust the target distribution by allocating a portion of the probability, α_t, to the [REJ] token, as defined below:

$$y_{\text{target}} = (1 - \alpha_t) \cdot y_t + \alpha_t \cdot \mathbf{1}_{\texttt{[REJ]}} \tag{2}$$

Intuitively, α_t should be 0 when the model accurately predicts the ground-truth token, and it should approach 1 when a significant discrepancy exists between the model's predictions and the ground truth. In this way, we calculate the shifted probability α_t as follows:

$$\alpha_t = \tau \cdot (1 - \frac{p_\theta(y_t \mid \mathbf{y}_{<t}, \mathbf{x})}{\max_w p_\theta(w \mid \mathbf{y}_{<t}, \mathbf{x})}) \tag{3}$$

Here, $p_\theta(y_t|\mathbf{y}_{<t}, \mathbf{x})$ denotes the probability of the ground truth, and $\max_w p_\theta(w|\mathbf{y}_{<t}, \mathbf{x})$ denotes the current maximum predicted probability. τ, a threshold within the range $[0, 1]$ caps the upper bound of target probability that can be assigned to the [REJ] token. In our experiments, we set $\tau = 0.5$.

Thus, the standard cross-entropy loss (Equation 1) is modified as follows:

$$\begin{aligned} \mathcal{L}_{nll} = -\sum_{t=1}^{|\mathbf{y}|} \big(&(1-\alpha_t)\log p_\theta(y_t|\mathbf{x}, \mathbf{y}_{<t}) \\ &+ \alpha_t \log p_\theta(\texttt{[REJ]}|\mathbf{x}, \mathbf{y}_{<t}) \big) \end{aligned} \tag{4}$$

In this situation, either *successfully* predicting the ground truth or *appropriately* abstaining from a prediction by predicting the [REJ] token can reduce the overall loss. This encourages the model to appropriately acknowledge its insufficiency of knowledge, rather than blindly assigning high probability to *all ground truth answers*. Furthermore, to prevent the model from excessively predicting the [REJ] token to reduce loss, we incorporate an additional regularization term that penalizes the model for abstaining when a correct prediction is feasible:

$$\mathcal{L}_{reg} = -\sum_{t=1}^{|\mathbf{y}|} \mathbb{I}_{\text{correct}} \cdot \log\left(1 - p_\theta\left(\texttt{[REJ]} \mid \mathbf{x}, \mathbf{y}_{<t}\right)\right) \tag{5}$$

The final loss function, combining the training objective with an abstention mechanism and the regularization term, is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{nll} + \mathcal{L}_{reg} \tag{6}$$

3.3 Abstention-aware Decoding

During the fine-tuning process, the [REJ] token serves as a *placeholder* that absorbs the uncertainty arising from knowledge discrepancies in model predictions. In this manner, these *unknown samples* effectively “turn trash into treasure”, endowing the [REJ] token with the role of reflecting the degree of model uncertainty. Moreover, our analysis reveals a significant correlation between the [REJ] probability and factuality (see Section §6): a higher [REJ] probability is associated with an increased likelihood of hallucinated content.

To encourage the generation of more factual answers, we introduce abstention-aware decoding. Concretely, we incorporate the uncertainty reflected by the [REJ] token into the search-based decoding strategy, *e.g.*, beam-search, to penalize uncertain predictions at each step. This strategy guides the model to navigate towards more confident and factual trajectories (Cao et al., 2022; Zhao et al., 2024). The decoding objective, incorporating an uncertainty penalty, is formalized as follows:

$$\mathbf{y}^* = \arg\max_{\mathbf{y}\in\mathcal{Y}} \left(\log p_\theta(\mathbf{y} \mid \mathbf{x}; \theta) - \lambda \cdot \mathcal{P}(\mathbf{y})\right) \tag{7}$$

where

$$\mathcal{P}(\mathbf{y}) = \frac{1}{|\mathbf{y}|}\sum_{t=1}^{|\mathbf{y}|} \log \frac{1}{1 - p_\theta\left(\texttt{[REJ]} \mid \mathbf{x}, \mathbf{y}_{<t}\right)} \tag{8}$$

At each decoding step t, the [REJ] token is used solely for regularization and will not be generated. λ quantifies the strength of the penalty applied. We set the beam size $\mathcal{B}$ to 8 and λ to 1.0.

4 Experiments

In this section, we detail our experimental setup designed to evaluate the effectiveness of SEAL in mitigating the knowledge misalignment issue. We first train the model using the constructed short-form and long-form QA training datasets separately, which include both *model-known* and *model-unknown* samples. Subsequently, we evaluate their performance under *out-of-distribution* settings.

4.1 Datasets

An overview of these datasets is shown in Table 8.

Training Datasets. To obtain high-quality training data for factual QA tasks, we utilize the June 1, 2024, Wikipedia snapshot as a reliable knowledge base. We then employ advanced open-source LLMs to generate QA pairs that are grounded in Wikipedia content. Specifically, for short-form QA tasks, we utilize few-shot prompting to guide the LLM in generating short-form QA pairs based on the abstract content of Wiki pages. For long-form QA tasks, we directly adopt the abstract content as long-form responses and leverage the LLM to generate corresponding instructions or questions. The *model-known* samples are determined by either checking the model’s accuracy or the popularity of Wikipedia pages. This process yields a total of 10,000 short-form QA pairs and 2,000 long-form QA pairs. For details on the construction of training data, please refer to Appendix A.

Evaluation Datasets. To conduct a comprehensive evaluation, we employ six mainstream factuality benchmarks, including four short-form and two long-form QA datasets. Specifically, for ***short-form QA***, we select TriviaQA (Joshi et al., 2017), Natural Questions (NQ) (Kwiatkowski et al., 2019), PopQA (Mallen et al., 2023) and SimpleQA (Wei et al., 2024a) for evaluation. As for ***long-form QA***, we evaluate performance using Biography (Min et al., 2023) and LongFact (Wei et al., 2024b). For detailed descriptions and specific examples of these datasets, please refer to Appendix B.

4.2 Evaluation

In the *short-form QA* task, we leverage accuracy (**Acc.**) to measure the extent of LLM hallucinations. Specifically, for TriviaQA, NQ, and PopQA, we follow Mallen et al. (2023) to assess correctness by determining whether ground-truth answers are included in the model generation. For SimpleQA, we follow the setting in (Wei et al., 2024a), using LLM-as-a-judge to compare the model's answer with the ground-truth. In the *long-form QA* task, for the biography dataset, we report the number of correct claims averaged per question (**# Correct**) and the **FActScore** (Min et al., 2023). FActScore is designed to evaluate the factuality of long-form responses by first decomposing them into atomic claims and then verifying them against retrieved Wikipedia paragraphs. For LongFact, we follow the evaluation metrics from (Wei et al., 2024b), reporting precision (**Prec.**), recall (**R@48**), and F1 score (**F1@48**). For more details about the evaluation metrics, please refer to Appendix C.

4.3 Baselines

We compare SEAL with the following baselines. To validate the generalizability of SEAL across different models, we select three representative LLMs for evaluation: Llama-3-8B (AI@Meta, 2024), Mistral-7B-v0.3 (Jiang et al., 2023), and Mistral-Nemo-12B (Mistral, 2024). Additional details about the baselines are provided in Appendix D.

Supervised Fine-Tuning (SFT) directly fine-tunes the pre-trained model on the constructed training dataset, aiming to maximize the likelihood of the ground-truth answers for the given questions.

POPULAR (Ghosal et al., 2024) only fine-tunes on a subset of the training dataset known to the model. For short-form QA, this subset is selected based on the accuracy of the pre-trained model's responses to questions. For long-form QA, average monthly Wikipedia page views (Mallen et al., 2023) are used as a proxy for judgment.

FLAME (Lin et al., 2024a) utilizes the pre-trained LLM as a source of supervision to first generate responses for the given questions, avoiding incorporating new factual knowledge. These self-generated responses are then utilized as ground-truth answers for SFT.

FACTTUNE (Tian et al., 2024) applies DPO (Rafailov et al., 2023) to the model fine-tuned via SFT to enhance factuality. Preference pairs are collected from the sampled outputs of the pre-trained model and annotated by either comparing with ground-truth answers or using FActScore.

4.4 Implementation Details

All experiments are conducted on eight NVIDIA A100-80GB GPUs, utilizing Deepspeed Stage 3 for multi-GPU distributed training with Bfloat16 precision enabled. To ensure a fair comparision, the training duration for all SFT-based baselines is set to 3 epochs with a learning rate of 5e-6, while the DPO-based baseline is trained for 2 epochs with a learning rate of 5e-7. For *short-form QA*, the total batch size is set to 128, and the maximum input sequence length is 128 tokens. As for *long-form QA*, the batch size is set to 32 with a maximum input length of 1024 tokens. Greedy decoding is used for all baselines to ensure the consistency of results. For more details, please refer to Appendix E.

5 Results

5.1 Main Results

We present the main results of three LLMs on both short-form and long-form QA tasks in Table 1.

SEAL achieves superior improvements in bridging the gap caused by unknown knowledge. As shown in Table 1, the *unknown samples* encountered during SFT negatively impact the factuality of pre-trained models, particularly in long-form QA scenarios. For example, `Llama-3-8B` experienced a decrease of 4.23% in average accuracy for short-form QA benchmarks, while its FActScore and F1@48 in long-form QA dropped by 24.51% and 6.32%, respectively. These findings not only align with recent research (Gekhman et al., 2024) but also extend their conclusion to long-form settings. Crucially, compared to the vanilla SFT training objective, our method achieves substantial improvements across six benchmarks, notably enhancing `Llama-3-8B` performance by an average of 10.98% on short-form QA and 19.24% (24.95 → 29.75 in FActScore), 4.17% (64.52 → 67.21 in F1@48) on long-form QA. This demonstrates the efficacy of SEAL in selectively rejecting tokens that misalign with the desired knowledge distribution, effectively bridging the gap caused by unknown knowledge.

SEAL demonstrates remarkable generalization across different models and tasks. Among all evaluated models of varying scales, SEAL consistently delivers improvements, highlighting its generalizability across different models. Notably, it also outperforms all strong baselines, achieving *state-of-the-art* performance in alleviating hallucinations induced by new knowledge. Addition-

Model	Short-form				Long-form				
	TriviaQA	NQ	PopQA	SimpleQA	Biography		LongFact		
	% Acc.	% Acc.	% Acc.	% Acc.	# Correct	FactScore	Prec.	R@48	F1@48
Llama-3-8B	67.03	37.31	33.17	9.22	29.26	33.05	73.03	70.67	68.87
+ SFT	64.09	35.68	32.30	8.71	13.87	24.95	68.79	**63.25**	64.52
+ POPULAR	65.40	38.31	35.83	9.43	17.15	28.24	66.67	61.60	63.08
+ FLAME	53.92	25.87	24.69	5.96	13.05	27.46	71.34	54.88	59.81
+ FACTTUNE	63.93	36.40	34.20	8.76	14.16	26.18	65.80	57.93	60.50
+ SEAL	**66.52** ↑ 3.79%	**39.50** ↑ 10.71%	**38.95** ↑ 20.59%	**9.48** ↑ 8.84%	**17.68**	**29.75** ↑ 19.24%	**72.82**	63.06	**67.21** ↑ 4.17%
Mistral-7B-v0.3	62.94	32.13	31.84	7.51	21.26	29.85	75.58	67.25	68.61
+ SFT	58.79	30.25	26.40	6.52	9.89	18.13	57.73	51.35	53.57
+ POPULAR	60.38	31.80	27.51	7.33	10.56	19.37	59.73	54.70	56.26
+ FLAME	41.39	18.56	18.83	4.44	9.92	18.34	**61.40**	50.62	52.67
+ FACTTUNE	59.41	30.14	26.64	6.66	10.28	19.33	60.08	51.41	54.53
+ SEAL	**60.58** ↑ 3.04%	**32.41** ↑ 7.14%	**29.04** ↑ 10.00%	**7.74** ↑ 18.71%	**13.51**	**21.68** ↑ 19.58%	60.32	**56.98**	**58.30** ↑ 8.83%
Mistral-Nemo-12B	70.16	42.44	36.90	10.89	28.89	32.30	79.39	73.70	73.62
+ SFT	68.09	37.59	33.09	9.71	15.92	28.08	64.45	57.37	59.53
+ POPULAR	68.31	39.09	32.81	9.73	14.92	28.19	68.98	59.34	62.01
+ FLAME	59.92	31.05	27.15	7.19	14.12	23.71	72.64	56.65	61.17
+ FACTTUNE	68.24	38.01	33.25	9.64	15.21	27.17	**73.22**	61.92	63.99
+ SEAL	**68.86** ↑ 1.13%	**40.33** ↑ 7.29%	**36.34** ↑ 9.82%	**9.90** ↑ 1.96%	**18.39**	**29.24** ↑ 4.13%	69.82	**65.90**	**64.79** ↑ 8.84%

Table 1: Experimental results on six short-form QA and long-form QA benchmarks. **Bold** and underline numbers indicate the best performance and second performance among all methods. And gray-colored text indicates the performance of pre-trained base models. **Arrows** indicate the relative improvement over SFT baselines.

Model	Short-form				Long-form	
	TQA	NQ	PQA	SQA	Bio.	LF.
	↑ % Acc.	↑ % Acc.	↑ % Acc.	↑ % Acc.	↑ FS.	↑ F1@48
Llama-3-8B (Ours)	**66.52**	**39.50**	**38.89**	**9.48**	**29.75**	**67.21**
w/o Decoding	65.61	36.93	36.34	9.36	27.40	65.43
w/o Tuning	64.09	35.68	32.30	8.71	24.95	64.52
Mistral-7B-v0.3 (Ours)	**60.58**	**32.41**	**29.04**	**7.74**	**21.68**	**58.30**
w/o Decoding	59.06	31.61	27.81	7.35	20.10	55.16
w/o Tuning	58.79	30.25	26.40	6.52	18.13	53.57
Mistral-Nemo-12B (Ours)	**68.86**	**40.33**	**36.34**	**9.90**	**29.24**	**64.79**
w/o Decoding	67.62	39.00	34.81	9.76	28.21	62.31
w/o Tuning	68.09	37.59	33.09	9.71	28.08	59.53

Table 2: Ablation studies results of various modules on short-form and long-form QA benchmarks. **Bold** numbers indicate the best performance among all variants.

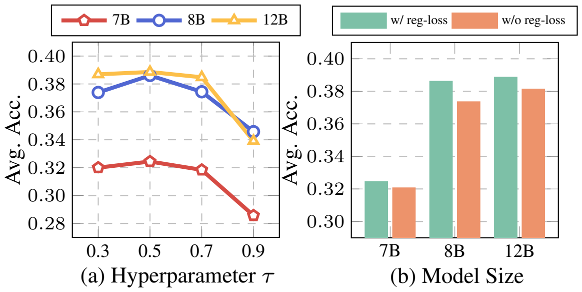
ally, we observe that the strongest baseline, POPULAR, while effective in short-form QA, struggles to maintain consistent performance in long-form QA and can even exacerbate hallucinations (64.52 → 63.08 for `Llama-3-8B` in LongFact). This decline suggests that long-form QA, which typically involves intricate factual details, poses significant challenges for filtering *model-known* data at the *sample-level*. In contrast, our approach employs selective *token-level* loss abstention based on discrepancies in model-predicted knowledge distributions, showcasing remarkable adaptiveness and robustness in complex long-form generation scenarios.

5.2 Ablation Study

To verify the effectiveness of SEAL, we conduct an extensive ablation study of its key components. We design two variants: (1) *w/o Decoding*, which replaces the abstention-aware decoding with a standard greedy decoding strategy; and (2) *w/o Tuning*, which further eliminates the abstention-tuning, reverting to the vanilla MLE objective. As shown in Table 2, it is clear that all variants underperform compared to the implementation of SEAL, indicating the effectiveness of each component. A more detailed ablation analysis is provided below.

Tuning of upper bound τ. SEAL employs a hyperparameter τ to set the upper bound of the target probability assigned to the `[REJ]` token. Typically, a higher τ encourages the model to allocate more predictions to `[REJ]`, preventing the model from overfitting to misaligned knowledge. Conversely, decreasing τ gradually degenerates into the cross-entropy loss. To investigate the impact of τ on reducing hallucinations, we adjust the value of τ from 0.3 to 0.9 and evaluate its influence on the average accuracy of three models across four short-form QA datasets. As depicted in Figure 3 (a), $\tau = 0.5$ emerges as the best choice across various settings, providing a balance that prevents the model from neglecting the learning of the downstream task while still providing sufficient signals to learn to predict `[REJ]` for confusing predictions.

Effective of regularization loss. Another key component of the abstention tuning is the regularization loss, designed to prevent the model from excessively predicting the `[REJ]` token even when a correct prediction is feasible. To validate the effectiveness of the regularization term, we provide a variant that omits this component during fine-



(a) Hyperparameter τ
(b) Model Size

Figure 3: Ablation study on hyperparameter τ and regularization loss in abstention tuning. The results displayed are the average accuracies of three LLMs on four short-form QA benchmarks. Full results are provided in Table 11 and Table 12.

Strategy	TQA	NQ	PQA	SQA
	↑ % Acc.	↑ % Acc.	↑ % Acc.	↑ % Acc.
Llama-3-8B				
DoLa-High	64.82	36.40	34.97	**9.89**
DoLa-Low	64.88	36.18	34.98	9.57
Activation	64.82	35.62	34.81	9.29
SEAL	**66.52**	**39.50**	**38.95**	9.48
Mistral-7B-v0.3				
DoLa-High	58.93	30.44	27.25	7.35
DoLa-Low	58.90	30.44	27.28	7.42
Activation	58.94	30.36	27.34	7.40
SEAL	**60.58**	**32.41**	**29.04**	**7.74**
Mistral-Nemo-12B				
DoLa-High	67.29	38.31	33.88	9.50
DoLa-Low	67.41	38.33	33.94	9.43
Activation	66.46	36.70	33.08	9.15
SEAL	**68.86**	**40.33**	**36.34**	**9.90**

Table 3: Ablation study results of different factuality-enhanced decoding strategies on short-form QA benchmarks. **Bold** numbers indicate the best performance among all decoding strategies.

tuning. The results, shown in Figure 3 (b), indicate a performance drop in all three models without the regularization term, highlighting its role in guiding appropriate target probability allocation and better aligning the [REJ] token with actual model uncertainty.

Effective of abstention-aware decoding. To further demonstrate the superiority of abstention-aware decoding, we compare it against other decoding strategies designed to enhance the factuality of LLMs: (1) DoLa (Chuang et al., 2024), which subtracts the logits in the contrastive layer to calibrate the final layer's logits. We consider two variants: DoLa-low, which leverages the first half of the layers to contrast with the final layer, and DoLa-high, which contrasts the second half with the final layer.

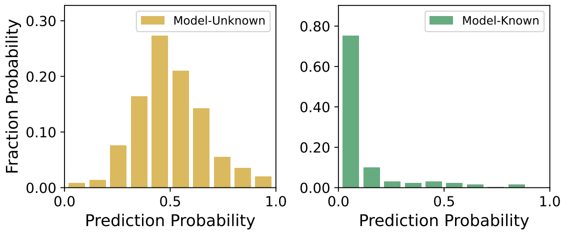


Figure 4: Histograms of the predicted [REJ] token probability for *model-unknown* (left) and *model-known* (right) samples. The x-axis, divided into 10 bins, represents the predicted probability of [REJ], and the y-axis shows the fraction of samples in each bin. The [REJ] token effectively distinguishes the two groups.

(2) Activation-decoding (Chen et al., 2024), which employs the sharpness of context activations within intermediate layers for next token prediction calibration. As shown in Table 3, abstention-aware decoding consistently outperforms other factuality-enhanced decoding strategies across four short-form QA tasks and three models after abstention tuning. Further details on additional ablation studies related to the beam size $\mathcal{B}$ and penalty α are available in Appendix F.3.

6 Analysis

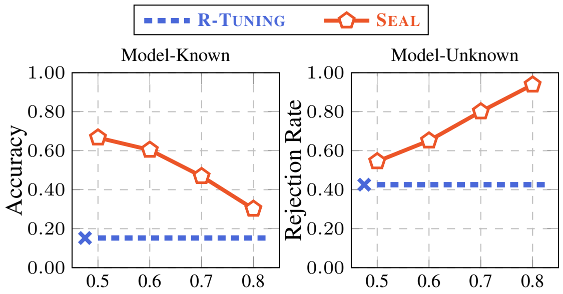
In this section, we present a deeper analysis of SEAL, exploring its potential applications and evaluating its impact and adaptability to real-world instruction-following tasks.

Calibration of [REJ] with hallucinated answers. The efficacy of our method relies heavily on the [REJ] token's ability to calibrate well with hallucinated answers. Ideally, we aim for a high probability of the [REJ] token for questions unknown to the model, indicating potential knowledge misalignment, while ensuring a relatively low probability for known questions to preserve accurate response generation. To validate this, we sample 500 data points from PopQA, evenly split between model-known and model-unknown[2]. We then manually convert the questions into a cloze format (*e.g.*, '*Question: What is George Rankin's occupation? Answer: George Rankin is a* ___') and calculate the probability of the next prediction being [REJ] for the model fine-tuned with abstention tuning. As illustrated in Figure 4, the probability of the [REJ] token distinctly differentiates between the two sce-

[2] We use Llama-3-8B as a proxy model and check its accuracy to determine whether questions are known to the model.

### Question: Who wrote he ain't heavy he's my brother lyrics?
### Answer: Bobby Scott
### Response:
SFT: The lyrics to he ain't heavy he's my brother were written by Bob Dylan.
Ours: ~~The lyrics to he ain't heavy he's my brother were written by[REJ]~~ I am not sure.

Table 4: A case study of extending SEAL in the answer refusal scenario, where the [REJ] token acts as an indicator of rejection, used to re-generate refusal responses.



(a) Hyperparameter τ (b) Hyperparameter τ

Figure 5: Comparison of SEAL and R-TUNING across varying τ, showing accuracy for *model-known* (left) and rejection rate for *model-unknown* (right).

Methods	*AlpacaEval*		*IFEval*	
	SFT	SEAL	SFT	SEAL
Llama-3-8b	70.43	**70.47** (+0.04)	53.84	**54.08** (+0.24)
Mistral-7B-v0.3	61.34	**61.70** (+0.36)	53.48	**53.96** (+0.48)
Mistral-Nemo-12B	76.56	**77.24** (+0.68)	57.79	**58.03** (+0.24)

Table 5: Results of three different LLMs on two instruction-following benchmarks

narios. Specifically, when the question is known to the model, the probability of [REJ] is exceptionally low, occurring in less than 0.1 in 75.2% of cases; conversely, for unknown question, there is a significant increase in the probability of the [REJ] token, appearing in the top 3 predictions in 77.8% of cases. These findings confirm that the [REJ] token works as expected, effectively calibrating with hallucinated answers.

Extend to answer refusal scenarios. In addition to utilizing the [REJ] token to guide more factual generation, inspired by its effective calibration, we can also extend our method to answer refusal scenarios, enabling LLMs to appropriately refuse to answer questions beyond their knowledge scope. Specifically, we adopt the greedy decoding strategy, allowing the model to generate the [REJ] token normally. Once the [REJ] token is generated, it signals the termination of the uncertain generation and prompts the model to re-generate a refusal response template *from scratch*. Upon completion, all tokens up to and including the [REJ] token are discarded before displaying the response to the user, as demonstrated in Table 4. To validate its effectiveness, we compare our method under different τ with R-TUNING (Zhang et al., 2024), which trains the model to refrain from responding to unknown questions using *model-specific* known/unknown data. We evaluate our approach from two dimensions: the accuracy of correctly answering *model-known* questions and the rejection rate of refusing to answer *model-unknown* questions. Results shown in Figure 5 demonstrate that our method's rejection rate continuously increases as τ grows, presenting a trade-off between *model-unknown* and *model-know* questions and consistently outperforming R-TUNING in both dimensions.

Impact on instruction following abilities. We further apply our training paradigm on diverse instruction fine-tuning datasets to evaluate their adaptability in instruction-following scenarios. Specifically, we fine-tune the model on the Deita dataset (Liu et al., 2024a), which consists of high-quality data selected from UltraChat (Ding et al., 2023), ShareGPT, and WizardLM (Xu et al., 2024). We evaluate our models using two widely used instruction-following benchmarks: AlpacaEval (Li et al., 2023b) and IFEval (Zhou et al., 2023b). For AlpacaEval, we use the default annotator weighted_alpaca_eval_gpt4_turbo for assessment, noted for its high human agreement, and report the raw win rate. For IFEval, we employ the instruction_loose metric for evaluation. As shown in Table 5, SEAL effectively maintains the instruction-following capabilities, confirming its adaptability across diverse downstream tasks.

7 Conclusion

This work introduces SEAL, a novel training objective with an abstention mechanism, enabling LLMs to selectively reject tokens that misalign with the desired knowledge distribution using a special token [REJ]. SEAL further enhances factuality by penalizing uncertain predictions using the uncertainty captured within [REJ]. Extensive experiments show that SEAL achieves notable improvements in both short-form and long-form QA, effectively mitigating hallucinations induced by new factual knowledge encountered during SFT.

Furthermore, further analysis demonstrates that SEAL can be extended to answer refusal scenarios and maintain instruction-following capabilities.

Limitations

This work exhibits several limitations worth noting. **Firstly**, during the training process, we use the ratio of the ground-truth probability to the current maximum predicted probability to model the LLM's confidence in its current prediction. This ratio serves as the basis for determining the shifted probability to the `[REJ]` token. While this approach is simple and effective, we have not yet explored other aggregation metrics to better capture the model's uncertainty, which can be explored in future work. **Secondly**, in the decoding stage, we incorporate the probability of the `[REJ]` token with search-based decoding strategies to further improve the model's factuality. Although effective, this may introduce greater inference overhead compared to traditional decoding strategies. Appropriately reducing the search space can alleviate this issue to some extent. Furthermore, we primarily focus on beam search due to its simplicity, but future work can explore the application of our method to other search-based decoding strategies, *e.g.*, Monte Carlo Tree Search (MCTS) (Zhao et al., 2024).

Ethics Statement

In this work, all data used for evaluating the factuality of LLMs derive from open-source public datasets, and no additional collection of sensitive information was conducted. Throughout the experimental process, all data and models were strictly utilized following their intended purposes and respective licenses. Our methodology aims to mitigate hallucinations induced by new factual knowledge encountered during fine-tuning, which has a positive impact on real-world applications by improving the factuality of LLMs. However, when deployed, our approach still carries inherent issues associated with LLMs, such as the potential for generating biased, harmful, or offensive output. Aside from this, to the best of our knowledge, there are no additional ethical issues associated with this paper.

Acknowledgments

Xiaocheng Feng is the corresponding author of this work. We thank the anonymous reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (NSFC) (grant 62276078, U22B2059), the Key R&D Program of Heilongjiang via grant 2022ZX01A32, and the Fundamental Research Funds for the Central Universities (Grant No.HIT.OCEF.2023018).

References

AI@Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: Part 3.1, knowledge storage and extraction. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Meng Cao, Yue Dong, Jingyi He, and Jackie Chi Kit Cheung. 2022. Learning with rejection for abstractive text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9768–9780. Association for Computational Linguistics.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. 2024. How do large language models acquire factual knowledge during pretraining? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. 2024. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yi Cheng, Xiao Liang, Yeyun Gong, Wen Xiao, Song Wang, Yuji Zhang, Wenjun Hou, Kaishuai Xu, Wenge Liu, Wenjie Li, Jian Jiao, Qi Chen, Peng Cheng, and Wayne Xiong. 2024. Integrative decoding: Improve factuality via implicit self-consistency. *Preprint*, arXiv:2410.01556.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. 2024. I don't know: Explicit modeling of uncertainty with an [IDK] token. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3029–3051. Association for Computational Linguistics.

Yuchun Fan, Yongyu Mu, Yilin Wang, Lei Huang, Junhao Ruan, Bei Li, Tong Xiao, Shujian Huang, Xiaocheng Feng, and Jingbo Zhu. 2025. SLAM: towards efficient multilingual reasoning via selective language alignment. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 9499–9515. Association for Computational Linguistics.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7765–7784. Association for Computational Linguistics.

Gaurav Rohit Ghosal, Tatsunori Hashimoto, and Aditi Raghunathan. 2024. Understanding finetuning for factual knowledge extraction. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Chao-Wei Huang and Yun-Nung Chen. 2024. Factalign: Long-form factuality alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 16363–16375. Association for Computational Linguistics.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yangfan Ye, Weihong Zhong, Yuxuan Gu, Baoxin Wang, Dayong Wu, Guoping Hu, and Bing Qin. 2025a. Improving contextual faithfulness of large language models via retrieval heads-induced optimization. *CoRR*, abs/2501.13573.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, and Bing Qin. 2024a. Learning fine-grained grounded citations for attributed large language models. *CoRR*, abs/2408.04568.

Lei Huang, Xiaocheng Feng, Weitao Ma, Liang Zhao, Yuchun Fan, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. 2024b. Advancing large language model attribution through self-improving. *CoRR*, abs/2410.13298.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2024. Unfamiliar finetuning examples control how language models hallucinate. *Preprint*, arXiv:2403.05612.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2024a. FLAME: factuality-aware alignment for large language models. *CoRR*, abs/2405.01525.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. 2024b. Rho-1: Not all tokens are what you need. *CoRR*, abs/2404.07965.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024a. What makes good data for alignment? A comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yujian Liu, Shiyu Chang, Tommi S. Jaakkola, and Yang Zhang. 2024b. Fictitious synthetic data can improve LLM factuality via prerequisite learning. *CoRR*, abs/2410.19290.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9802–9822. Association for Computational Linguistics.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. A discrete hard EM approach for weakly supervised question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2851–2864. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12076–12100. Association for Computational Linguistics.

Mistral. 2024. Mistral nemo.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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. *CoRR*, abs/2305.18290.

John Schulman. 2023. Reinforcement learning from human feedback: Progress and challenges. Talk given at the University of California, Berkeley on April 19, 2023.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2024. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024a. Measuring short-form factuality in large language models. *Preprint*, arXiv:2411.04368.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024b. Long-form factuality in large language models. *CoRR*, abs/2403.18802.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-tuning: Instructing large language models to say 'i don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7113–7139. Association for Computational Linguistics.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *CoRR*, abs/2411.14405.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. LIMA: less is more for alignment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023b. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911.

A Training Datasets

A.1 Data Construction Process

Given that our experimental evaluation covers both short-form and long-form QA tasks, we specifically construct high-quality training data for these two types of tasks. The high quality of the data stems from two main aspects: (1) We use Wikipedia[3] as a reliable source of knowledge, ensuring the factuality of the generated QA pairs; (2) The construction of the training data leverages one of the state-of-the-art open-source LLMs[4] through few-shot prompting, which is shown to be of high quality via our pilot human evaluation. We provide the detailed processes for constructing the training data below.

Short-form QA: For short-form QA, we use the abstract field content from Wikipedia as context, allowing the LLM to generate factual and objective questions with concise and indisputable answers based on the given context. Specifically, we sample 5 QA pairs that GPT-4o can answer correctly from the SimpleQA dataset (Wei et al., 2024a), along with the provided knowledge sources as context, to form demonstrations for few-shot prompting. The prompt template for generating short-form QA pairs is as follows:

```
Generate a short-form QA based on the
following context from Wikipedia, adhering
to these criteria: The question must seek
factual, objective knowledge with a single,
indisputable answer and specify the scope to
avoid ambiguity (e.g., “which city,” “what
year”). Ensure the reference answer is
concise, evergreen, and remains valid over
time, using precise language when needed
(e.g., specifying a season or event). Besides,
provide a one-sentence statement response to
the question.

Input: {Wikipedia Context}

Output: {Short-form QA Pairs}
```

We adopt an over-generate-then-filter strategy, initially generating 20,000 high-quality QA pairs. Given that our focus is on mitigating hallucinations induced by new factual knowledge encountered during SFT, it is crucial to maintain a balance between *model-known* and *model-unknown* samples in the training data. Therefore, we use the `Llama-3-8B` model as a proxy model, determining whether a question is known to the model based on the correctness of its answers. Ultimately, we select 6,000 *model-known* questions and 4,000 *model-unknown* questions, totaling 10,000 QA pairs as our final short-form QA training data. An example of the generated QA pair is as follows:

An Example of Short-form Wiki-QA
Question: What was the title of Chantal Kreviazuk’s debut studio album, first released in Canada in 1996? **Answer:** Under These Rocks and Stones

Table 6: An example of *short-form QA* training datasets.

Long-form QA: Similarly, for long-form QA, we also use few-shot prompting to enable the LLM to construct long-form QA pairs based on Wikipedia content. The difference lies in that we directly use the abstract field content of Wikipedia as the ground-truth answer and generate questions or instructions related to specific topics or events of the Wikipedia abstract. This ensures the reliability and factuality of the answers. Additionally, to avoid overly lengthy responses, we only select Wikipedia pages with abstract field content lengths between 180 and 220 words to construct QA pairs. Notably, to determine whether the topics in QA pairs are known or unknown to the model, we follow previous work (Mallen et al., 2023) by using the average monthly page views of the Wikipedia page as a criterion. Specifically, when the average monthly page views exceed 10,000, we consider the topic to be known by the model, and when it is less than 100, we consider the topic to be unknown. A total of 2,000 long-form QA pairs are selected, including 1,200 known and 800 unknown QA pairs. The prompt template for generating long-form QA pairs is as follows:

```
Given a specific topic (such as a person,
place, event, act, company etc.) along with
a related Wikipedia biography below, generate
a question or instruction about the topic. Do
not ask about any particular aspect of the
object (such as history, significance, impact,
role, purpose, etc.). Ensure to follow the
question styles in the provided examples.

Topic: {Wikipedia Topic}

Context: {Wikipedia Context}

Question: {Long-form Question}
```

We also provide an example of the constructed long-form training data as follows:

[3] We utilize the 2024-06-01 Wikipedia snapshot.

[4] Specifically, we use `Llama-3.1-70B`.

An Example of Long-form Wiki-QA

Question: Tell me about the movie Spaceballs.
Answer: Spaceballs is a 1987 American space opera parody film co-written, produced, and directed by Mel Brooks. (... omit part of the answer) Meanwhile, Spaceball commanders Dark Helmet (Moranis) and Colonel Sandurz (Wyner) lead the search for them but are hindered by their own incompetence. The film was released by Metro-Goldwyn-Mayer (MGM) on June 24, 1987.

Table 7: An example of *long-form QA* training datasets.

Datasets	**Statistic**	
	Type	**# Examples**
Training Datasets		
Short-Wiki	short-form	10,000
Long-Wiki	long-form	2,000
Evaluation Datasets		
TriviaQA (Joshi et al., 2017)	short-form	11,313
NQ (Kwiatkowski et al., 2019)	short-form	3,610
PopQA (Mallen et al., 2023)	short-form	14,267
SimpleQA (Wei et al., 2024a)	short-form	4,326
Biography (Min et al., 2023)	long-form	138
LongFact (Wei et al., 2024b)	long-form	114

Table 8: Data statistics of the *training* and *evaluation* datasets, covering long-form and short-form QA tasks.

A.2 Prompt Template

The prompt template for short-form QA datasets:

```
Given the following factual question, generate
an accurate and concise answer.

Question: {Question}

Answer: {Answer}
```

The prompt template for long-form QA datasets:

```
Given the following question or topic, generate
a comprehensive and detailed long-form response.
Include key details and relevant aspects to
help understand the topic.

Question: {Question}

Answer: {Answer}
```

B Evaluation Datasets

The datasets for evaluating the factuality of LLMs primarily consist of two aspects: ***short-form QA***, where questions typically test knowledge of a single factoid and answers often appear as short-form entities; and ***long-form QA***, where the questions generally inquire about a concept or object within a specific topic, often requiring long-form responses that include multiple detailed factoids. The detailed descriptions of these two datasets are as follows.

B.1 Short-form QA

We select the following four representative short-form QA datasets.

TriviaQA (Joshi et al., 2017) is a widely used QA dataset for evaluating the world knowledge of LLMs, with questions sourced from trivia and quiz-league websites. Since the ground-truth answers of the TriviaQA test set are not publicly available, we follow previous research (Min et al., 2019; Asai et al., 2024) by using the TriviaQA development set containing 11,313 samples for evaluation.

NQ (Kwiatkowski et al., 2019) is a popular open-domain QA dataset designed to reflect real-world information-seeking questions. We directly use the test set of this dataset for evaluation, which contains 3,610 questions sourced from the Google search engine, with human-annotated short-form answers grounded in Wikipedia.

PopQA (Mallen et al., 2023) is an entity-centric open-domain QA dataset that aims at evaluating the factuality of LLMs on long-tail factual knowledge. The dataset consists of 14,267 questions about long-tail entities sourced from Wikipedia, covering 16 diverse relationship types.

SimpleQA (Wei et al., 2024a) is a challenging QA dataset designed to evaluate the factuality of LLMs, comprising 4,326 short, fact-seeking questions. These questions are adversarially collected based on GPT-4's responses, and each question's answer is independently annotated by two human annotators, with only a single indisputable answer.

B.2 Long-form QA

We select the following two long-form QA datasets.

Biography (Min et al., 2023) is a long-form factuality QA dataset consisting of a set of prompts that require LLMs to generate lengthy biographies covering specific information. The dataset contains 183 annotated and 500 unannotated human entities sampled from Wikipedia articles, covering varying frequency levels. We directly use the 183 annotated prompts for evaluation, following Lin et al. (2024a).

LongFact (Wei et al., 2024b) consists of two subtasks: LongFact-Concepts and LongFact-Objects. The former involves prompts designed to inquire about general concepts, while the latter focuses more on specific objects. Each subtask includes 38 manually selected topics, with 30 unique prompts generated for each topic, totaling 1,140 prompts per task. In our evaluation, we follow prior study (Huang and Chen, 2024) by focusing on the more challenging LongFact-Objects task. Considering the high cost of evaluating LongFact due to the significant number of API calls required, we follow the setting in Cheng et al. (2024) by sampling 114 prompts to form our final evaluation dataset.

B.3 Examples

We provide specific examples of short-form and long-form QA evaluation datasets as shown in Table 9 and Table 10.

Short-form QA Examples	
TriviaQA	**Question:** To the nearest two, how many tennis Grand Slam titles did Jimmy Connors win? **Answer:** ["10", "ten"]
NQ	**Question:** When was the last time anyone was on the moon? **Answer:** ["14 December 1972 UTC", "December 1972"]
PopQA	**Question:** What is Fritz Goos's occupation? **Answer:** ["astronomer", "physicist"]
SimpleQA	**Question:** Who received the IEEE Frank Rosenblatt Award in 2010? **Answer:** Michio Sugeno

Table 9: Examples of *short-form QA* evaluation datasets.

Long-form QA Examples	
Biography	**Question:** Tell me a bio of Kang Ji-hwan.
LongFact	**Question:** What is the Agilent High Performance Liquid Chromatography (HPLC) device used for in analytical chemistry?

Table 10: Examples of *long-form QA* evaluation datasets.

C Evaluation

C.1 Evaluation for Short-form QA

In our constructed training data, short-form answers are presented in the form of complete statements. For example, for the question *'When was the last time anyone was on the moon?'*, after SFT training, the model's response would likely be: *'The last time anyone was on the moon was during the Apollo 17 mission in December 1972'*. For TriviaQA, NQ, and PopQA, since their answers consist of multiple candidate short-form entities (as shown in Table 9), when evaluating model performance, we follow previous research (Mallen et al., 2023) by directly determine whether a gold answers is included in the model generations instead of requiring an exact matching. For SimpleQA, we directly follow its evaluation setup (Wei et al., 2024a), using LLM as a grader to compare the model's responses with the unique ground truth answer and grades responses as either "correct", "incorrect" or "not attempted". Since there are no refusal responses in the SFT data, the model can hardly refuse to answer. Therefore, we directly use the percentage of all questions answered correctly as the model's accuracy. Additionally, considering the extensive use of GPT-4o's API during evaluation, we adopt a more cost-effective solution by using Llama-3.1-70B, one of the most powerful open-source models available, as the grader. This evaluation approach has demonstrated a high correlation with GPT-4o in our preliminary experiments.

C.2 Evaluation for Long-form QA

The evaluation process for long-form responses typically involves two steps: (1) breaking down the long-form generation into a series of atomic facts, and (2) assessing the factuality of each fact with external retrieval. For Biography, we follow the evaluation setup in (Min et al., 2023), using *FActScore* for evaluation. We first utilize `Llama-3.1-70B-Instruct` for atomic facts decomposition, following previous research (Cheng et al., 2024). When assessing factuality, we employ `retrieval+llama+npm` as the evaluator due to its strong correlation with human judgment. For LongFact, we adhere to its original evaluation setting, employing metrics including the proportion of truthful facts (*Precision*), the number of truthful facts divided by 48 (*Recall@48*), and a combination of the two, *F1@48*. During the evaluation process, we also utilize `Llama-3.1-70B-Instruct` to break down the model generation into atomic facts, which are then assessed their factuality by DeepSeek-V3, considering its strong capability and cost-effectiveness.

D Baselines

The detailed descriptions and implementation details of all baselines are as follows:

SFT: We directly fine-tune pre-trained LLMs using the standard supervised fine-tuning objective (Fan et al., 2025), with a total of 10,000 samples fine-tuned for short-form QA and 2,000 samples for long-form QA.

POPULAR (Ghosal et al., 2024): We directly fine-tune the pre-trained model on questions it is already familiar with, avoiding the introduction of unfamiliar knowledge. However, the key issue lies in how to distinguish between problems that the model knows and those it does not. For short-form QA, we let the model perform greedy decoding to generate corresponding responses, which are then compared with ground-truth answers. Using `Llama-3.1-70B` as a judge to determine the correctness of the responses, questions that the model can answer correctly are considered known by the model. Specifically, from the original 10,000 short-form QA training data entries, 6,000, 4,920, and 4,215 *model-specific* known samples are finally selected for training `Llama-3-8B`, `Mistral-7B-v0.3`, and `Mistral-Nemo-12B`, respectively. As for long-form QA, following previous work (Mallen et al., 2023), we use Wikipedia page views per month as a proxy; entities with an average monthly view count greater than 10,000 are considered familiar to the model, while those with less than 100 views per month are deemed unfamiliar entities. Therefore, for all models, the number of *model-known* and *model-unknown* samples is kept consistent, at 1200 and 800, respectively.

FLAME (Lin et al., 2024a): For both short-form and long-form QA tasks, we additionally sample five question-answering pairs from outside the constructed training dataset as demonstrations for few-shot prompting. With 5-shot demonstrations, we use vanilla pre-trained models (`Llama-3-8b`, `Mistral-7b-v0.3` or `Mistral-Nemo-12b`) to sample five responses for each question. During sampling, we set the temperature to 1.0 and `top_p` to 0.95. As a result, we obtain a total of 50,000 samples for short-form QA and 10,000 samples for long-form QA. We follow the original setup in Lin et al. (2024a) and directly use these sampled responses as ground-truth answers to perform supervised fine-tuning on pre-trained models.

FACTTUNE (Tian et al., 2024): We directly utilize the samples generated from FLAME on top of the SFT model and annotate them using reference-based truthfulness metrics. For short-form QA, we compare the ground-truth answers with the sampled responses, selecting one correct and one incorrect sample as preference pairs. Questions where all samples are either correct or incorrect are skipped. For long-form QA, we score long-form generations using FactScore, choosing the samples with the highest and lowest FactScore as preference pairs. During the DPO training process, we set β to 0.1 and the learning rate to 5e-7, with the entire process fine-tuned for two epochs.

E Implementation Details

In the abstention tuning stage, we adopt the same setups as the SFT baseline and train the model on the constructed short-form and long-form QA training datasets separately. Each QA pair in the training dataset is transformed into an (instruction, response) pair following the template outlined in Appendix A.2. For training, we configure the upper bound τ to 0.5, set the number of training epochs to 3, and set the learning rate to 5e-6. Additionally, the warmup ratio is set to 0.03 with a linear learning rate scheduler. The total batch size is 128 for short-form QA, while for long-form QA, it is set to 32.

In the decoding stage, we apply the chat template used in the training stage across all evaluation datasets. For all baselines, we employ a greedy decoding strategy to ensure consistency in the generated outputs. Additionally, we utilize the vLLM framework (Kwon et al., 2023) for efficient inference. For short-form QA, the `max_new_tokens` is set to 128, whereas for long-form QA, it is set to 1024. For our proposed abstention-aware decoding, we do not allow the model to generate the `[REJ]` token for all main experiments, setting its generation probability to negative infinity. We use a beam size of 8 for short-form QA while considering the substantial inferent cost for long-form generation, and we appropriately reduce the search space for long-form QA, setting the beam size to 6. As for the penalty coefficient λ, we set it to 1.0 for both short-form and long-form QA.

F Ablation Study

In this section, we present the complete results of the ablation studies on the upper bound and regularization loss, along with additional analyses of the beam size and rejection penalty.

F.1 Tuning of Upper Bound τ

The full results of the ablation study regarding the tuning of the upper bound τ are shown in Table 11. Among all the settings, $\tau = 0.5$ emerges as the best choice.

Model	TQA	NQ	PQA	SQA	Avg.
	↑ % Acc.	↑ % Acc.	↑ % Acc.	↑ % Acc.	
Llama-3-8B					
$\tau = 0.3$	65.48	37.84	37.02	9.20	37.39
$\tau = 0.5$	**66.52**	**39.50**	**38.95**	**9.48**	**38.61**
$\tau = 0.7$	65.44	37.98	36.97	9.36	37.44
$\tau = 0.9$	61.13	34.21	34.98	8.00	34.58
Mistral-7B-v0.3					
$\tau = 0.3$	60.05	31.58	28.93	7.44	32.00
$\tau = 0.5$	**60.58**	**32.41**	**29.04**	**7.74**	**32.44**
$\tau = 0.7$	58.95	31.83	29.54	7.03	31.84
$\tau = 0.9$	54.50	28.09	25.46	6.24	28.57
Mistral-Nemo-12B					
$\tau = 0.3$	68.19	**40.41**	36.26	9.89	38.69
$\tau = 0.5$	**68.86**	40.33	**36.34**	**9.90**	**38.86**
$\tau = 0.7$	67.48	39.92	36.61	9.92	38.48
$\tau = 0.9$	61.60	34.49	31.56	8.00	33.91

Table 11: The results of ablation studies with varying τ on the short-form QA benchmarks. **Bold** numbers indicate the best performance among all settings.

F.2 Regularization Loss

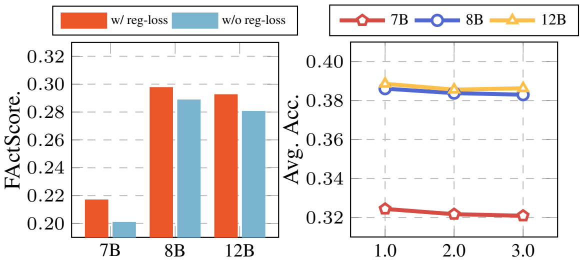
The full results of the regularization loss ablation study across four short-form QA datasets are presented in Table 12. Additionally, considering that evaluating the LongFact dataset requires a substantial number of API calls, we provide the ablation results on the Biography dataset for long-form QA in Figure 6 (a). The results are consistent with those observed in short-form QA, further underscoring the importance of the regularization loss.

Model	TQA	NQ	PQA	SQA	Avg.
	↑ % Acc.	↑ % Acc.	↑ % Acc.	↑ % Acc.	
Llama-3-8B	**66.52**	**39.50**	**38.95**	**9.48**	**38.61**
w/o Reg	65.58	37.31	37.24	9.27	37.35
Mistral-7B-v0.3	**60.58**	**32.41**	**29.04**	**7.74**	**32.44**
w/o Reg	60.10	32.35	28.75	7.03	32.06
Mistral-Nemo-12B	**68.86**	**40.33**	**36.34**	**9.90**	**38.86**
w/o Reg	68.04	39.09	35.53	9.87	38.13

Table 12: The results of ablation studies on the regularization loss on short-form QA datasets. **Bold** numbers indicate the best performance among all variants.

F.3 Beam Size $\mathcal{B}$ and Rejection Penalty λ.

We provide the ablation study for beam size $\mathcal{B}$ in Table 13 and rejection penalty λ in Figure 6 (b).



(a) Model Size (b) Hyperparameter λ

Figure 6: Ablation study on regularization loss on the Biography dataset and rejection penalty λ in abstention-aware decoding.

Model	TQA	NQ	PQA	SQA	Avg.
	↑ % Acc.	↑ % Acc.	↑ % Acc.	↑ % Acc.	
Llama-3-8B					
$\mathcal{B} = 4$	66.47	38.98	38.82	9.50	38.44
$\mathcal{B} = 6$	66.48	39.28	38.89	**9.57**	38.56
$\mathcal{B} = 8$	**66.52**	**39.50**	**38.95**	9.48	**38.61**
Mistral-7B-v0.3					
$\mathcal{B} = 4$	60.36	32.05	**29.27**	7.56	32.31
$\mathcal{B} = 6$	60.44	**32.58**	29.12	7.61	32.44
$\mathcal{B} = 8$	**60.58**	32.41	29.04	**7.74**	**32.44**
Mistral-Nemo-12B					
$\mathcal{B} = 4$	68.74	40.11	36.17	8.88	38.48
$\mathcal{B} = 6$	68.78	40.25	36.28	9.71	38.76
$\mathcal{B} = 8$	**68.86**	**40.33**	**36.34**	**9.90**	**38.86**

Table 13: The results of ablation studies with varying beam size $\mathcal{B}$ on short-form QA benchmarks. **Bold** numbers indicate the best performance among all settings.

Considering the substantial cost of evaluating long-form QA, we primarily conducted ablation experiments of beam size $\mathcal{B}$ on short-form QA. As shown in Table 13, increasing the beam size improves the average performance across four short-form QA datasets for three different model sizes. This is because the increase in $\mathcal{B}$ enlarges the hypothesis search space, thereby increasing the likelihood of finding potential answers. However, increasing the beam size also brings greater overhead. Thus, there is a trade-off between decoding efficiency and factual accuracy.

As for the rejection penalty, these results show that the $\lambda = 1.0$ we selected is the optimal setting. This also highlights the importance of an appropriate uncertainty penalty term.