UALIGN: Leveraging Uncertainty Estimations for Factuality Alignment on Large Language Models

Anonymous ACL submission

Abstract

Despite demonstrating impressive capabilities, Large Language Models (LLMs) still often struggle to accurately express the factual knowledge they possess, especially in cases where the LLMs' knowledge boundaries are ambiguous. To improve LLMs' factual expressions, we propose the UALIGN framework, which leverages **U**ncertainty estimations to represent knowledge boundaries, and then explicitly incorporates these representations as input features into prompts for LLMs to **Align** with factual knowledge. First, we prepare the dataset on knowledge question-answering (QA) samples by calculating two uncertainty estimations, including confidence score and semantic entropy, to represent the knowledge boundaries for LLMs. Subsequently, using the prepared dataset, we train a reward model that incorporates uncertainty estimations and then employ the Proximal Policy Optimization (PPO) algorithm for factuality alignment on LLMs. Experimental results indicate that, by integrating uncertainty representations in LLM alignment, the proposed UALIGN can significantly enhance the LLMs' capacities to confidently answer known questions and refuse unknown questions on both in-domain and out-of-domain tasks, showing reliability improvements and good generalizability over various prompt- and training-based baselines.

1 Introduction

Despite the remarkable proficiency of large language models (LLMs) across a diverse range of tasks (Touvron et al., 2023; OpenAI, 2023; Chiang et al., 2023), they still frequently face challenges in accurately expressing factual knowledge that they learned from the pre-training stage but are uncertain about. In such cases, the knowledge boundaries are somewhat ambiguous by LLMs, remaining a gap between "known" and "expression" (Lin et al., 2024; Zhang et al., 2024b; Li et al., 2024), which may lead to the hallucination problem and undermine the reliability and applicability to users.

LLMs typically generate responses ("expression") based on knowledge distributions learned during pre-training ("known"). However, much of the knowledge acquired during this phase exhibits vague boundaries, comprising numerous learned but uncertain knowledge pieces (*weakly known, light green area of spectrum* in Fig. 1 (a)) (Gekhman et al., 2024). Hence, LLMs may not confidently convey accurate information in downstream tasks even though they hold relevant knowledge but don't make sure (Zhang et al., 2024b). Additionally, LLMs may exhibit overconfidence in the knowledge they are unfamiliar with (*unknown, the gray area of spectrum* in Fig. 1 (a)), leading to fabricated or hallucinatory content (Zhang et al., 2024a; Liu et al., 2024). This issue primarily arises from that LLMs don't properly reconcile the knowledge boundaries with factual accuracy during alignment (Tian et al., 2024). Unlike previous works that focused on reinforcement learning (RL) through knowledge feedback or factuality alignment (Liang et al., 2024; Xu et al., 2024a; Tian et al., 2024; Lin et al., 2024; Zhang et al., 2024b; Yang et al., 2024), our objective is to elicit LLMs' weakly known facts and extend beyond merely discerning unknown facts by explicitly utilizing knowledge boundaries in alignment. We aim to leverage the knowledge boundary information of LLMs to instruct LLMs to confidently express their known yet uncertain information and firmly refuse questions beyond their knowledge as in Fig. 1 (b). Based on improvements of "known", LLMs' expressions are more truthful and reliable, thereby minimizing the discrepancy between "known" and "expression" (Lin et al., 2024; Zhang et al., 2024b; Li et al., 2024).

Inspired by the aforementioned analysis, we propose the UALIGN framework, which strategically models **U**ncertainty regarding knowledge boundary representations, subsequently **Align**ing these esti-

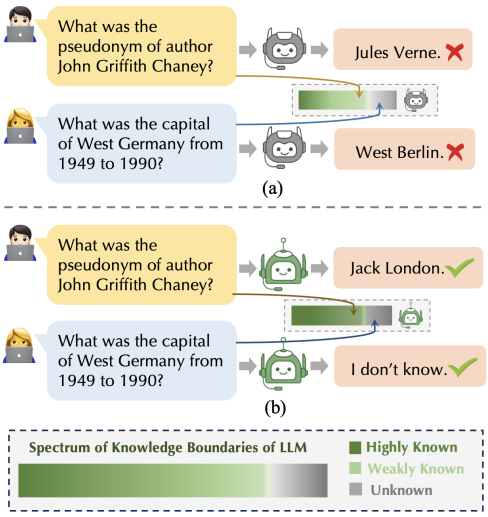


Figure 1: Examples of LLMs with (a) ambiguous and (b) explicit knowledge boundaries to answer questions.

mations with factuality. Therefore, the UALIGN framework focuses on two pivotal issues: how to capture the knowledge boundary representations and how to align with factuality.

First, we prepare the dataset that incorporates knowledge boundary information for alignment in the UALIGN framework. Knowledge boundaries always indicate the known level of factual knowledge, generally implemented using uncertainty estimation methods on LLMs (Ren et al., 2023). To precisely capture the intrinsic perception of knowledge boundary representations given the knowledge QA datasets, we adopt two uncertainty estimations of accuracy-based confidence score (Xiong et al., 2024) and semantic entropy (Kuhn et al., 2023) respectively. We sample multiple responses to a question using varied prompting and temperature sampling to approximate actual knowledge boundaries by calculating the confidence and entropy of each question. The two measures (Kuhn et al., 2023; Xiong et al., 2024), as complementary, can reflect the convince and dispersion of generated responses to a question based on LLMs' internal knowledge. Questions with at least one correct sampled answer are regarded as "known", and those with all incorrect sampled responses are considered "unknown". We revise ground-truth answers to unknown questions to refusal responses to delineate known and unknown facts (Zhang et al., 2024a).

Second, following Ouyang et al. (2022), we explicitly leverage the uncertainty estimations to align with factuality on the prepared dataset using both supervised fine-tuning (SFT) and reinforcement learning (RL). We employ SFT to train two uncertainty estimation models to predict confidence and entropy, and then train a reward model to evaluate the correctness of the generated answer conditioned on the input comprising the question, the generated response, and two uncertainty estimations regarding the knowledge boundary. With the reward model, we further adopt the Proximal Policy Optimization (PPO) (Schulman et al., 2017) algorithm for LLM alignment by feeding both questions and two measures as prompts to elicit the policy LLM's factual expressions to improve the reliability.

Experiments are conducted to evaluate in-domain and out-of-domain performance on a range of knowledge QA datasets. The results demonstrate our proposed UALIGN method significantly enhances the reliability and generalization for LLMs over several baseline methods to accurately express known factual knowledge and refuse unknown questions, suggesting that leveraging the two employed uncertainty estimations in alignment can notably improve LLMs' factuality.

In summary, our contributions are as follows.

1) To the best of our knowledge, UALIGN is the first to explicitly leverage the uncertainty estimations representing knowledge boundaries for LLM alignment, heralding a promising direction for future research of LLM training [1].

2) We demonstrate that jointly incorporating confidence and semantic entropy into prompts can provide precise knowledge boundary information to elicit LLMs' factual expressions.

3) We conduct main experiments by comparing our UALIGN with various baselines as well as ablation studies, validating the reliability improvements and robust generalization of the UALIGN method.

2 Methodology

The proposed UALIGN framework is introduced in this section with two parts: The Sec. 2.1 involves the UALIGN dataset preparation process, including strategies to collect multiple responses, as well as uncertainty measures to capture intrinsic representations of knowledge boundary on knowledge-based QA pairs as illustrated in Fig. 2. The Sec. 2.2 utilizes the obtained UALIGN dataset to train the uncertainty estimation models, and further explicitly incorporate the estimations as input features to elicit LLMs to generate factual responses using SFT- and PPO-based alignment methods as shown in Fig. 3 and Algorithm 1.

[1]The codes will be released on GitHub.

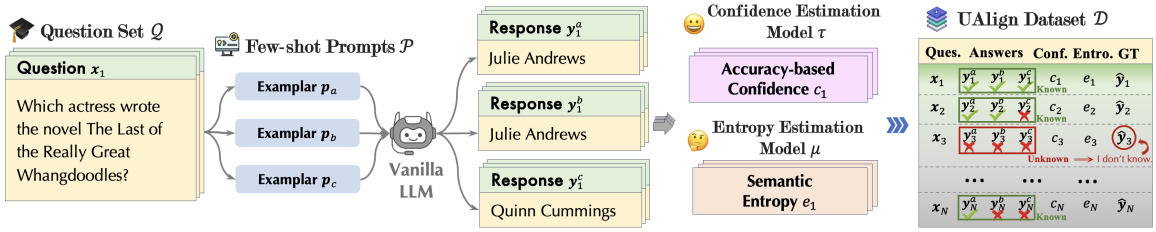


Figure 2: Illustration of UALIGN dataset preparation process.

2.1 Dataset Preparation

2.1.1 Responses Sampling Strategy

As in Fig. 2, to explore the knowledge boundary of the LLM given a question, we sample multiple responses by repeating the generation procedure several times. In this phase, the preparation process can be represented in a tuple $(\mathcal{Q}, \mathcal{P}, \mathcal{A})$. $\mathcal{Q}$ contains a batch of N QA pairs $\{(\boldsymbol{x}_i, \hat{\boldsymbol{y}}_i)\}_{i=1}^{N}$ where $\boldsymbol{x}_i$ and $\hat{\boldsymbol{y}}_i$ denote the i-th question and ground-truth answer respectively. To mitigate context sensitivity, we utilize different few-shot prompts in $\mathcal{P}$ with temperature $T = 0.2$ to make a trade-off between the accuracy and diversity to represent knowledge boundaries (Gekhman et al., 2024). The few-shot prompt set $\mathcal{P}$ consists of K different 1-shot exemplars in this work which is enough for LLMs to generate answers in the correct format. We present the few-shot prompts for sampling on TriviaQA and SciQ datasets as exemplified in Appendix J.

In the k-th sampling process for the i-th question $\boldsymbol{x}_i$, we employ each few-shot exemplar $\boldsymbol{p}_k \in \mathcal{P}$ with the question $\boldsymbol{x}_i$ to the LLM to generate the k-th response ${\boldsymbol{y}_i}^{(k)}$. By taking K times of the sampling process, we can obtain an answer set $\boldsymbol{Y}_i = \{{\boldsymbol{y}_i}^{(k)}\}_{k=1}^{K}$ to $\boldsymbol{x}_i$. We set the labels $\boldsymbol{Z}_i = \{{z_i}^{(k)}\}_{k=1}^{K}$ by comparing each generated answer ${\boldsymbol{y}_i}^{(k)}$ with the ground-truth $\hat{\boldsymbol{y}}_i$ to indicate the correctness (${z_i}^{(k)} \in \{0, 1\}$, 1 for *True* and 0 for *False*). We collect and format the data in $(\boldsymbol{x}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i, \hat{\boldsymbol{y}}_i)$ in an extended dataset and calculate the uncertainty measures subsequently. Note that since fine-tuning LLMs on unknown knowledge will encourage hallucinations (Zhang et al., 2024a; Gekhman et al., 2024), we revise the ground-truth answer to the question with ${z_i}^{(k)} = 0, \forall {z_i}^{(k)} \in \boldsymbol{Z}_i$ to "*Sorry, I don't know.*" to teach LLMs to refuse the questions beyond their knowledge (Zhang et al., 2024a).

2.1.2 Uncertainty Measures

In order to quantify the knowledge boundaries, we can leverage some uncertainty estimation methods. The knowledge boundary of LLMs in this work is defined in two aspects. The first involves the prior judgment to a question $\boldsymbol{x}_i$ regardless of the answers (Ren et al., 2023) which indicates the certainty level of $\boldsymbol{x}_i$. The second entails the dispersion measure to the distribution of the generated responses in $\boldsymbol{Y}_i$ to $\boldsymbol{x}_i$. Accordingly, we adopt accuracy-based confidence (Xiong et al., 2024) and semantic entropy (Kuhn et al., 2023) to jointly determine and represent the actual knowledge boundary information.

Accuracy-based Confidence A natural idea of aggregating varied responses is to measure the accuracy among the candidate outputs to denote confidence scores (Manakul et al., 2023; Xiong et al., 2024). Given a question $\boldsymbol{x}_i$, the accuracy of candidate responses in $\boldsymbol{Y}_i$ by comparing with the ground-truth answer $\hat{\boldsymbol{y}}_i$ serves as the confidence score c_i, computed as follows.

$$c_i = \mathsf{Conf}(\boldsymbol{x}_i) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}\left(\tilde{\boldsymbol{y}}_i = {\boldsymbol{y}_i}^{(k)}\right) \quad (1)$$

Semantic Entropy Due to the variable length and semantically equivalent generated sequences in sentence-level output spaces, Kuhn et al. (2023) proposes semantic entropy to capture uncertainty on the semantic level to quantify the degree of dispersion of sentence meanings. The semantic entropy e_i given $\boldsymbol{x}_i$ and $\boldsymbol{Y}_i$ is calculated as

$$p(s|\boldsymbol{x}_i) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}\left[{\boldsymbol{y}_i}^{(k)} \in s\right] \quad (2)$$

$$e_i = \mathsf{SE}(\boldsymbol{x}_i) = -\sum_{s} p(s|\boldsymbol{x}_i) \log p(s|\boldsymbol{x}_i) \quad (3)$$

where s denotes a set of sentences in semantic equivalent space. As illustrated in Fig. 1, semantic entropy is calculated by clustering semantically equivalent responses, as a measure to quantify the dispersion of generations to confirm the correct answer despite the low confidence, which will be further analyzed with the experimental results in Sec. 4.2. We calculate the confidence score and semantic entropy for both known and unknown questions.

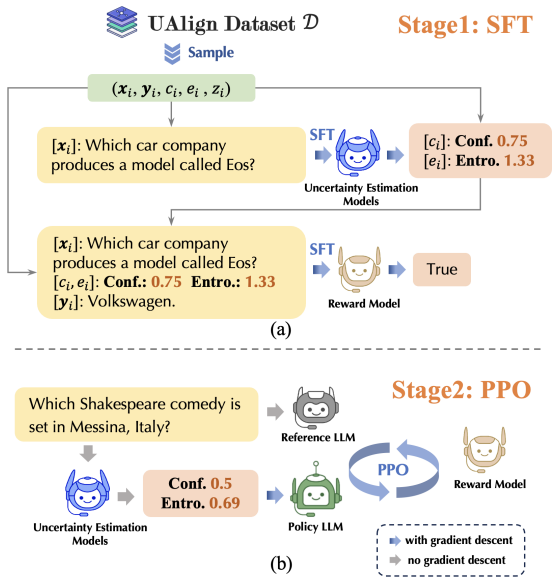


Figure 3: Illustration of (a) SFT and (b) PPO alignment processes of UALIGN framework. Note that for simplicity, we only present one estimation model in the figure but there are actually two.

Algorithm 1 UALIGN Training Algorithm

1: **Input:** UALIGN dataset $\mathcal{D}$, uncertainty models τ μ, reward model θ, initial policy π_o.
2: **Output:** Optimized policy π_θ.
3: **Stage 1: UALIGN SFT**
4: Train uncertainty models τ μ on $\mathcal{D}$ to predict c_i, e_i by feeding $\boldsymbol{x}_i$ using Eq. 4 and 5.
5: Train reward model θ on $\mathcal{D}$ to predict z_i by feeding $\boldsymbol{x}_i, c_i, e_i, \boldsymbol{y}_i^{(k)}$ using Eq. 6.
6: **Stage 2: UALIGN PPO**
7: Collect reward r including the reward signal r_1 by θ and KL-penalty r_2 between policy π_θ and initial policy π_o as Eq. 7.
8: Update policy π_θ using the collected reward r.

Then we update a UALIGN dataset $\mathcal{D}$ by formatting the i-th sample in $(\boldsymbol{x}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i, \hat{\boldsymbol{y}}_i, c_i, e_i)$.

2.2 UALIGN Training Process

2.2.1 UALIGN SFT: Uncertainty Estimation and Reward Models Training

As presented in Fig. 3 (a) and Algorithm 1, given dataset $\mathcal{D}$, UALIGN SFT is to train uncertainty estimation models to explicitly learn the two estimations given specific questions. Uncertainty estimation models of τ and μ are utilized to predict the confidence score and semantic entropy respectively, which are continuously used to train a reward model. When training τ and μ, we only feed a question $\boldsymbol{x}_i$ to the models to generate two uncertainty estimations. The training objectives are to minimize the cross-entropy losses $\mathcal{L}_\tau$ and $\mathcal{L}_\mu$ as

$$\arg\min_{\tau} \mathcal{L}_\tau, \arg\min_{\mu} \mathcal{L}_\mu,$$

$$\mathcal{L}_\tau = -\mathbb{E}_{(\boldsymbol{x}_i, c_i)\sim\mathcal{D}} \left[\log p_\tau(c_i|\boldsymbol{x}_i)\right] \quad (4)$$

$$\mathcal{L}_\mu = -\mathbb{E}_{(\boldsymbol{x}_i, e_i)\sim\mathcal{D}} \left[\log p_\mu(e_i|\boldsymbol{x}_i)\right] \quad (5)$$

where the models can explicitly learn and express the uncertainty estimations which represent more accurate knowledge boundary information.

Subsequently, the reward model is introduced as a binary evaluator to determine if a generated answer $\boldsymbol{y}_i^{(k)} \in \boldsymbol{Y}_i$ is correctly conditioned on the question $\boldsymbol{x}_i$, confidence c_i, and entropy e_i. Both c_i and e_i are explicitly used as additional auxiliary features to improve the accuracy of the reward model. The binary cross-entropy loss $\mathcal{L}_\theta$ for the reward model θ is minimized as follows.

$$\arg\min_{\theta} \mathcal{L}_\theta, \mathcal{L}_\theta = -\mathbb{E}_{(\boldsymbol{x}_i, \boldsymbol{y}_i^{(k)}, z_i^{(k)}, c_i, e_i)\sim\mathcal{D}}[\mathcal{L}_\theta^{(i)}]$$

$$\mathcal{L}_\theta^{(i)} = -z_i^{(k)} \log p_\theta(z_i^{(k)}|\boldsymbol{x}_i, c_i, e_i, \boldsymbol{y}_i^{(k)}) - (1 - z_i^{(k)}) \log(1 - p_\theta(z_i^{(k)}|\boldsymbol{x}_i, c_i, e_i, \boldsymbol{y}_i^{(k)})) \quad (6)$$

2.2.2 UALIGN PPO: Policy Model Training

The UALIGN PPO is to elicit the LLM's factual expressions to a question with the uncertainty measures using obtained models. Inspired by the progress of reinforcement learning from human feedback (RLHF) technique (Ouyang et al., 2022; Ziegler et al., 2019), we employ proximal policy optimization (PPO) (Schulman et al., 2017) for LLM optimization with the reward model θ. As illustrated in Fig. 3 (b), the LLM to be optimized is used as the policy π_θ. During this phase, we iteratively feed the question $\boldsymbol{x}$, and the predicted confidence c and entropy e to both the policy π_θ and the reference π_o, and the reward function r will facilitate reliable expressions of $\boldsymbol{y}$ of the policy model π_θ. Model update details are further specified in Appendix B.1. The training objective is to maximize the following reward function r as

$$\arg\max_{\pi_\theta} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}, c\sim\tau(\boldsymbol{x}), e\sim\mu(\boldsymbol{x}), \boldsymbol{y}\sim\pi_\theta(\boldsymbol{x}, c, e)} [r]$$

$$r = \underbrace{\theta(\boldsymbol{x}, \boldsymbol{y}, c, e)}_{r_1} - \beta \underbrace{\mathrm{KL}[\pi_\theta(\boldsymbol{x}, c, e)||\pi_o(\boldsymbol{x})]}_{r_2} \quad (7)$$

where the reward function r contains a reward signal r_1 from θ and a KL-penalty r_2 to make sure the generated answers $\boldsymbol{y}$ by policy π_θ don't diverge too much from the original policy π_o. The hyperparameter β is the coefficient of KL-penalty.

3 Experimental Setting

3.1 Datasets

The UALIGN training set is comprised of three widely used knowledge-intensive QA datasets: **TriviaQA** (**TVQA**) (Joshi et al., 2017) which contains closed-book trivia QA pairs to gauge models' factual knowledge, **SciQ** (Johannes Welbl, 2017) requiring scientific professional knowledge, and **NQ-Open** (Kwiatkowski et al., 2019) which is constructed by Google Search queries along with annotated short answers or documents.

For testing, we evaluate the in-domain (ID) performance on the corresponding validation/test sets and generalization on an out-of-domain (OOD) test set **LSQA** (Xue et al., 2024) which contains multilingual language-specific QA pairs. More dataset details and statistics are presented in Appendix C.

3.2 Evaluation Metrics

To evaluate LLMs' reliability, we employ two metrics: *Precision* (*Prec.*) and *Truthfulness* (*Truth.*). *Precision* is defined as the proportion of correctly answered questions among all the known questions, representing LLMs' ability to accurately express their known factual knowledge. *Truthfulness* represents the proportion of the sum of correctly answered known and refused unknown questions among all questions, indicating LLMs' honesty level. Details can be referred to Appendix D.1.

To ascertain the correctness of the LLM-generated answer $\boldsymbol{y}$ with the ground truth $\hat{\boldsymbol{y}}$, we employ a string-matching approach. Exact matching (EM) of $\boldsymbol{y} \equiv \hat{\boldsymbol{y}}$ always misjudges some correct answers with slight distinctions on such closed-book QA tasks. Therefore, we replace EM with a variant of $\boldsymbol{y} \in \hat{\boldsymbol{y}} \vee \hat{\boldsymbol{y}} \in \boldsymbol{y}$ to evaluate the accuracy. The specific illustrations of evaluation formulas and comparisons of several EM variants we tested with human evaluations are in Appendix D.2.

3.3 Baselines

We present several baselines in four categories below. To clearly delineate the differences between our proposed method and other baselines, we have illustrated all methods in Fig. 7 in Appendix E.

Prompt-based We present two prompt-based baselines namely In-Context Learning (**ICL**), In-Context Learning with Refusal Examples (**ICL-IDK**), and In-Context Learning Chain-of-Thought (**ICL-CoT**) (Wei et al., 2022). The few-shot prompt templates are presented in Appendix F.

SFT-based We employ standard Supervised Fine-Tuning (**SFT**) by training an LLM to generate answers for all questions. We also introduce **R-Tuning** (Zhang et al., 2024a) which teaches LLM to refuse their unknown questions.

RL-based Following RLHF technique (Ouyang et al., 2022), we first train a reward model to determine correctness by SFT. Then we employ PPO to optimize the policy model with the reward model (**RL-PPO**). We also introduce an advanced variant called reinforcement learning from knowledge feedback (**RLKF**) (Liang et al., 2024) which leverages knowledge probing and consistency checking to train the reward model. Following Zhang et al. (2024b); Tian et al. (2024); Lin et al. (2024), we also construct the factuality preference dataset to conduct direct preference optimization (**RL-DPO**) to enhance the factuality of LLMs.

Inference-based Another branch of work focuses on shifting the output distribution to improve factuality during inference. Li et al. (2023) (**ITI**) intervenes in the activations in attention heads to the "truthfulness" direction.

3.4 Implementation Details

Experiments are conducted on two LLMs: **Llama-3-8B** (Llama-3) [2] (AI@Meta, 2024) and **Mistral-7B** (Mistral) [3] (Jiang et al., 2023). When preparing the UALIGN dataset, we sample 10 responses for each question on $K = 10$ different 1-shot prompts. The sampling temperature T is set to 0.2 to achieve a trade-off between the diversity and factuality of the answer set. During training, all the LLMs are trained using LoRA (Hu et al., 2022) with rank $r = 16$. Both the uncertainty estimation models and the reward model utilize the vanilla LLM as their bases and are trained using LoRA with rank $r = 4$. ADAM parameter update is used in a mini-batch mode. Uncertainty estimation models and the reward model are trained using SFT on the UALIGN dataset. The UALIGN PPO algorithm and all the RL-based baselines are implemented by `trl` [4]. All training hyper-parameters are presented in Appendix G. When decoding, the temperature is also set to 0.2 to be consistent with the sampling setting. All the experiments are conducted on 4 × NVIDIA A100-40GB GPUs.

[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B
[3] https://huggingface.co/mistralai/Mistral-7B-v0.1
[4] https://github.com/huggingface/trl

Method	TVQA (ID)		SciQ (ID)		NQ-Open (ID)		Avg. (ID)		LSQA (OOD)	
	Prec. ↑	*Truth.* ↑	*Prec.* ↑	*Truth.* ↑	*Prec.* ↑	*Truth.* ↑	*Prec.* ↑	*Truth.* ↑	*Prec.* ↑	*Truth.* ↑
Llama-3-8B										
ICL	76.15	56.55	70.43	44.30	50.28	20.11	65.62	40.32	77.35	52.98
ICL-IDK	69.17	54.10	68.36	43.00	45.43	20.72	60.98	39.27	66.67	50.24
ICL-CoT	66.68	53.37	72.34	45.90	**57.34**	23.60	65.45	40.95	73.96	49.37
SFT	70.80	52.57	72.18	45.40	41.41	16.57	61.46	38.18	68.09	46.63
R-Tuning	72.93	55.44	71.38	44.90	47.81	18.12	64.04	39.48	71.54	52.15
RL-PPO	76.32	55.19	75.70	45.80	54.07	24.19	68.03	41.72	72.18	48.43
RL-DPO	72.08	53.96	71.23	44.20	49.65	19.18	64.32	39.11	71.09	48.88
RLKF	77.12	56.07	72.36	44.90	54.86	22.15	68.11	41.04	74.95	52.46
ITI	71.09	53.97	72.35	43.80	43.20	17.13	62.21	38.30	68.52	46.99
UALIGN	**79.14**	57.04	**76.44**	**48.00**	56.60	26.09	**70.72**	43.71	**79.56**	**55.88**
(w/o **Conf.**)	74.13	54.45	74.05	45.00	54.19	23.60	67.45	41.01	74.25	52.06
(w/o **Entro.**)	78.43	**57.69**	75.39	47.50	56.68	**27.56**	70.16	**44.25**	76.14	54.43
Mistral-7B										
ICL	77.92	55.14	68.62	42.20	52.09	17.95	66.21	38.43	74.09	47.71
ICL-IDK	72.59	51.37	63.74	39.20	51.13	17.67	62.48	36.20	72.27	47.32
ICL-CoT	76.73	54.78	71.87	44.20	**54.47**	18.22	67.69	39.06	**79.24**	52.59
SFT	74.57	54.77	65.85	42.50	50.82	14.42	63.74	37.08	68.33	44.00
R-Tuning	67.70	52.25	64.44	40.10	46.33	15.52	59.49	36.29	64.67	44.05
RL-PPO	79.23	55.08	71.35	44.10	53.76	19.19	68.11	39.45	74.49	49.67
RL-DPO	72.20	52.98	66.44	41.80	50.95	16.42	63.19	37.06	67.82	43.77
RLKF	80.43	56.92	70.66	43.90	52.09	18.24	67.72	39.68	74.19	49.23
ITI	74.65	55.16	66.90	44.90	51.12	16.68	64.22	38.91	67.73	46.20
UALIGN	**82.10**	**59.05**	**73.21**	**46.70**	54.17	**19.64**	**70.82**	**41.79**	76.29	**52.89**
(w/o **Conf.**)	76.44	55.13	69.84	43.50	50.30	17.88	65.52	38.83	73.15	47.06
(w/o **Entro.**)	80.18	57.64	72.90	45.60	52.21	18.44	68.43	40.56	75.34	50.15

Table 1: Experiments of Precision (*Prec.*) and Truthfulness (*Truth.*) on four datasets on Llama-3 and Mistral.

4 Results and Analysis

4.1 Main Experimental Results

We present the results of UALIGN and several baselines on three ID and one OOD test sets as shown in Table 1. Several findings are listed below.

Reliability Significant improvements are consistently achieved on diverse datasets using the proposed UALIGN framework over other baseline methods on both Llama-3 and Mistral. We highlight the supreme Precision and Truthfulness performance using grey highlights among the all baselines of each column in Table 1. The core idea of our UALIGN framework is the utilization of uncertainty estimation models. Compared with the most relevant baselines of RL-PPO and RLKF, both the reward model and policy model in UALIGN generate predictions and responses conditioned on uncertainty estimations regarding the knowledge boundaries to questions, thereby yielding better reliability performance. It can be attributed that by explicitly appending uncertainty measures following the question, LLMs can assist LLMs in eliciting more accurate responses based on intrinsic knowledge boundary representations.

Generalization We also introduced an OOD test set to assess the generalization capability of the UALIGN method. The results in Table 1 indicate that most training-based baselines (SFT, RL, Inference) are unstable and result in performance decreasing compared with prompt-based baselines when generalizing on the OOD test set. However, comparable reliability performances are obtained on two LLMs using the proposed UALIGN in comparison with prompt-based methods, demonstrating strong generalization capability.

Conf.	Entro.	ID TVQA	ID SciQ	ID NQ-Open	OOD LSQA
Llama-3-8B					
✗	✗	82.31	79.00	67.45	70.12
✓	✗	85.41	84.30	70.37	**75.09**
✗	✓	82.05	77.90	67.85	70.40
✓	✓	**86.73**	**86.40**	**72.00**	74.59
Mistral-7B					
✗	✗	84.53	77.30	65.24	68.31
✓	✗	86.80	79.50	72.10	72.95
✗	✓	85.24	74.60	66.64	71.22
✓	✓	**88.06**	**79.80**	**75.14**	**73.61**

Table 2: Accuracy of reward model varying different uses of uncertainty measures Conf. and Entro. in UALIGN dataset on Llama-3 and Mistral.

4.2 Effects of Uncertainty Estimation Models

Setting To investigate the effects of introducing uncertainty estimations as input features to reward models, we report the accuracy of reward models that vary in different uses of two measures on ID

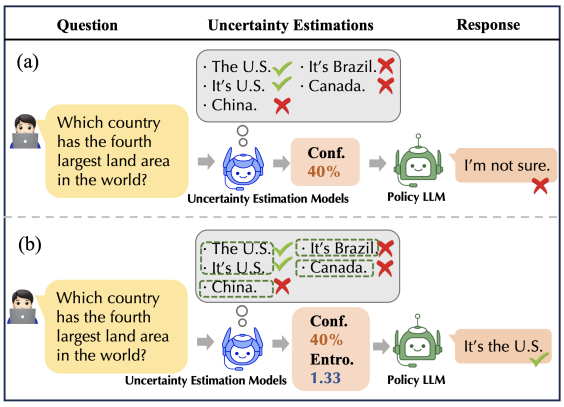


Figure 4: Illustration of the effects of different uses of uncertainty estimations under varying knowledge boundaries perceived by LLMs.

and OOD tasks. The reward models are trained on the UALIGN dataset on both Llama-3 and Mistral.

Results As in Table 2, we present the results of the accuracy of reward models. Significant accuracy improvements of reward models are obtained that predominantly benefit from the use of confidence scores across both ID and OOD test sets on two LLMs, validating the effectiveness of our proposed UALIGN framework. The isolated use of semantic entropy does not guarantee a stable improvement but may even lead to a performance decrease on some test sets. However, when semantic entropy is employed in combination with confidence measures, it can facilitate further enhancements, achieving optimal results across most test sets as highlighted grey cells for two LLMs.

Analysis In the UALIGN framework, both confidence score and semantic entropy are introduced to quantify the intrinsic knowledge boundary of LLMs to questions. The explicit introduction of the knowledge boundary representations in prompts can be regarded as the added thinking step like CoT. The combined use of confidence and semantic entropy can achieve supreme prediction performance in Table 2. We illustrate the mechanism as follows.

As demonstrated in Fig. 4 (a), by sampling multiple responses to a question, we can approximate LLM's intrinsic knowledge boundary, where the certainty level of the answer "*The U.S.*" is 40%. In previous work (Zhang et al., 2024a) which only considers the confidence level, the correct answer that the LLM knows but is not sure will be discarded and the LLM will refuse to answer. However, as in Fig. 4 (b), the LLM can perceive that even though its certainty level to the correct answer is low, other answers are more uncertain and the dispersion level of answers is relatively high which is quantified by semantic entropy. After UALIGN PPO training, the ability to generate correct answers conditioned on questions and estimations is well enhanced. As a result, the correct but unsure knowledge will be elicited in the responses.

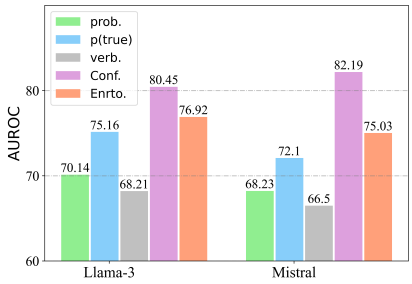


Figure 5: Results of AUORC↑ of several uncertainty estimation methods on TVQA using Llama-3 and Mistral.

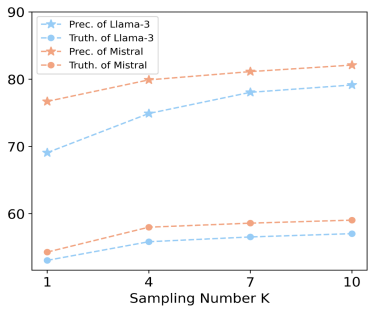
4.3 Reliability of Uncertainty Estimations

Setting Evaluating the performance of confidence score and semantic entropy is essential to the UALIGN method. We present the AUROC (Detailed in Appendix D) results of two estimations in comparison with three confidence/uncertainty estimation methods (one probability-based method (Prob.), two prompt-based methods including p(True) and verbalized (Verb.) as illustrated in Fig. 8) on TriviaQA on two LLMs. Results on other datasets are remained in Appendix I. Details of baseline estimation baselines are presented in Sec. 5, Appendix H, and Fig. 8.

Results In Fig. 5, both the confidence and entropy prediction consistently outperform other baseline uncertainty estimation methods. Optimal AUROC performances are obtained using confidence on both Llama-3 (80.45) and Mistral (82.19).

Analysis After UALIGN SFT stage, the uncertainty estimation models are converged on the UALIGN dataset to predict both confidence and entropy, indicating the models possess the ability to predict the two measures. Practically, our utilized confidence and semantic entropy incorporate the advantages of both sampling- and training-based uncertainty estimations. Multiple sampling can better approximate the actual knowledge boundaries of LLMs, while the training-based approach enables the LLMs to learn to perceive their intrinsic knowledge boundaries. Compared to other baselines that suffer from overconfidence issues with

low AUROC scores, our utilized methods yield more reliable estimates, thereby ensuring improved performance for both the reward model and the policy model in the following stages.



Sampling time cost on 10k queries on Llama-3

Sampling Number	Time Cost
1	18 min
4	22 min
7	24 min
10	25 min

Figure 6: Experiments of *Prec.*, *Truth.* (left), and time costs (right) of various sampling number K of 1, 4, 7, and 10 on TVQA on both Llama-3 and Mistral.

4.4 Effects of Sampling Number and Cost

Setting The sampling number K is a crucial hyper-parameter in the UALIGN method. Different values of K can significantly affect the precision of the knowledge boundary measurements. To evaluate the effects, we compare performances using various K of 1, 4, 7, and 10. Experimental results on TVQA are presented in Fig. 6 and Appendix I.

Findings Results in Fig. 6 indicate that when using small sampling numbers, increasing K leads to significant improvements in both *Prec.* and *Truth.*. However, as K increases, the reliability improvement tends to plateau, exhibiting convergence. Therefore, we opt $K = 10$ as the optimal setting and don't experiment using larger K.

We also report the sampling time costs to construct the training set in Fig. 6, and further specify cost analysis of UALIGN construction and inference in Appendix B.2 and B.3 respectively. We showcase that with various acceleration and quantization methods, time costs of UALIGN can be significantly reduced when scaling to larger models or datasets, exhibiting both efficiency and efficacy.

Analysis The results in Fig. 6 demonstrate that while the sampling number K increases linearly, the performance improvements are non-linear. This may be attributed to utilizing non-linear metrics, or it could suggest that $K = 10$ can approximate the actual knowledge boundaries, resulting in a gradual slowdown in performance gains. Consequently, setting K to 10 in this work makes a trade-off between performance gains and computation expense.

5 Related Works

Knowledge Boundary Previous works investigate the knowledge boundary (Yin et al., 2024) to identify the known level of a knowledge piece of LLMs by quantifying uncertainty estimations like output consistency (Cheng et al., 2024), prompting methods (Ren et al., 2023) or knowledge probing (Ji et al., 2024). Generally, knowledge boundary measures derive from uncertainty estimations.

Uncertainty Estimation for LLMs We categorize uncertainty estimation methods on LLMs into four classes as illustrated in Figure 8. ① *Likelihood-based methods* Vazhentsev et al. (2023) directly quantify sentence uncertainty over token probabilities; ② *Prompting-based methods* instruct LLMs to express uncertainty in words (Lin et al., 2022a; Xiong et al., 2024) or to self-evaluate its correctness on $p(\text{True})$ (Kadavath et al., 2022); ③ *Sampling-based methods* aggregate sampled responses to calculate consistency (Xiong et al., 2024) or semantic entropy (Kuhn et al., 2023); ④ *Training-based methods* (Lin et al., 2022a) propose to train LLMs to improve linguistic uncertainty expressions.

Factuality Alignment LLM alignment is to guide human preference through Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022a). Distinct from recent studies that apply RL to improve LLMs' factuality (Zhang et al., 2024b; Lin et al., 2024; Liang et al., 2024; Xu et al., 2024a), this work improves LLMs' reliability by explicitly leveraging the uncertainty estimations for LLM alignment.

Due to the space limitation, detailed investigations of related works are shown in Appendix H.

6 Conclusion

In this paper, we present a UALIGN framework to explicitly leverage uncertainty estimations to elicit LLMs to accurately express factual knowledge that LLMs cannot constantly answer correctly due to ambiguous knowledge boundaries. We introduce the dataset preparation process and UALIGN training strategies of factuality alignment by incorporating uncertainty estimations of the confidence score and semantic entropy as input features into prompts. Experiments on several knowledge QA tasks affirm the efficacy of UALIGN to enhance the LLMs' reliability and generalizability, demonstrating significant improvements over various baselines.

Limitations

The limitations and future work of this study are listed as follows:

Task Expansion: The dataset used in this paper is solely based on factual knowledge QA tasks, with a simple and fixed template and response format. However, the UALIGN methodology has not been further validated on other factual knowledge-based tasks such as open-form instruction-following tasks, long-form generation like biography, or even knowledge reasoning tasks, where the uncertainty estimations remain challenging. In future works, we plan to extend the UALIGN framework to open-ended generation tasks to enhance the LLMs' factual expressions.

Computational Cost: The current method for constructing the UALIGN dataset relies on multiple samplings, requiring additional computational cost that linearly increases with the number of sampling instances K and a higher number of samplings is preferable to accurately approximate the knowledge boundaries. As we have adopted a range of acceleration and quantization methods to reduce the time cost during both constructing the dataset and inference as presented in Appendix B.2 and B.3, there remains potential for exploration to further alleviate computational resources requirements.

Ethical Statement

In this paper, three evaluators are employed to annotate the correctness of four EM variants on selected samples, which aims to select the optimal EM variant to evaluate the correctness of the generated answer and the ground-truth label as presented in Appendix D.2. All the evaluators are M.Phil. or Ph.D. students possessing sufficient expertise to carry out the evaluation. We meticulously adhered to legal and ethical standards throughout the human evaluation process, prioritizing privacy and obtaining informed consent. Evaluators were furnished with comprehensive details regarding the study's objectives, data collection methodologies, and associated risks or benefits. They were afforded the opportunity to seek clarifications and voluntarily provide consent before their involvement. All the human evaluation results were exclusively utilized for research purposes.

Acknowledgments

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297.

AI@Meta. 2024. Llama 3 model card. *AI@Meta*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jiuhai Chen and Jonas Mueller. 2023. Quantifying uncertainty in answers from any language model via intrinsic and extrinsic confidence assessment. *arXiv preprint arXiv:2308.16175*.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can ai assistants know what they don't know? *Preprint*, arXiv:2401.13275.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion

Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim GJ Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. 2019. Benchmarking bayesian deep learning with diabetic retinopathy diagnosis. *Preprint at https://arxiv.org/abs/1912.10481*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Yarin Gal et al. 2016. Uncertainty in deep learning. *Ph.D. Thesis*.

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*, pages 7765–7784, Miami, Florida, USA. Association for Computational Linguistics.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2023. A survey of language model confidence estimation and calibration. *Preprint*, arXiv:2311.08298.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced self-training (rest) for language modeling. *Preprint*, arXiv:2308.08998.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. *Preprint*, arXiv:1706.04599.

Haixia Han, Tingyun Li, Shisong Chen, Jie Shi, Chengyu Du, Yanghua Xiao, Jiaqing Liang, and Xin Lin. 2024. Enhancing confidence expression in large language models through learning from past experience. *Preprint*, arXiv:2404.10315.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. LLM internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 88–104, Miami, Florida, US. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Matt Gardner Johannes Welbl, Nelson F. Liu. 2017. Crowdsourcing multiple choice science questions. In *arXiv*.

Mandar Joshi, Eunsol Choi, Daniel 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 (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur 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. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Siheng Li, Cheng Yang, Taiqiang Wu, Chufan Shi, Yuji Zhang, Xinyu Zhu, Zesen Cheng, Deng Cai, Mo Yu, Lemao Liu, Jie Zhou, Yujiu Yang, Ngai Wong, Xixin Wu, and Wai Lam. 2024. A survey on the honesty of large language models. *Preprint*, arXiv:2409.18786.

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

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for large language models. *Preprint*, arXiv:2405.01525.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

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

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Genglin Liu, Xingyao Wang, Lifan Yuan, Yangyi Chen, and Hao Peng. 2024. Examining llms' uncertainty expression towards questions outside parametric knowledge. *Preprint*, arXiv:2311.09731.

Jianqiao Lu, Zhiyang Dou, Hongru Wang, Zeyu Cao, Jianbo Dai, Yingjia Wan, and Zhijiang Guo. 2024. Autopsv: Automated process-supervised verifier. *Preprint*, arXiv:2405.16802.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2024. Calibrating large language models with sample consistency. *Preprint*, arXiv:2402.13904.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. Association for Computational Linguistics.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Preprint*, arXiv:2405.20003.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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. *Preprint*, arXiv:2305.18290.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *Preprint*, arXiv:2307.11019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in rlhf. *Preprint*, arXiv:2310.03716.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *Preprint*, arXiv:2408.03314.

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*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023a. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023b. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Artem Vazhentsev, Akim Tsvigun, Roman Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko, and Artem Shelmanov. 2023. Efficient out-of-domain detection for sequence to sequence models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1430–1454, Toronto, Canada. Association for Computational Linguistics.

Hao Wang and Dit-Yan Yeung. 2020. A survey on bayesian deep learning. *ACM Comput. Surv.*, 53(5).

Hongru Wang, Boyang Xue, Baohang Zhou, Tianhua Zhang, Cunxiang Wang, Guanhua Chen, Huimin Wang, and Kam fai Wong. 2024. Self-dc: When to retrieve and when to generate? self divide-and-conquer for compositional unknown questions. *Preprint*, arXiv:2402.13514.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7273–7284, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024a. Rejection improves reliability: Training LLMs to refuse unknown questions using RL from knowledge feedback. In *First Conference on Language Modeling*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024b. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Boyang Xue, Shoukang Hu, Junhao Xu, Mengzhe Geng, Xunying Liu, and Helen Meng. 2022. Bayesian neural network language modeling for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2900–2917.

Boyang Xue, Hongru Wang, Rui Wang, Sheng Wang, Zezhong Wang, Yiming Du, Bin Liang, and Kam-Fai Wong. 2024. A comprehensive study of multilingual confidence estimation on large language models. *Preprint*, arXiv:2402.13606.

Boyang Xue, Weichao Wang, Hongru Wang, Fei Mi, Rui Wang, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2023. Improving factual consistency for knowledge-grounded dialogue systems via knowledge enhancement and alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7829–7844, Singapore. Association for Computational Linguistics.

Yuchen Yang, Houqiang Li, Yanfeng Wang, and Yu Wang. 2023. Improving the reliability of large language models by leveraging uncertainty-aware in-context learning. *Preprint*, arXiv:2310.04782.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024. Alignment for honesty. *Preprint*, arXiv:2312.07000.

Xunjian Yin, Xu Zhang, Jie Ruan, and Xiaojun Wan. 2024. Benchmarking knowledge boundary for large language model: A different perspective on model evaluation. *arXiv preprint arXiv:2402.11493*.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'i don't know'. *Preprint*, arXiv:2311.09677.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024b. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. *Preprint*, arXiv:2402.09267.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

Notation	Description
$\mathcal{Q}$	Dataset containing n Question-Answering pairs. ($\lvert\mathcal{Q}\rvert = n$)
$\mathcal{P}$	Set of few-shot exemplars.
$\boldsymbol{x}_i$	The i-th question sample in $\mathcal{Q}$.
$\hat{\boldsymbol{y}}_i$	The i-th ground-truth answer in $\mathcal{Q}$.
${\boldsymbol{y}_i}^{(k)}$	The k-th sampled response to the i-th question in $\mathcal{Q}$.
$\boldsymbol{p}_k$	k-th few-shot exemplar to sample ${\boldsymbol{y}_i}^{(k)}$.
$\boldsymbol{Y}_i$	Answering set containing K sampled response $\left\{{\boldsymbol{y}_i}^{(k)}\right\}$ for the i-th question $\boldsymbol{x}_i$.
${z_i}^{(k)}$	The label of ${\boldsymbol{y}_i}^{(k)}$ (${z_i}^{(k)} \in \{0, 1\}$, 1 for *True* and 0 for *False*).
$\boldsymbol{Z}_i$	Label set corresponding to $\boldsymbol{Y}_i$.
c_i	The confidence score for the i-th question $\boldsymbol{x}_i$.
e_i	The semantic entropy for the i-th question $\boldsymbol{x}_i$.
$\mathcal{D}$	Constructed UALIGN training set containing N tuple samples $(\boldsymbol{x}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i, \hat{\boldsymbol{y}}_i, c_i, e_i)$.
τ	Uncertainty estimation model trained to calculate confidence score by feeding $\boldsymbol{x}$.
μ	Uncertainty estimation model trained to calculate semantic entropy by feeding $\boldsymbol{x}$.
θ	Binary classifier by feeding $(\boldsymbol{x}, c, e, \boldsymbol{y})$ as the reward model.
$\mathcal{L}_{\mathcal{M}}$	Training loss functions for three models respectively where $\mathcal{M} \in \{\tau, \mu, \theta\}$.
r	Final reward signal consisted of reward score r_1 and KL-penalty r_2.
β	Coefficient for the KL-penalty r_2.
π_θ	Policy model to be optimized using r by PPO.
π_o	Reference model initialized by the original policy.
T	Sampling temperatue.
K	Number of sampled responses.
N	Number of QA pairs.

Table 3: Summarized notations in this work.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Protocols

A.1 Definition of Notations

The definitions of the notations in this work are summarized in Table 3.

A.2 Terminology Use

- In this work, "UALIGN" in small caps font specifically indicates the proposed framework, which indicates methodology like UALIGN dataset, UALIGN SFT and UALIGN PPO.

B Method Specification and Supplement

B.1 Model Update during PPO

During the PPO process, only the policy model π_θ is optimized while the uncertainty models μ, τ do not need to be updated because the reward model θ updates are offline. As discussed and demonstrated in Sec. 4.2 and Table 2, uncertainty models are directly associated with and benefit the reward model. In our UAlign PPO algorithm, by incorporating the two uncertainty estimations, the reward model θ can provide more precise reward scores, thereby guiding LLMs π to generate more factual responses. Since the reward model is offline during PPO, the uncertainty models also do not require online updates.

In addition, due to the KL-divergence constraint, the knowledge distribution of policy LLMs may not diverge too much from the initial policy. Both uncertainty models and reward models are trained on data generated by sampling from the vanilla LLMs, and their combined effect is to elicit the LLMs' capacity for factual expression, evolving towards improved reliability. During the PPO process, with the KL-divergence constraint in Equation 7, the knowledge distribution of policy LLMs may not shift too much from the initial policy. We demonstrate the accuracy-based confidence distribution of Llama-3 before and after UALIGN training on TriviaQA validation sets as follows.

Conf. Range	Before UAlign	After UAlign
$[0, 0.25)$	2404	2116
$[0.25, 0.5)$	1786	1628
$[0.5, 0.75)$	1509	1747
$[0.75, 1.0]$	4261	4469

Table 4: Accuracy-based confidence distribution of Llama-3 before and after UALIGN training on TriviaQA validation sets.

Since the knowledge distribution does not shift too much from the initial policy, we can still achieve good performance without updating the uncertainty model for simplicity. Compared to traditional RLHF that solely utilize reward models, our

proposed UALIGN introduces uncertainty models that leverage knowledge boundary representations to benefit reward model and finally enhance LLMs, leading to signification improvements in reliability and generalization of knowledge QA tasks.

B.2 Computation Cost of Constructing UAlign Dataset

As mentioned in Sec. 4.4, we test the time costs to construct the UAlign dataset $\mathcal{D}$ in different sampling numbers. We present different sampling time costs on 10000 QA samples on Llama-3-8B and Llama-3-70B (AI@Meta, 2024) [5] on 4×40G A100 GPUs loaded in fp16. We have tried to address the computation cost problem by introducing many effective acceleration or quantization packages like vllm [6], bitsandbytes [7], etc that are widely used to drive test time scaling law (Snell et al., 2024). As presented in Table 5, the results demonstrate the efficiency of our proposed UALIGN method even though scaling on larger models.

Model	Sampling Number	Time Cost
Llama-3-8B	1	18 min
	4	22 min
	7	24min
	10	25min
Llama-3-70B	1	1h 18min
	4	1h 33min
	7	1h 40min
	10	2h 12min

Table 5: Time cost in different sampling numbers of UALIGN on Llama-3-8B and Llama-3-70B.

In addition, the relatively low computation costs when sampling can be attributed that experiments are conducted on knowledge QA datasets. The answer spans are entity-level and each answer only needs to generate a few tokens. Since the output form is relatively simple, sampling ten times is sufficient and cost-saving to accurately fit the knowledge boundaries as presented in Sec. 4.4.

Furthermore, Test Time Scaling Law (Snell et al., 2024) has attracted much attention recently which proposed to consider allocating more computation resources in inference to generate high-quality responses. These LLMs' self-generated data can be further used for LLM training to self-improve LLMs (Gulcehre et al., 2023). Many works validate that incorporating data multiply sampled on LLMs in inference can benefit LLMs for further improvements, which is a new trend for LLM training. In this way, our proposed UALIGN provides a novel insight of the test time scaling law to represent knowledge boundaries by calculating uncertainty estimations on the sampled responses, and further explicitly leverages the uncertainty estimations for factuality alignment, heralding a promising view of test time scaling law. Although few additional computation costs are required, our proposed UALIGN is still efficient to be utilized practically with significant reliability improvements.

B.3 Computation Cost of Inference of UAlign

Following B.2, we subsequently analyze the computation cost during inference of UALIGN. Our proposed UALIGN barely increases additional inference memory and time budget as follows.

First, uncertainty models also share the base LLMs with their respective plug-in LoRA modules with rank r=4. Additional parameters introduced only account for less than 1% of the base model parameters.

Second, uncertainty models only predict two tokens of uncertainty estimations in inference. We report the inference time cost on four test sets of vanilla Llama-3 which only generates the answer to the question and UALIGN trained Llama-3 which predicts uncertainty estimations and then generates the answers on a single A100 GPU Card.

Dataset	Time Cost	
	Vanilla ICL	UAlign
TVQA	58 min	1h 6min
NQ-Open	28 min	32m in
SciQ	6 min	7 min
LSQA	5 min	6 min

Table 6: Inference time cost on four test sets of Llama-3 using vanilla ICL prompt-based and UALIGN methods. Note the inference time costs on all the baseline methods in Sec. 3.3 are comparable to the vanilla ICL prompt-based baseline method.

Therefore, with the slight increase in additional memory and time cost in inference, UALIGN significantly outperforms other baseline methods, demonstrating the reliability and efficiency on such tasks.

[5] https://huggingface.co/meta-llama/Meta-Llama-3-70B
[6] https://github.com/vllm-project/vllm
[7] https://github.com/bitsandbytes-foundation/bitsandbytes

C Dataset Details

TriviaQA The TriviaQA dataset (Joshi et al., 2017) [8] is a comprehensive reading comprehension dataset of QA resource consisting of approximately 650,000 question-answer-evidence triples sourced from 95,000 documents on Wikipedia and various other websites. This dataset is distinguished by its complexity and serves as an effective benchmark for evaluating machine comprehension and open-domain QA systems. Unlike standard QA benchmark datasets, where answers are directly retrievable, TriviaQA presents a more rigorous challenge as it requires deeper inference to derive answers.

When constructing the UALIGN dataset, we preprocess and extract 76,523 QA samples from the TriviaQA training set and 9,960 from the development set to contribute to the UALIGN training and in-domain test set respectively. Since approximating the knowledge distribution of a question requires multiple sampling where the computation cost is linearly increasing with the sampling time K, to simplify the setup and conserve computation resources, we conducted experiments using half of the training data points from the original dataset.

SciQ The SciQ dataset (Johannes Welbl, 2017) [9] contains 13,679 crowd-sourced science exam questions about physics, chemistry and biology, among others. The original dataset was divided, with 11,679 samples allocated as the training set and an additional 1,000 samples designated as the validation set. These were subsequently incorporated into our UALIGN training set and in-domain test set, respectively.

NQ-Open The NQ-Open dataset is derived from Natural Question (Kwiatkowski et al., 2019) [10], which is a QA dataset consisting of real queries issued to the Google search engine. We employ the training and development set of NQ-Open, which contains 87,925 and 3,610 samples respectively, to further enhance the UALIGN training and in-doamin test set. Since data construction is highly expensive, we also randomly sample half of the QA pairs from the source training data. We mix the selected training samples to construct the UALIGN dataset, which is further used for U2Align SFT+PPO training.

[8] https://huggingface.co/datasets/mandarjoshi/trivia_qa
[9] https://huggingface.co/datasets/allenai/sciq
[10] https://huggingface.co/datasets/google-research-datasets/nq_open

LSQA The LSQA dataset is a multilingual knowledge-intensive QA dataset pertaining to language-dominant knowledge covering specific social, geographical, and cultural language contexts for the UK & US, France, China, Japan, and Thailand respectively. In this study, we only input the QA pairs in English from each LSQA subset which includes 1,025 samples as the out-of-domain test set.

D Evaluation Details

D.1 Precision and Truthfulness

Notation	Indication
KC	Known and answered correctly
KI	Known but answered incorrectly
KR	Known but refused to answer
UC	Unknown but answered correctly
UI	Unknown but answered correctly
UR	Unknown and refused to answer

Table 7: Denotation of different answer types.

Explanations and Equations As defined in Table 7, "*Truthfulness*" is the proportion of questions the LLM either the known answered correctly or the unknown refused to answer, which measures the honesty of LLMs. Some unknown but correctly guessed answers will not be included. The equation of *Truthfulness* is as follows.

$$\text{Truthfulness} = \frac{\text{UR} + \text{KC}}{\text{KC} + \text{KI} + \text{KR} + \text{UC} + \text{UI} + \text{UR}} \quad (8)$$

Precision is defined as the proportion of correctly answered questions among all the known questions, representing LLMs' ability to accurately express their known factual knowledge. The equation of *Precision* is as follows.

$$\text{Precision} = \frac{\text{KC}}{\text{KI} + \text{KC} + \text{KR}} \quad (9)$$

Clarifications of Use of Truthfulness To avoid the over-conservative problem incurred by using precision only, we employ "truthfulness" as complementary to measure the proportion of questions the model either known answered correctly or unknown refused to answer, which reflects the honesty of the model. Therefore, as demonstrated in

Sec. 4 and Table 1, the previous methods like R-Tuning which may lead models to be overly conservative perform well in precision but poor in truthfulness. The employed two metrics of precision and truthfulness can comprehensively measure the reliability of different methods, thereby comprehensively demonstrating the superiority of our method over other baselines from these two perspectives.

D.2 Accuracy

For closed-book QA evaluation, we observe that simply applying EM may misjudge the correct answers. We compare several variants of EM as in Table 8 and report their successful judgments on responses of 20 selected samples that are misjudged using EM, where PEM, RRM, and PREM indicate Positive-EM, Recall-EM, and Positive-Recall-EM and the mathematical explanations are presented in Table 8. Upon human discrimination, EMPR exhibits the lowest failure rate and is therefore selected as the evaluation metric for this work.

Variant	Explanation	# Fail
EM	$\boldsymbol{y} \equiv \hat{\boldsymbol{y}}$	20
PEM	$\boldsymbol{y} \in \hat{\boldsymbol{y}}$	16
REM	$\hat{\boldsymbol{y}} \in \boldsymbol{y}.$	6
PREM	$\boldsymbol{y} \in \hat{\boldsymbol{y}} \vee \hat{\boldsymbol{y}} \in \boldsymbol{y}.$	2

Table 8: Number of failed judgments by human check for different EM variants.

D.3 AUROC

Area Under the Receiver Operator Characteristic Curve (AUROC) assesses the effectiveness of confidence estimation (Filos et al., 2019) by quantifying how likely a randomly chosen correct answer possesses a higher confidence score than an incorrect one, yielding a score within the range of [0, 1], implemented by `sklearn` toolkit [11]. A higher AUROC score implying higher reliability is preferred.

E Baseline Details

Prompt-based For all in-context learning methods, we extract the examples from the respective training set to mitigate the knowledge distribution shift between different datasets. For example, the demonstrated examples in Appendix J are derived from the TriviaQA training set and are specifically used when inferring on the TriviaQA validation set. For LSQA without the training set, we use the same examples as TriviaQA as their knowledge domains largely overlap.

- **ICL**: Few-shot prompts containing m examples are utilized for answer generation with temperature $T = 0.2$ where m is set to 2 as presented in the Template F.
- **ICL-IDK**: Two examples are included in the few-shot prompt while one is selected from the ICL-used example, and another is an unknown question whose answer is revised to "*Sorry, I don't know.*" as presented in the Template F.
- **ICL-CoT**: We also employ the Chain-of-Thought in few-shot examples by recalling the relevant knowledge piece of LLMs and incorporating it into thinking steps before answering the question as presented in the Template F.
- **SFT**: The standard supervised fine-tuning (SFT) is implemented by minimizing the negative log-likelihood of the ground-truth $\hat{\boldsymbol{y}}$ conditioned on input question $\boldsymbol{x}$ on model π.

$$\arg\min_{\pi} \mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(\boldsymbol{x}_i, \hat{\boldsymbol{y}}_i)\sim\mathcal{D}} \left[\log p_{\pi}(\hat{\boldsymbol{y}}|\boldsymbol{x})\right] \tag{10}$$

- **R-Tuning**: R-Tuning (Zhang et al., 2024a) is implemented in the same way as SFT which only revises the ground-truth label of unknown questions to the refusal answers. The unknown questions are determined if all the sampled responses in the UALIGN dataset are incorrect.
- **RL-PPO**: Following (Ouyang et al., 2022), we develop the RL-PPO by training a reward model using the LLM-generated incorrect responses as negative samples. Then we conduct the PPO (Schulman et al., 2017) algorithm with the obtained reward model. In other word, the RL-PPO baseline is a variant of UALIGN which discards the uncertainty estimations.
- **RLKF**: Following (Liang et al., 2024), we employ the RLKF baseline by training the reward model on the LLMs' internal states with the knowledge probes and further conduct PPO

[11] https://github.com/scikit-learn/scikit-learn/blob/main/sklearn/metrics/_ranking.py

using the reward model. The knowledge probing setting and implementations are referred to as Liang et al. (2024).

- **RL-DPO**: All Tian et al. (2024); Lin et al. (2024); Zhang et al. (2024b) focus on long-context generation like biography. We still utilize the LLMs' generated incorrect responses as negative samples to construct the preference data to conduct the DPO (Rafailov et al., 2023) algorithm.
- **ITI**: We replicate (Li et al., 2023) by training a head probe in the attention layer to intervene in the activations to the "truthfulness" direction. To be consistent with the original work, we also train the head on TruthfulQA (Lin et al., 2022b) with our prepared UALIGN dataset to decode in the "truthfulness" direction. Then we further train the LLM using LoRA by SFT to adapt QA tasks. Therefore, the replicated ITI can be regarded as conducting SFT on LLMs with an additional "truthfulness" head.

F Prompt Template

ICL Prompt

You are an excellent Question-Answering assistant. Please answer the following question based on your knowledge.

Question ###: `{demo_question_1}`
Answer ###: `{demo_answer_1}`

Question ###: `{demo_question_2}`
Answer ###: `{demo_answer_2}`

Question ###: `{input_question}`
Answer ###:

ICL-IDK Prompt

You are an excellent Question-Answering assistant. Please answer the following question based on your knowledge.

Question ###: `{demo_question_1}`
Answer ###: `{demo_answer_1}`

Question ###: `{demo_question_2}`
Answer ###: `{refusal}`

Question ###: `{input_question}`
Answer ###:

ICL-CoT Prompt

You are an excellent Question-Answering assistant. Please answer the following question based on your knowledge.

Question ###: `{demo_question_1}`
Recall ###: `{knowledge_1}`
Answer ###: `{demo_answer_1}`

Question ###: `{demo_question_2}`
Recall ###: `{knowledge_2}`
Answer ###: `{demo_answer_2}`

Question ###: `{input_question}`
Answer ###:

G Training Setting Details

To conserve memory overhead and accelerate computation, all the models are quantified using `float16 (fp16)` to load and save parameters during both the training and inference phases. During the training stage, the batch sizes for the LLM, uncertainty estimation models, and reward models are set at 4, 16, and 16, respectively. The initial learning rate of 1e-4 is utilized with the 0.05 warm-up ratio and 0.01 weight decay of the ADAM optimizer. We set the training epoch to 2 and ensure that all the models can be trained to convergence by increasing additional training steps if necessary. The dropout rate is set at 0.05 during all model updates to reduce overfitting. In the RL phase, all the hyper-parameters related to PPO algorithm are default values by the `trl` PPOConfig recipe [12] except the epoch, learning rate, and batch size which are set at 2, 1e-5, and 2, respectively.

H Detailed Related Works

H.1 Knowledge Boundary

Previous works investigate the knowledge boundary to identify the known level of a knowledge piece of LLMs by quantifying the confidence or uncertainty estimations like output consistency (Cheng et al., 2024), prompting methods (Ren et al., 2023) or knowledge probing (Ji et al., 2024). Researchers are examining the limits of parametric knowledge in LLMs with the objective of delineating the extent of the LLMs' knowledge and identifying their capability boundaries. Present studies on the knowledge boundary primarily focus on

[12] https://github.com/huggingface/trl/blob/main/trl/trainer/ppo_config.py

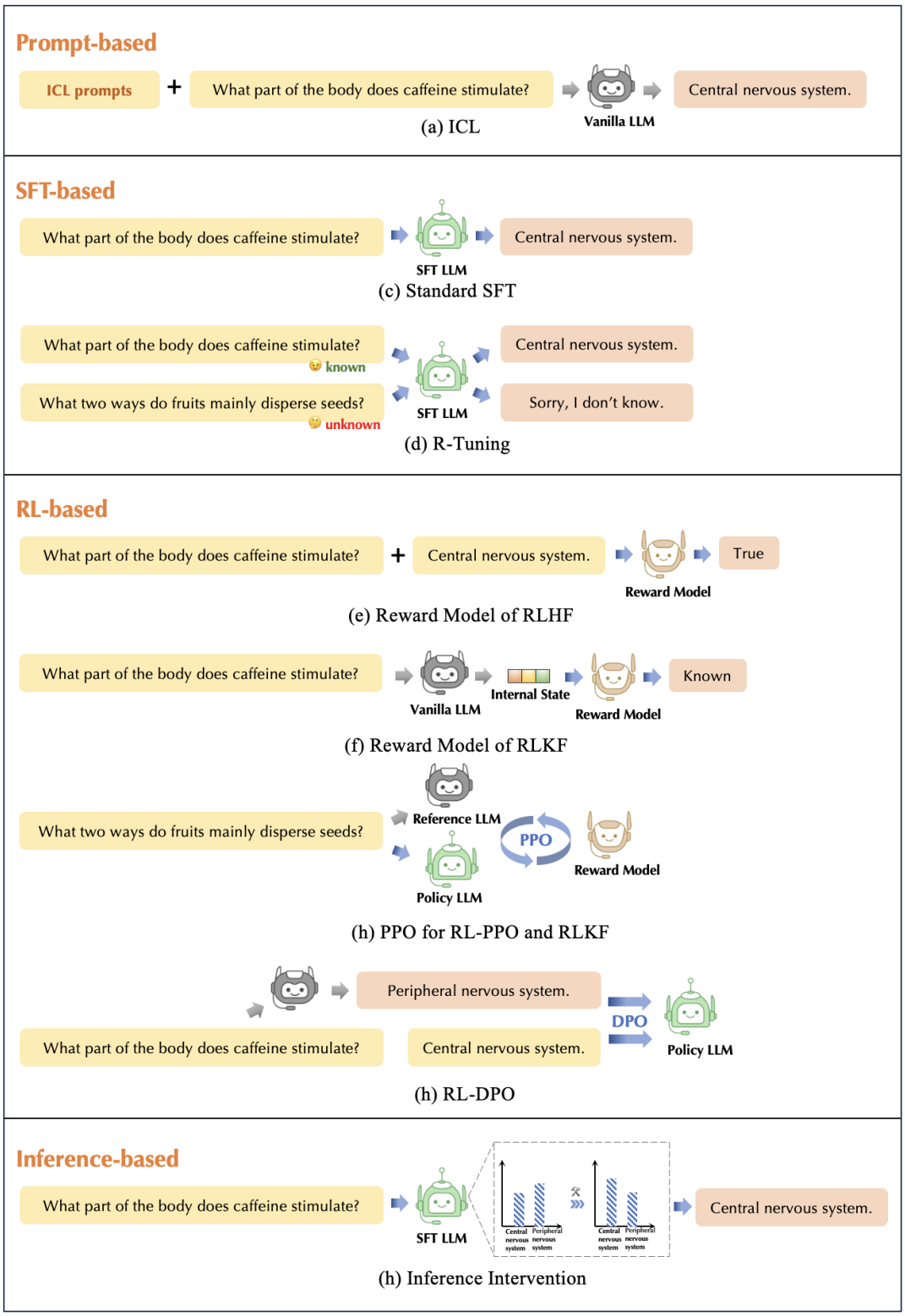


Figure 7: Illustration of several baselines as in Sec. 3.3.

measuring the knowledge boundaries using confidence or uncertainty estimations on specialized tasks. The ambiguity of knowledge boundaries can be attributed to the knowledge distribution learned from the pre-training stage or the influence of external knowledge leading to knowledge conflict (Xu et al., 2024b) and inconsistency (Xue et al., 2023).

H.2 Uncertainty Estimation of LLMs

To alleviate over-confidence and enhance the reliability of LLMs, reliable uncertainty estimation is essential to determine whether a question is known

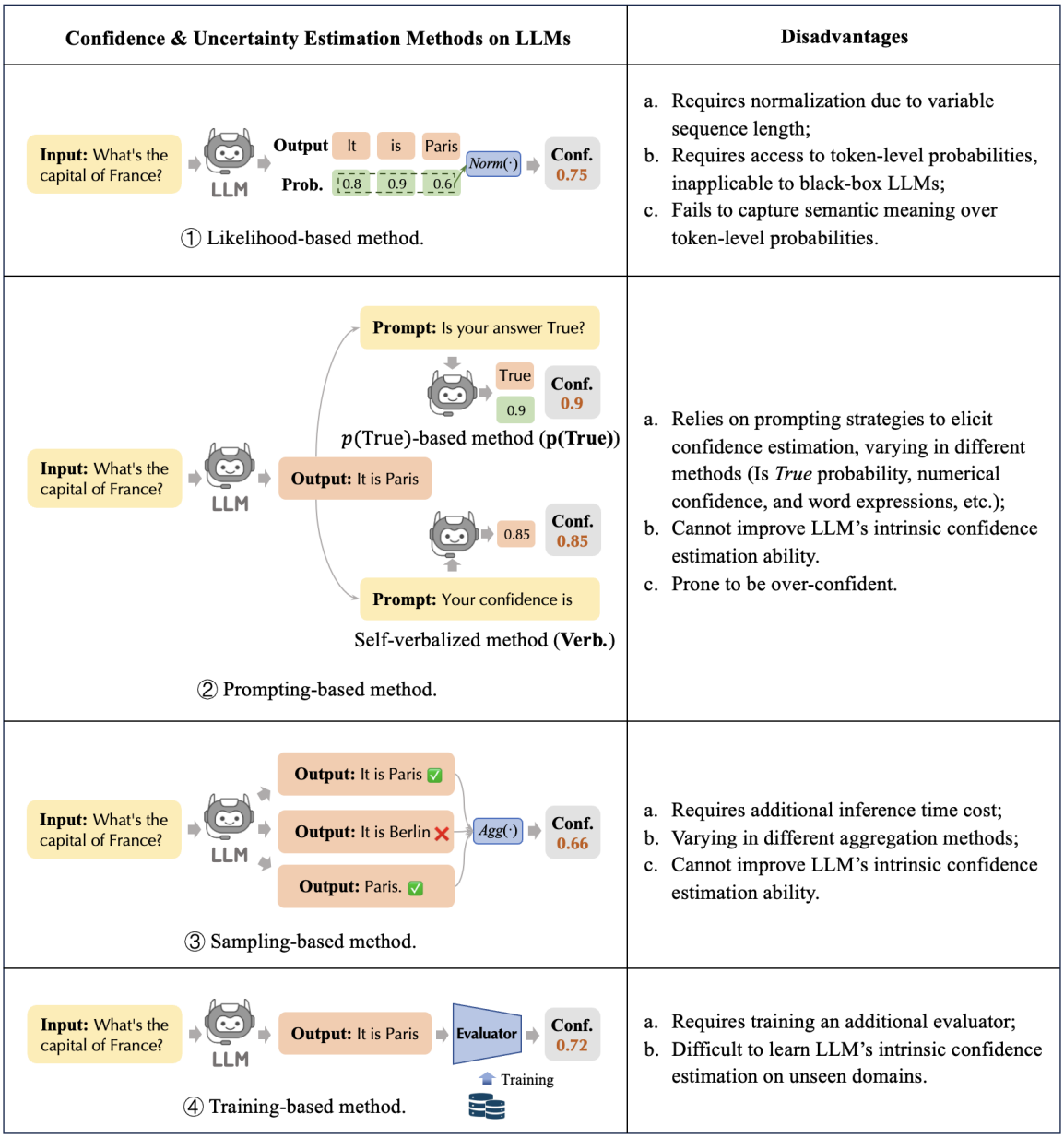


Figure 8: Several uncertainty estimation methods for Generative LLMs.

or not to the LLM (Geng et al., 2023). Both *Uncertainty* and *Confidence* estimations can indicate the reliability degree of the responses generated by LLMs, and are generally used interchangeably (Xiao et al., 2022; Chen and Mueller, 2023; Geng et al., 2023; Lu et al., 2024). In this part, we investigate several commonly used *confidence* & *uncertainty* estimation methods for generative LLMs as mentioned in Sec. 5. Specifically, we denote $\text{Conf}(\boldsymbol{x}, \boldsymbol{y})$ as the confidence score associated with the output sequence $\boldsymbol{y} = [y_1, y_2, \ldots, y_N]$ given the input context $\boldsymbol{x} = [x_1, x_2, \ldots, x_M]$. We also illustrate the summarized estimation methods as well as their disadvantages in Fig. 8.

Likelihood-based Methods: Following model calibration on classification tasks (Guo et al., 2017), Vazhentsev et al. (2023); Xue et al. (2024); Varshney et al. (2023); Wang et al. (2024) intermediately quantify sentence uncertainty over token probabilities. In traditional discriminative models, except likelihood-based methods, confidence estimations also include ensemble-based and Bayesian methods (Lakshminarayanan et al., 2017; Gal and Ghahramani, 2016; Xue et al., 2022; Wang and Yeung, 2020; Gal et al., 2016; Abdar et al., 2021), and density-based methods (Lee et al., 2018). However, this likelihood-based method requires access to token probabilities and thus being limited to

white-box LLMs. The likelihood-based confidence is estimated by calculating the joint token-level probabilities over $\boldsymbol{y}$ conditioned on $\boldsymbol{x}$. As longer sequences are supposed to have lower joint likelihood probabilities that shrink exponentially with length, the product of conditional token probabilities of the output should be normalized by calculating the geometric mean by the sequence length (Murray and Chiang, 2018; Malinin and Gales, 2021), and the confidence score can be represented as:

$$\mathsf{Conf}(\boldsymbol{x}, \boldsymbol{y}) = \left(\prod_{i}^{N} p(y_i | \boldsymbol{y}_{<i}, \boldsymbol{x}) \right)^{\frac{1}{N}} \quad (11)$$

Similarly, the arithmetical average of the token probabilities is adopted in Varshney et al. (2023):

$$\mathsf{Conf}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{N} \sum_{i}^{N} p(y_i | \boldsymbol{y}_{<i}, \boldsymbol{x}) \quad (12)$$

Furthermore, a low probability associated with even one generated token may provide more informative evidence of uncertainty (Varshney et al., 2023). Hence, the minimum of token probabilities is also employed.

$$\mathsf{Conf}(\boldsymbol{x}, \boldsymbol{y}) = \min \{ p(y_1 | \boldsymbol{x}), \ldots, p(y_N | \boldsymbol{y}_{<N}, \boldsymbol{x}) \} \quad (13)$$

Prompting-based Methods: Recently, LLMs' remarkable instruction-following ability (Brown et al., 2020) provides a view of instructing LLMs to self-estimate their confidence level to previous inputs and outputs including expressing uncertainty in words (Lin et al., 2022a; Zhou et al., 2023; Tian et al., 2023a; Xiong et al., 2024), or instructing the LLM to self-evaluate its correctness on $p(\text{True})$ (Kadavath et al., 2022). The $P(\text{True})$ confidence score is implemented by simply asking the model itself if its first proposed answer $\boldsymbol{y}$ to the question $\boldsymbol{x}$ is true (Kadavath et al., 2022), and then obtaining the probability $p(\text{True})$ assigned by the model, which can implicitly reflect self-reflected certainty as follows.

$$\mathsf{Conf}(\boldsymbol{x}, \boldsymbol{y}) = p(\text{True}) = p(\boldsymbol{y} \text{ is True?} | \boldsymbol{x}) \quad (14)$$

Another method is to prompt LLMs to linguistically express tokens of confidence scores in verbalized numbers or words (Lin et al., 2022a; Mielke et al., 2022; Zhou et al., 2023; Tian et al., 2023b; Xiong et al., 2024).

The sampling-based method refers to randomly sampling multiple responses given a fixed input $\boldsymbol{x}$ using beam search or temperature sampling strategies (Manakul et al., 2023; Xiong et al., 2024; Lyu et al., 2024). Various aggregation methods are adopted on sampled responses to calculate the consistency level as the confidence score. Moreover, some uncertainty quantification methods are used to calculate the entropy indicating the dispersion level of multiple outputs (Kuhn et al., 2023; Lin et al., 2023; Nikitin et al., 2024).

Training-based Methods: For training methods, an external evaluator trained on specific datasets is introduced to output a confidence score given an input and an output. The evaluator can be a pre-trained NLI model (Mielke et al., 2022), or a value head connected to the LLM output layer (Lin et al., 2022a; Kadavath et al., 2022), or the LLM itself (Han et al., 2024).

However, both self-verbalized and sampling methods for uncertainty estimations using extrinsic prompting or aggregation strategies with additional time costs fail to improve LLMs' intrinsic capability of uncertainty estimation. Recent works investigate confidence learning methods to enhance the reliability of LLMs (Han et al., 2024). Li et al. (2023) introduces Inference-Time Intervention (ITI) to enhance the truthfulness of LLMs by shifting model activations during inference. Yang et al. (2023) proposes an uncertainty-aware in-context learning method leveraging uncertainty information to refine the responses but cannot improve uncertainty estimation. (Zhang et al., 2024a) proposes R-tuning to instruct LLMs to refuse unknown questions considering uncertainty estimations as binary indicators. In contrast, our proposed UALIGN framework not only obtains more reliable uncertainty estimations regarding knowledge boundary information but also elicits accurate responses of LLMs.

H.3 Factuality Alignment of LLMs

Alignment is a standard procedure to improve LLMs' helpfulness and factuality (Bai et al., 2022a). The main goal of LLM alignment is to guide human preference through Supervised Fine-Tuning (SFT), Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022a) or AI feedback (Bai et al., 2022b), which may also guide LLMs to output detailed

and lengthy responses (Singhal et al., 2023) but inevitably encourage hallucination. Therefore, many works explore to apply RL to improve LLMs' factuality through Proximal Policy Optimization (PPO) (Schulman et al., 2017) with the trained reward model (Liang et al., 2024; Xu et al., 2024a) or Direct Preference Optimization (DPO) Rafailov et al. (2023) with the constructed preference dataset (Zhang et al., 2024b; Lin et al., 2024) to align with factuality preferences annotated by human beings. Xu et al. (2024a) encourage LLM to reject unknown questions using the constructed preference data by leveraging knowledge boundary feedback.

I Experiments

I.1 Experiments of Reliability of Uncertainty Estimations

Due to the page limitation in the main part, we present the AUROC performance results of the used confidence and entropy compared with other baseline uncertainty estimations on SciQ, NQ-Open, and LSQA as in Fig. 9, 10, and 11.

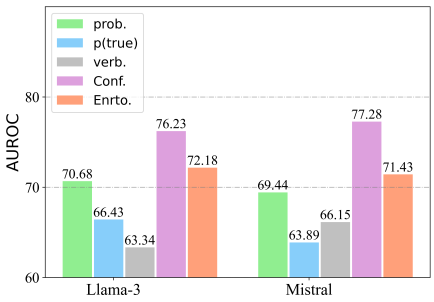


Figure 9: Results of AUORC↑ across several confidence/uncertainty estimation methods on SciQ on Llama-3 and Mistral.

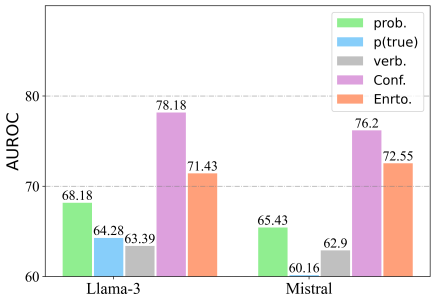


Figure 10: Results of AUORC↑ across several confidence/uncertainty estimation methods on NQ-Open on Llama-3 and Mistral.

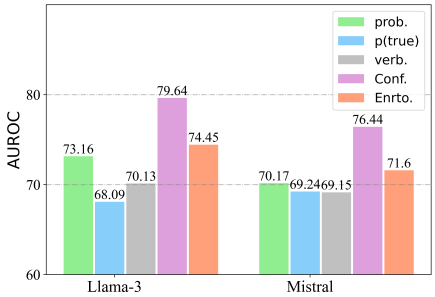


Figure 11: Results of AUORC↑ across several confidence/uncertainty estimation methods on LSQA on Llama-3 and Mistral.

J Few-shot Prompt Examples

10 different few-shot prompts for sampling on TriviaQA are demonstrated in Table 9.

Examplar ID	Examples
1	Q: Which William wrote the novel Lord Of The Flies? A: Golding.
2	Q: Where in England was Dame Judi Dench born? A: York, UK.
3	Q: Neil Armstrong was a pilot in which war? A: Korean.
4	Q: How many home runs did baseball great Ty Cobb hit in the three world series in which he played? A: None.
5	Q: Who had a big 60s No 1 with Tossin' and Turnin'? A: Bobby Lewis.
6	Q: Which Disney film had the theme tune A Whole New World? A: 'Ala' ad Din.
7	Q: In basketball where do the Celtics come from? A: City of Boston.
8	Q: Which element along with polonium did the Curies discover? A: Radium.
9	Q: Who was the Egyptian king whose tomb an treasures were discovered in the Valley of the Kings in 1922? A: Tutanhamon.
10	Q: Where were the 2004 Summer Olympic Games held? A: Atina, Greece.

Table 9: Demonstrations of 1-shot examples for TriviaQA sampling to construct UALIGN dataset.

Examplar ID	Examples
1	Q: What type of organism is commonly used in preparation of foods such as cheese and yogurt? A: mesophilic organisms.
2	Q: What phenomenon makes global winds blow northeast to southwest or the reverse in the northern hemisphere and northwest to southeast or the reverse in the southern hemisphere? A: coriolis effect.
3	Q: Changes from a less-ordered state to a more-ordered state (such as a liquid to a solid) are always what? A: exothermic.
4	Q: What is the least dangerous radioactive decay? A: alpha decay.
5	Q: Kilauea in hawaii is the world's most continuously active volcano. very active volcanoes characteristically eject red-hot rocks and lava rather than this? A: smoke and ash.
6	Q: When a meteoroid reaches earth, what is the remaining object called? A: meteorite.
7	Q: What kind of a reaction occurs when a substance reacts quickly with oxygen? A: combustion reaction.
8	Q: Organisms categorized by what species descriptor demonstrate a version of allopatric speciation and have limited regions of overlap with one another, but where they overlap they interbreed successfully? A: ring species.
9	Q: Alpha emission is a type of what? A: radioactivity.
10	Q: What is the stored food in a seed called? A: endosperm.

Table 10: Demonstrations of 1-shot examples for SciQ sampling to construct UALIGN dataset.