

U2ALIGN: Mitigating Hallucinations for Large Language Models via Uncertainty-Aware Alignment

Anonymous ACL submission

Abstract

Although Large Language Models (LLMs) have exhibited impressive capability across various tasks, they are still over-confident in generating certain but incorrect responses to unknown questions, leading to hallucinations. Uncertainty estimation becomes essential to improve the reliability of LLMs. Existing methods to improve uncertainty expressions in LLMs have not significantly improved **intrinsic reliability** and often lack explainability. Additionally, these methods do not effectively leverage uncertainty estimations to elicit accurate responses. In this work, we propose an **Uncertainty-Aware Alignment** (U2ALIGN) framework by utilizing uncertainty-aware knowledge boundary estimations to elicit accurate responses in LLM alignment. We first create U2ALIGN dataset to integrate LLMs’ intrinsic representation of knowledge boundaries using two uncertainty-aware measures. We also introduce additional unknown questions in the dataset to enhance LLMs’ ability to refuse unknown questions. Subsequently, we train a reward model using the dataset using supervised fine-tuning (SFT) and apply reinforcement learning (RL) to elicit accurate expressions by incorporating uncertainty-aware knowledge boundary information into prompts. Experimental results demonstrate our proposed U2ALIGN effectively enhances LLMs’ reliability of uncertainty estimations and mitigates hallucinations on several tasks.

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 frequently exhibit overconfidence, generating certain yet incorrect responses to unknown questions, resulting in hallucinations

(Ye et al., 2023; Liu et al., 2024). To alleviate overconfidence and enhance the reliability of LLMs, reliable uncertainty estimation is essential (Geng et al., 2023; Xiong et al., 2024).

However, previous efforts to elicit uncertainty expressions in LLMs through prompting or sampling strategies have not significantly improved the intrinsic uncertainty estimation capabilities (Xiong et al., 2024; Zhou et al., 2023; Chen and Mueller, 2023). While some approaches teach LLMs to verbally express uncertainty (Lin et al., 2022a; Han et al., 2024), these methods lack an explainable view for uncertainty estimation and still tend to be over-confident when verbalizing uncertainty (Xiong et al., 2024). Moreover, previous work leveraging uncertainty estimations is only limited in refusal to unknown questions (Zhang et al., 2024a), and does not effectively elicit accurate responses of LLMs.

If an LLM cannot discern whether a question pertains to known or unknown, it is likely to produce fabricated or hallucinated contents (Liu et al., 2024). This tendency arises from the LLMs’ inaccurate perception of the knowledge boundary, leading to over-confidence (Ren et al., 2023). This insight highlights the importance of capturing reliable uncertainty of responses by leveraging the intrinsic representation of LLMs’ knowledge boundaries. Furthermore, while reliable uncertainty estimations can help refrain from answering unknown questions (Zhang et al., 2024a), LLMs may occasionally falter in conveying accurate information on partially known questions (Zhang et al., 2024b). Since alignment is a standard procedure to improve LLMs’ helpfulness and factuality (Bai et al., 2022a), it is promising to explicitly leverage uncertainty scores that incorporate knowledge boundary information to elicit factually correct expressions during LLM alignment stages.

Specifically, we begin our method by pinpointing two essential research questions. **RQ1: How**

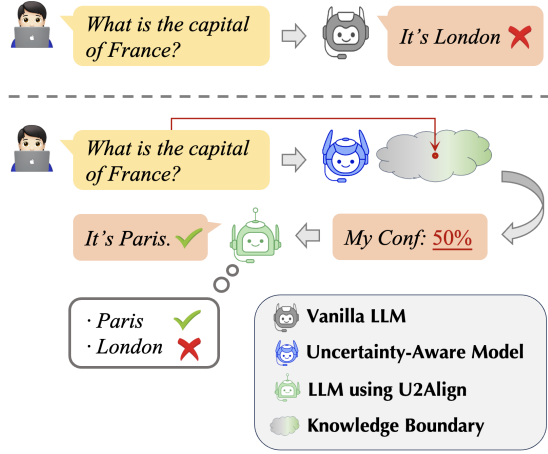


Figure 1: Examples of answering a question on the vanilla LLM (upper) and LLM using U2ALIGN method.

to capture reliable uncertainty estimations by leveraging LLMs’ inherent perception of knowledge boundary? **RQ2: How to explicitly apply the obtained uncertainty estimations to elicit LLMs’ accurate responses?** Inspired by the viewpoints, we propose an **Uncertainty-Aware Alignment (U2ALIGN)** framework. Our framework can be divided into two steps as follows.

First, we incorporate knowledge boundary information into uncertainty estimations and create an U2ALIGN dataset for LLM alignment for **RQ1**. Ren et al. (2023) demonstrates the prior judgments of knowledge boundaries tend to be over-confident. To accurately represent the perception of knowledge boundaries, our uncertainty estimation is conducted by sampling multiple responses to a question with varied prompting and decoding strategies. We adopt two uncertainty-aware measures: consistency-based confidence score and semantic entropy (Kuhn et al., 2023). The confidence score reflects factual accuracy, while semantic entropy indicates the dispersion of generated responses based on internal knowledge. In comparison, Han et al. (2024); Xiong et al. (2024); Zhang et al. (2024b) have only considered confidence scores to represent uncertainty estimation. Recognizing that LLMs may produce hallucinations when encountering new or unknown information during instruction-following fine-tuning (Lin et al., 2024), we also introduce additional unknown questions to bolster LLMs’ generalization to refuse unknown questions.

Then regarding **RQ2**, utilizing the constructed U2ALIGN dataset, we employ supervised fine-tuning (SFT) to explicitly learn the two uncertainty-aware knowledge boundary measures using an

uncertainty-aware model. Conditioned on the question, the generated response, and uncertainty-aware knowledge boundary estimations, we further train a reward model to judge the correctness of the response using SFT. Subsequently, we apply reinforcement learning (RL) to elicit LLMs’ accurate expressions using the reward model by explicitly incorporating uncertainty-aware knowledge boundary information into the prompts using Proximal Policy Optimization (PPO) (Schulman et al., 2017) algorithm. As illustrated in Figure 1, our proposed method can explicitly learn the uncertainty-aware knowledge boundary. Despite only 50% known to a question, the LLM equipped with U2ALIGN framework can elicit the accurate answer, mitigating the hallucinated generations.

In summary, our contributions are as follows.

1) To the best of our knowledge, this work is the first to leverage the intrinsic representation of knowledge boundary for uncertainty estimation. We construct an U2ALIGN dataset to explicitly learn uncertainty-aware measures on LLMs.

2) Using the uncertainty-aware knowledge boundary measures and U2ALIGN dataset, we devise an alignment framework including both SFT and RL methods to elicit LLMs to produce accurate responses conditioned on uncertainty estimations to the given questions.

3) We conduct experiments using LLaMA-2 and Vicuna on several datasets to validate the effectiveness of our proposed framework. The experimental results show that our method significantly improves the LLMs’ reliability by eliciting accurate responses.

2 Related Works

Uncertainty Estimation for LLMs 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). Despite there are marginal distinctions between them (Lin et al., 2023), we will further clarify in Sec. 3.1.2. This paper categorizes previous uncertainty estimation methods on LLMs into four distinct classes, as illustrated in Figure 4. ① **Likelihood-based method**: Following traditional calibration research on classification tasks (Guo et al., 2017), some works intermediately quantify sentence uncertainty over token probabilities (Vazhentsev et al., 2023; Varshney et al., 2023;

Xue et al., 2024). ② **Self-verbalized method**: Recently, LLMs’ remarkable instruction-following ability provides a view of eliciting uncertainty expressions in words (Lin et al., 2022a; Zhou et al., 2023; Tian et al., 2023a; Xiong et al., 2024), or instructing LLMs to self-evaluate its correctness by regenerating and accessing $p(\text{True})$ (Kadavath et al., 2022). ③ **Sampling-based method**: By sampling multiple responses to a given question, Xiong et al. (2024); Lyu et al. (2024); Chen and Mueller (2023) aggregate all the confidence scores as the indicator. Kuhn et al. (2023) proposes semantic entropy to quantify uncertainty at the semantic level. ④ **Training-based method**: Mielke et al. (2022); Lin et al. (2022a) propose to train an external NLI model or LLM itself to improve linguistic uncertainty expression;

However, both self-verbalized and sampling methods elicit more reliable uncertainty estimations for LLMs using extrinsic prompting or aggregation strategies with additional inference-time costs, failing to improve LLMs’ intrinsic capability of uncertainty estimation. Recent works investigate confidence learning methods to enhance the reliability of LLMs on single-token generation (Han et al., 2024). Yang et al. (2023) proposes an uncertainty-aware in-context learning method which leverages 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 U2ALIGN framework not only obtains more reliable uncertainty estimations regarding knowledge boundary information but also elicits accurate responses of LLMs.

LLM Alignment 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; Ziegler et al., 2020; Bai et al., 2022a) or AI feedback (Bai et al., 2022b). Distinct from recent studies that apply RL to enhance LLMs’ factuality (Zhang et al., 2024b; Lin et al., 2024; Tian et al., 2024), this work focuses on improving LLMs’ reliability with an uncertainty-aware alignment framework using SFT and PPO (Schulman et al., 2017).

3 Methodology

We introduce our proposed uncertainty-aware alignment (U2ALIGN) framework, which is di-

vided into two steps: The first step involves constructing an U2ALIGN dataset which incorporates uncertainty-aware knowledge boundary information of question-answering (QA) pairs as illustrated in Figure 2. The second step is to utilize the obtained U2ALIGN dataset to explicitly learn the uncertainty-aware measures for LLMs regarding knowledge boundaries by SFT, and further elicit LLMs to produce accurate responses conditioned on obtained uncertainty estimations using PPO as shown in Figure 3.

3.1 U2ALIGN Dataset Construction

In this section, we elaborate the construction of U2ALIGN dataset, including known questions to explore knowledge boundaries of the questions, uncertainty-aware measures to leverage intrinsic representations of knowledge boundary, and unknown questions to enhance the ability of LLMs to refrain unknown questions.

3.1.1 Known Question Construction

As shown in Figure 2, to explore the knowledge boundary of the LLM given a question, we sample multiple responses to the same question by repeating the generation procedure several times. In this phase, the known question construction process can be represented in a tuple $(Q, \mathcal{P}, \mathcal{A})$. Q contains a batch of n question-answering pairs $\{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) | i \in \{1, 2, \dots, n\}\}$ where \mathbf{x}_i and $\tilde{\mathbf{y}}_i$ denote i -th question and ground-truth answer respectively. We also devise various prompting strategies in \mathcal{P} and decoding strategies in \mathcal{A} to mitigate context sensitivity and capture a more accurate representation of the knowledge boundaries.

Prompt Strategies We employ three prompting strategies \mathcal{P} including vanilla QA prompt, chain-of-thought (CoT) (Wei et al., 2022), and one-shot method (Brown et al., 2020). The designed prompt templates can be referred in Appendix.

Decoding Strategies We also introduce three decoding strategies in \mathcal{A} for LLM inference, including greedy search, and the random sampling of Top-k and Top-p methods.

Multiple Sampling By combining previous techniques with the prepared questions, we capture the LLMs’ intrinsic perception of knowledge boundaries on Q . In the k -th sampling process for the i -th question \mathbf{x}_i , we randomly select one prompting strategy $p_i^{(k)} \in \mathcal{P}$ and one decoding strat-

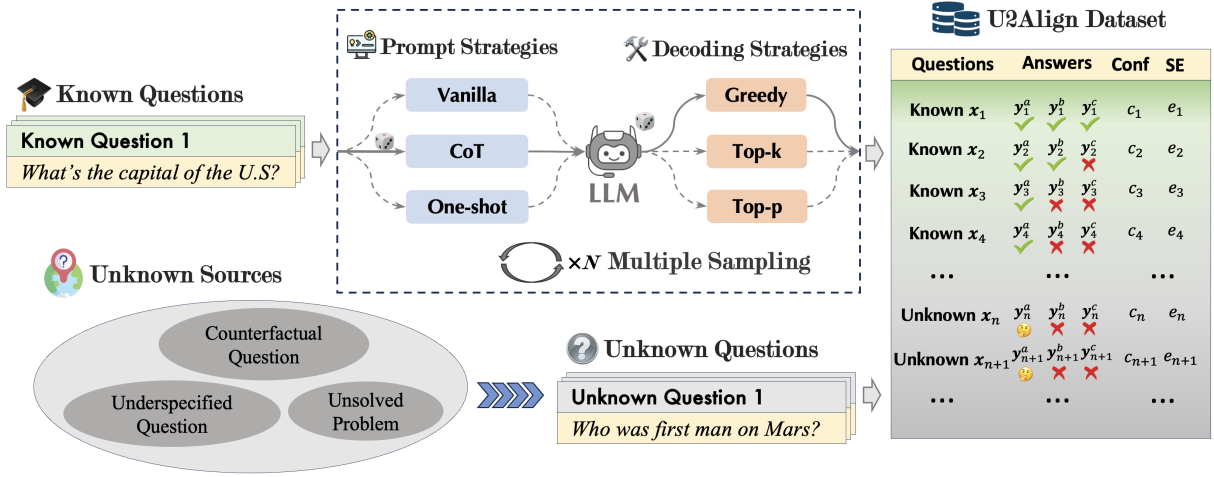


Figure 2: Illustration of U2ALIGN dataset construction.

egy $a_i^{(k)} \in \mathcal{A}$ to generate a response $y_i^{(k)}$ using the LLM. By taking K times of the sampling process, we can obtain an answer set $\mathbf{Y}_i = \{y_i^{(k)} | k \in \{1, \dots, K\}\}$ to question x_i .

3.1.2 Uncertainty-Aware Measures

In order to quantify the intrinsic representations of knowledge boundaries, we can leverage the uncertainty estimation methods to indicate the known level. We develop uncertainty-aware measures in two aspects, including the certainty level to a question x_i and the dispersion of the generated answers in the set \mathbf{Y}_i . Therefore, we adopt self-consistency based confidence (Xiong et al., 2024) and semantic entropy (Kuhn et al., 2023) to jointly determine the intrinsic knowledge boundary information.

Self-consistency based Confidence A natural idea of aggregating varied responses is to measure the degree of agreement among the candidate outputs and integrate the inherent uncertainty (Manakul et al., 2023; Xiong et al., 2024). Given a question x_i , the ground-truth \tilde{y}_i , and the associated answer set \mathbf{Y}_i , the agreement between these candidate responses and the ground-truth answer then serves as a measure of confidence score c_i , computed as follows:

$$c_i = \text{Conf}(x_i) = \frac{1}{K} \sum_{k=1}^K \mathbb{1} \{ \tilde{y}_i = y_i^{(k)} \} \quad (1)$$

Semantic Entropy Due to the variable length of generated sequences leading to limitless 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 x_i and \mathbf{Y}_i is calculated as

$$p(c|x_i) = \sum_{y_i^{(k)} \in c} p(y_i^{(k)}|x_i) \quad (2)$$

$$e_i = \text{SE}(x_i) = - \sum_c p(c|x_i) \log p(c|x_i) \quad (3)$$

where c denotes a set of sentences in semantic equivalent space. We format the data point in $(x_i, \mathbf{Y}_i, c_i, e_i)$ as the i -th sample and append in the U2ALIGN dataset \mathcal{D} .

3.1.3 Unknown Question Collection

Since LLMs are prone to produce hallucinations when faced with unknown information in instruction-following SFT stage (Lin et al., 2024), it is essential to introduce definitely unknown questions to enhance LLMs' refusal ability (Zhang et al., 2024a). we have identified a series of unknown source categories following (Amayuelas et al., 2023) as presented in Figure 2. For unknown questions, we set the ground-truth label as refusal responses such as "I don't know ..." or "I am not sure ..." while we also add certain answers as negative samples for the penalty. Likewise, we also calculate uncertainty-aware measures and semantic entropy for unknown questions and format as known samples in U2ALIGN dataset \mathcal{D} . Assume that there are totally N samples in \mathcal{D} including n knowns and $N - n$ unknowns.

3.2 U2ALIGN Training Strategies

In this section, we demonstrate the training processes with the obtained U2ALIGN dataset \mathcal{D} . The U2ALIGN training strategy is divided into

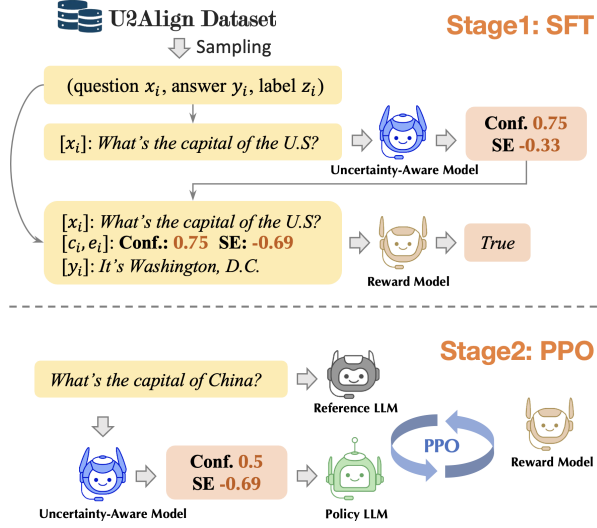


Figure 3: Illustration of SFT and PPO alignment processes of U2ALIGN framework.

two stages as illustrated in Figure 3: Stage 1 - U2ALIGN SFT is to train an uncertainty-aware model to explicitly learn the uncertainty-aware measures given specific questions and a reward model as the evaluator. Stage 2 - U2ALIGN PPO is to make the LLM always elicit the accurate answer to a question using the uncertainty-aware model and the reward model using PPO algorithm.

3.2.1 Stage 1 - U2ALIGN SFT

This fine-tuning process empowers the model to autonomously adjust its responses based on the calculated uncertainty scores.

Supervised Fine-Tuning (SFT) represents a straightforward yet efficient alignment technique. We directly employ the UAlign dataset for the Supervised Fine-tuning of LLMs. Given that the UAlign dataset comprises both question-answer pairs and associated uncertainty scores, it serves as a conditional generation task. We feed the questions into the model and task the model with predicting the responses. The training of our model is conducted using the standard sequence-to-sequence loss.

$$\arg \min P(y|x, \text{Conf}(x_i), \text{SE}(x_i)) \quad (4)$$

For each question, we elicit the knowledge by generating the uncertainty-aware knowledge boundary representations.

3.2.2 Stage 2 - U2ALIGN PPO

Inspired by the recent progress of reinforcement learning from human feedback (RLHF) technique

to align human preferences (Ouyang et al., 2022; Ziegler et al., 2019), based on our reward model, we can use proximal policy optimization (PPO) (Schulman et al., 2017) to optimize the LLM. We prefer reliable responses that are actually aligned with the uncertainty. As illustrated in Figure 3, the LLM to be optimized is used as the policy model. The response generated by the policy model is fed into the reward model to obtain the consistency reward score which is mainly used to align the uncertainty-aware preferences. The reward model will return a higher score for the confident estimations to facilitate the reliable uncertainty expressions of the policy model. Furthermore, a reference model is also introduced. The inputs used for PPO training are identical to those used for reward modeling, but sample responses in an online fashion.

4 Experimental Settings

4.1 Datasets

Our experiments consider three extensively adopted open-domain question-answering (QA) benchmarks, including **Natural Questions** (NQ) (Kwiatkowski et al., 2019), **TriviaQA** (Joshi et al., 2017), and **TruthfulQA** (Lin et al., 2022b). NQ is constructed by Google Search queries along with annotated short answers or documents (long answers). TriviaQA is an open-form trivia question dataset that gauges models’ factual knowledge. TruthfulQA is a dataset that tests whether models generate truthful answers to questions specifically designed to induce false answers.

4.2 Evaluation Metrics

AUROC Following Hendrycks and Gimpel (2018); Filos et al. (2019), we employ Area Under the Receiver Operator Characteristic Curve (AUROC) score to assess the effectiveness of uncertainty estimations, implement by sklearn toolkit¹. AUROC quantifies how likely a correct answer possesses a higher uncertainty score than an incorrect one, yielding a score within the range of [0, 1]. A higher AUROC score is preferred.

Accuracy We measure the typical accuracy on the test split of the datasets between the generated answer and gold reference, which is assessed using GPT-3.5 Turbo (Correct scores 1 and Incorrect

¹https://github.com/scikit-learn/scikit-learn/blob/main/sklearn/metrics/_ranking.py

scores 0) by calling OpenAI’s API ². The prompt template is presented in Appendix.

4.3 Baselines

Vanilla

SFT

SFT+PPO

4.4 Implementation Details

To construct the answer set, we generate 10 responses for each question. The temperature T is set to 1.0. We conduct experiments on both the Vicuna-1.5 7B (Chiang et al., 2023) and LLaMA-2 7B (Touvron et al., 2023) models. ADAM parameter update is used in a mini-batch mode for all models. During U2ALIGN SFT stage, we fine-tune the LLMs using LoRA and set epoch to 4, learning rate to 1e-5, and batch size to 4. The U2ALIGN RL alignment is implemented by trl³, and all the hyper-parameters related to PPO algorithm are default values by the trl PPOConfig recipe⁴ while we set epoch to 3, learning rate to 1e-5, and batch size to 2. All the experiments are conducted with NVIDIA A100-40GB GPUs.

5 Experimental Results and Analysis

5.1 Main Results

Methods	TriviaQA	TruthfulQA	NQ
LLaMA2			
Vanllia	43.50	22.10	24.80
SFT	54.60	42.70	33.70
SFT+PPO	62.30	53.40	37.60
U2ALIGN	68.20	65.50	45.70
Vicuna			
Vanllia	32.60	25.00	21.30
SFT	40.50	33.70	31.40
SFT+PPO	43.60	38.90	36.10
U2ALIGN	48.70	42.20	44.60

Table 1: Main results of on several QA datasets LLaMA2 and Vicuna.

5.2 Ablation Study and Analysis

5.2.1 The Effect of Uncertainty Estimation Methods

Settings Likelihood method, $P(\text{True})$ method, self-verbalized method. The details can be referred in Appendix B

²<https://platform.openai.com/docs/api-reference/introduction>

³<https://github.com/lvwerra/trl>

⁴https://github.com/huggingface/trl/blob/main/trl/trainer/ppo_config.py

As shown in Table 2, several trends can be found.

Findings 1) The sampling-based method consistently outperforms other baseline uncertainty estimation methods.

Methods	TriviaQA	TruthfulQA	NQ
LLaMA-2			
Log-Norm	73.34	75.35	69.45
$P(\text{True})$	80.23	79.20	76.22
Verbal.	77.34	76.37	73.14
Conf (U2ALIGN)	85.36	86.45	82.06
SE (U2ALIGN)	82.21	87.26	83.49
Vicuna			
Log-Norm	64.25	61.00	58.34
$P(\text{True})$	69.13	65.27	62.20
Verbal.	67.44	66.48	62.55
Conf (U2ALIGN)	73.16	72.59	70.34
SE (U2ALIGN)	71.83	74.59	71.11

Table 2: Experimental results of AUROC \uparrow of uncertainty measure.

Analysis

5.2.2 The Effect of Different Ratio of Known-Unknown Questions

Settings As shown in Table 2, several trends can be found.

Findings 1) Introducing unknown questions can greatly enhance the effectiveness and reliability of models when dealing with unknown questions.

Ratio	TriviaQA	TruthfulQA	NQ
LLaMA-2			
1:1	79.22	79.89	77.64
1:0.5	82.47	83.22	80.15
1:0.33	85.36	86.45	82.06
1:0.25	83.25	84.33	82.04
Vicuna			
1:1	70.44	71.23	68.57
1:0.5	72.59	71.94	70.13
1:0.33	73.16	72.74	70.34
1:0.25	73.29	71.92	70.08

Table 3: Ratio of Known-Unknown Questions in U2ALIGN Dataset.

Analysis

5.2.3 The Effect of Uncertainty Measures in SFT Stage

Settings As shown in Table 2, several trends can be found.

Findings 1) The sampling-based method consistently outperforms other baseline uncertainty estimation methods.

Conf	SE	TriviaQA	TruthfulQA	NQ
LLaMA-2				
✗	✗	90.10	88.20	85.30
✓	✗	93.60	90.20	89.40
✗	✓	92.30	89.50	86.70
✓	✓	93.70	92.30	91.10
Vicuna				
✗	✗	86.40	83.80	84.10
✓	✗	89.40	85.90	86.60
✗	✓	88.70	86.20	87.70
✓	✓	90.50	88.50	89.20

Table 4: Ratio of Known-Unknown Questions in U2ALIGN Dataset.

Analysis

6 Discussions

7 Conclusion

Acknowledgments

References

- Alfonso Amayuelas, Liangming Pan, Wenhui Chen, and William Wang. 2023. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *arXiv preprint arXiv:2305.13712*.
- Rie Kubota Ando and Tong Zhang. 2005. [A framework for learning predictive structures from multiple tasks and unlabeled data](#). *Journal of Machine Learning Research*, 6:1817–1853.
- Galen Andrew and Jianfeng Gao. 2007. [Scalable training of \$L_1\$ -regularized log-linear models](#). In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.
- 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*.
- 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](#).
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Maric, Shane Legg, and Dario Amodei. 2017. [Deep reinforcement learning from human preferences](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Farshid Faal, Ketra Schmitt, and Jia Yuan Yu. 2023. Reward modeling for mitigating toxicity in transformer-based language models. *Applied Intelligence*, 53(7):8421–8435.
- 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*.
- Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koepl, Preslav Nakov, and Iryna Gurevych. 2023. [A survey of language model confidence estimation and calibration](#). *Preprint*, arXiv:2311.08298.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural networks](#). *Preprint*, arXiv:1706.04599.
- Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.
- 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.
- Dan Hendrycks and Kevin Gimpel. 2018. [A baseline for detecting misclassified and out-of-distribution examples in neural networks](#). *Preprint*, arXiv:1610.02136.
- 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

567	Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,	621
568	Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natu-	Carroll Wainwright, Pamela Mishkin, Chong Zhang,	622
569	ral questions: A benchmark for question answering	Sandhini Agarwal, Katarina Slama, Alex Ray, John	623
570	research . <i>Transactions of the Association for Compu-</i>	Schulman, Jacob Hilton, Fraser Kelton, Luke Miller,	624
571	tational Linguistics , 7:452–466.	Maddie Simens, Amanda Askill, Peter Welinder,	625
572	Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan	Paul F Christiano, Jan Leike, and Ryan Lowe. 2022.	626
573	Xiong, Jimmy Lin, Wen tau Yih, and Xilun Chen.	Training language models to follow instructions with	627
574	2024. Flame: Factuality-aware alignment for large	human feedback . In <i>Advances in Neural Information</i>	628
575	language models . <i>Preprint</i> , arXiv:2405.01525.	<i>Processing Systems</i> , volume 35, pages 27730–27744.	629
576	Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a.	Curran Associates, Inc.	630
577	Teaching models to express their uncertainty in	Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015.	631
578	words . <i>Transactions on Machine Learning Research</i> .	Yara parser: A fast and accurate dependency parser .	632
579	Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b.	<i>Computing Research Repository</i> , arXiv:1503.06733.	633
580	TruthfulQA: Measuring how models mimic human	Version 2.	634
581	falsehoods . In <i>Proceedings of the 60th Annual Meet-</i>	Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin	635
582	<i>ing of the Association for Computational Linguistics</i>	Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen,	636
583	<i>(Volume 1: Long Papers)</i> , pages 3214–3252, Dublin,	and Haifeng Wang. 2023. Investigating the factual	637
584	Ireland. Association for Computational Linguistics.	knowledge boundary of large language models with	638
585	Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023.	retrieval augmentation . <i>Preprint</i> , arXiv:2307.11019.	639
586	Generating with confidence: Uncertainty quantifi-	John Schulman, Filip Wolski, Prafulla Dhariwal,	640
587	cation for black-box large language models. <i>arXiv</i>	Alec Radford, and Oleg Klimov. 2017. Proxi-	641
588	<i>preprint arXiv:2305.19187</i> .	mal policy optimization algorithms. <i>arXiv preprint</i>	642
589	Genglin Liu, Xingyao Wang, Lifan Yuan, Yangyi Chen,	<i>arXiv:1707.06347</i> .	643
590	and Hao Peng. 2024. Examining llms’ uncertainty ex-	Katherine Tian, Eric Mitchell, Huaxiu Yao, Christo-	644
591	pression towards questions outside parametric knowl-	pher D Manning, and Chelsea Finn. 2024. Fine-	645
592	edge . <i>Preprint</i> , arXiv:2311.09731.	tuning language models for factuality . In <i>The Twelfth</i>	646
593	Qing Lyu, Kumar Shridhar, Chaitanya Malaviya,	<i>International Conference on Learning Representa-</i>	647
594	Li Zhang, Yanai Elazar, Niket Tandon, Marianna	<i>tions</i> .	648
595	Apidianaki, Mrinmaya Sachan, and Chris Callison-	Katherine Tian, Eric Mitchell, Allan Zhou, Archit	649
596	Burch. 2024. Calibrating large language models with	Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn,	650
597	sample consistency . <i>Preprint</i> , arXiv:2402.13904.	and Christopher Manning. 2023a. Just ask for cali-	651
598	Andrey Malinin and Mark Gales. 2021. Uncertainty	bration: Strategies for eliciting calibrated confidence	652
599	estimation in autoregressive structured prediction . In	scores from language models fine-tuned with human	653
600	<i>International Conference on Learning Representa-</i>	feedback . In <i>Proceedings of the 2023 Conference</i>	654
601	<i>tions</i> .	<i>on Empirical Methods in Natural Language Process-</i>	655
602	Potsawee Manakul, Adian Liusie, and Mark Gales. 2023.	<i>ing</i> , pages 5433–5442, Singapore. Association for	656
603	SelfCheckGPT: Zero-resource black-box hallucina-	Computational Linguistics.	657
604	tion detection for generative large language models .	Katherine Tian, Eric Mitchell, Allan Zhou, Archit	658
605	In <i>Proceedings of the 2023 Conference on Empiri-</i>	Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn,	659
606	<i>cal Methods in Natural Language Processing</i> , pages	and Christopher D Manning. 2023b. Just ask for cali-	660
607	9004–9017, Singapore. Association for Computa-	bration: Strategies for eliciting calibrated confidence	661
608	tional Linguistics.	scores from language models fine-tuned with human	662
609	Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-	feedback . <i>arXiv preprint arXiv:2305.14975</i> .	663
610	Lan Boureau. 2022. Reducing conversational agents’	Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier	664
611	overconfidence through linguistic calibration . <i>Trans-</i>	Martinet, Marie-Anne Lachaux, Timothée Lacroix,	665
612	<i>actions of the Association for Computational Linguis-</i>	Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal	666
613	<i>tics</i> , 10:857–872.	Azhar, Aurelien Rodriguez, Armand Joulin, Edouard	667
614	Kenton Murray and David Chiang. 2018. Correcting	Grave, and Guillaume Lample. 2023. Llama: Open	668
615	length bias in neural machine translation . In <i>Proceed-</i>	and efficient foundation language models . <i>Preprint</i> ,	669
616	<i>ings of the Third Conference on Machine Translation:</i>	arXiv:2302.13971.	670
617	<i>Research Papers</i> , pages 212–223, Brussels, Belgium.	Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jian-	671
618	Association for Computational Linguistics.	shu Chen, and Dong Yu. 2023. A stitch in time saves	672
619	OpenAI. 2023. Gpt-4 technical report . <i>Preprint</i> ,	nine: Detecting and mitigating hallucinations of	673
620	arXiv:2303.08774.	llms by validating low-confidence generation. <i>arXiv</i>	674
		<i>preprint arXiv:2307.03987</i> .	675
		Artem Vazhentsev, Akim Tsvigun, Roman Vashurin,	676
		Sergey Petrakov, Daniil Vasilev, Maxim Panov,	677

678	Alexander Panchenko, and Artem Shelmanov. 2023.	Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B	734
679	Efficient out-of-domain detection for sequence to sequence models . In <i>Findings of the Association for Computational Linguistics: ACL 2023</i> , pages 1430–	Brown, Alec Radford, Dario Amodei, Paul Chris-	735
680	1454, Toronto, Canada. Association for Computa-	tiano, and Geoffrey Irving. 2019. Fine-tuning lan-	736
681	tional Linguistics.	guage models from human preferences. <i>arXiv</i>	737
682		<i>preprint arXiv:1909.08593</i> .	738
683			
684	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten	Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B.	739
685	Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le,	Brown, Alec Radford, Dario Amodei, Paul Chris-	740
686	and Denny Zhou. 2022. Chain-of-thought prompt-	tiano, and Geoffrey Irving. 2020. Fine-tuning lan-	741
687	ing elicits reasoning in large language models . In	guage models from human preferences . <i>Preprint</i> ,	742
688	<i>Advances in Neural Information Processing Systems</i> ,	arXiv:1909.08593.	743
689	volume 35, pages 24824–24837. Curran Associates,		
690	Inc.		
691	Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie		
692	Neiswanger, Ruslan Salakhutdinov, and Louis-		
693	Philippe Morency. 2022. Uncertainty quantification		
694	with pre-trained language models: A large-scale em-		
695	pirical analysis . In <i>Findings of the Association for</i>		
696	<i>Computational Linguistics: EMNLP 2022</i> , pages		
697	7273–7284, Abu Dhabi, United Arab Emirates. As-		
698	sociation for Computational Linguistics.		
699	Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie		
700	Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs		
701	express their uncertainty? an empirical evaluation of		
702	confidence elicitation in LLMs . In <i>The Twelfth Inter-</i>		
703	<i>national Conference on Learning Representations</i> .		
704	Boyang Xue, Hongru Wang, Weichao Wang, Rui Wang,		
705	Sheng Wang, Zeming Liu, and Kam-Fai Wong. 2024.		
706	A comprehensive study of multilingual confidence		
707	estimation on large language models . <i>Preprint</i> ,		
708	arXiv:2402.13606.		
709	Yuchen Yang, Houqiang Li, Yanfeng Wang, and		
710	Yu Wang. 2023. Improving the reliability of large		
711	language models by leveraging uncertainty-aware in-		
712	context learning . <i>Preprint</i> , arXiv:2310.04782.		
713	Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and		
714	Weiqiang Jia. 2023. Cognitive mirage: A review of		
715	hallucinations in large language models . <i>Preprint</i> ,		
716	arXiv:2309.06794.		
717	Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung,		
718	Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji,		
719	and Tong Zhang. 2024a. R-tuning: Instructing large		
720	language models to say ‘i don’t know’ . <i>Preprint</i> ,		
721	arXiv:2311.09677.		
722	Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou,		
723	Lifeng Jin, Linfeng Song, Haitao Mi, and Helen		
724	Meng. 2024b. Self-alignment for factuality: Mit-		
725	igating hallucinations in llms via self-evaluation .		
726	<i>Preprint</i> , arXiv:2402.09267.		
727	Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto.		
728	2023. Navigating the grey area: How expressions		
729	of uncertainty and overconfidence affect language		
730	models . In <i>Proceedings of the 2023 Conference on</i>		
731	<i>Empirical Methods in Natural Language Processing</i> ,		
732	pages 5506–5524, Singapore. Association for Com-		
733	putational Linguistics.		

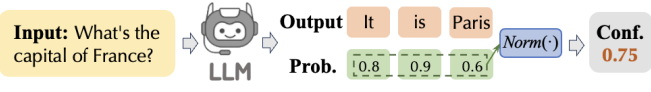
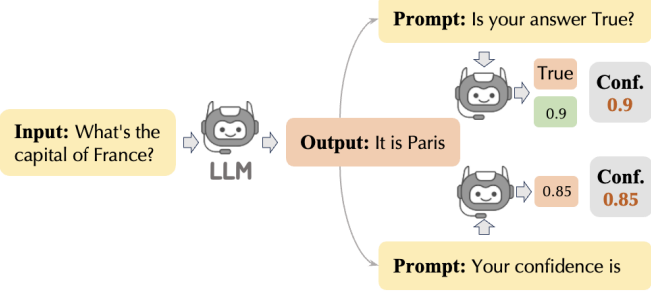
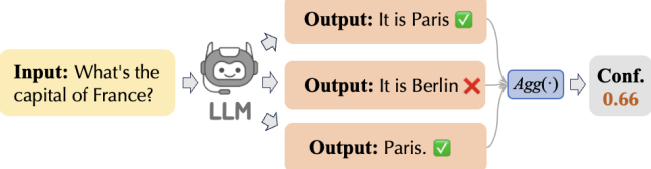
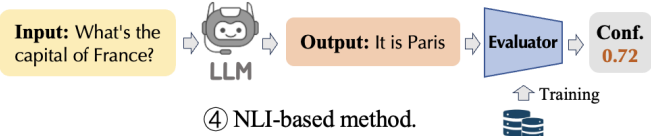
Confidence & Uncertainty Estimation Methods on LLMs	Disadvantages
 <p>① Likelihood-based method.</p>	<ul style="list-style-type: none"> a. Requires normalization due to variable sequence length; b. Requires access to token-level probabilities, inapplicable to black-box LLMs; c. Fails to capture semantic meaning over token-level probabilities.
 <p>② Self-verbalized method.</p>	<ul style="list-style-type: none"> a. Relies on prompting strategies to elicit confidence estimation, varying in different methods (Is <i>True</i> probability, numerical confidence, and word expressions, etc.); b. Cannot improve LLM's intrinsic confidence estimation ability. c. Prone to be over-confident.
 <p>③ Sampling-based method.</p>	<ul style="list-style-type: none"> a. Requires additional inference time cost; b. Varying in different aggregation methods; c. Cannot improve LLM's intrinsic confidence estimation ability.
 <p>④ NLI-based method.</p>	<ul style="list-style-type: none"> a. Requires training an additional evaluator; b. Difficult to learn LLM's intrinsic confidence estimation on unseen domains.

Figure 4: Uncertainty Estimation in Generative LLMs.

A Example Appendix

This is an appendix.

B Baseline Details

B.1 Confidence Estimation Methods

In this section, we investigate several commonly used confidence estimation methods for generative LLMs. Specifically, we denote $\text{Conf}(x, y)$ as the confidence score associated with the output sequence $y = [y_1, y_2, \dots, y_N]$ given the input context $x = [x_1, x_2, \dots, x_M]$.

Likelihood-based Confidence: The likelihood-based confidence is estimated by calculating the

joint token-level probabilities over y conditioned on x . As longer sequences are supposed to have lower joint likelihood probabilities that shrink exponentially with length, we calculate the geometric mean by normalizing the product of conditional token probabilities in the output by the sequence length (**Likelihood-Norm**) (Murray and Chiang, 2018; Malinin and Gales, 2021), and the confidence can be represented as:

$$\text{Conf}(x, y) = \left(\prod_i^N p(y_i | y_{<i}, x) \right)^{\frac{1}{N}} \quad (5)$$

Similarly, we also take the average (**Likelihood-Avg**) of the probabilities of tokens arithmetically:

$$\text{Conf}(\mathbf{x}, \mathbf{y}) = \frac{1}{M} \sum_i^M p(y_i | \mathbf{y}_{<i}, \mathbf{x}) \quad (6)$$

Furthermore, a low probability associated with even one generated token may provide more informative evidence of uncertainty (Varshney et al., 2023). Hence, we also employ the minimum of token probabilities. (**Likelihood-Min**).

$$\text{Conf}(\mathbf{x}, \mathbf{y}) = \min \{p(y_1 | \mathbf{x}), \dots, p(y_N | \mathbf{y}_{<N}, \mathbf{x})\} \quad (7)$$

True Probability Confidence: The **True Probability** confidence score is implemented by simply asking the model itself if its first proposed answer \mathbf{y} to the question \mathbf{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.

$$\text{Conf}(\mathbf{x}, \mathbf{y}) = p(\text{True}) = p(\mathbf{y} \text{ is True} | \mathbf{x}) \quad (8)$$

Self-verbalized Confidence: Recent works pay particular attention to linguistic confidence via prompting LLMs to express certainty 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). We adopt verbalized numerical probability (**Verbal Number**) and words (**Verbal Word**) in token-level space as LLM’s confidence estimations. The verbalized word contains a set of five words (e.g. “lowest”, “low”, “medium”, “high”, “highest”) indicating the confidence degrees.

C Training

C.1 U2ALIGN PPO

We prefer reliable responses that are actually aligned with the uncertainty. Aligning with the factual consistency preference can implicitly encourage LLMs to convey reliable responses. Inspired by the recent progress of reinforcement learning from human feedback (RLHF) technique to align human preferences (Ouyang et al., 2022; Ziegler et al., 2019; Christiano et al., 2017) like mitigating toxicity (Faal et al., 2023). As illustrated in Figure 3, the LLM to be optimized is used as the policy model. The response generated by the policy model is fed into the reward model to obtain the consistency reward score r_1 which is mainly used

to align the uncertainty-aware preferences. The reward model will return a higher score for the confident estimations to facilitate the reliable uncertainty expressions of the policy model. Furthermore, a reference model generating a response is also introduced. The KL divergence r_2 between the outputs of the reference model and the policy model is used as an extra reward signal to make sure the generated responses don’t diverge too far from the originals. The optimization objective $r = r_1 + r_2$ is utilized for RL training via the Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017). Utilizing our reward model, we employ proximal policy optimization (PPO) to optimize the model. The inputs used for PPO training are identical to those used for reward modeling, yet responses are sampled in an online fashion.

Command	Output	Command	Output
<code>{\`a}</code>	ä	<code>{\c c}</code>	ç
<code>{\^e}</code>	ê	<code>{\u g}</code>	ğ
<code>{\`i}</code>	ì	<code>{\l}</code>	ł
<code>{\.I}</code>	İ	<code>{\~n}</code>	ñ
<code>{\o}</code>	ø	<code>{\H o}</code>	ő
<code>{\'u}</code>	ú	<code>{\v r}</code>	ř
<code>{\aa}</code>	å	<code>{\ss}</code>	ß

Table 5: Example commands for accented characters, to be used in, *e.g.*, BibT_EX entries.

Please ensure that BibT_EX records contain DOIs or URLs when possible, and for all the ACL materials that you reference. Use the `doi` field for DOIs and the `url` field for URLs. If a BibT_EX entry has a URL or DOI field, the paper title in the references section will appear as a hyperlink to the paper, using the `hyperref` L^AT_EX package.

C.2 References

The L^AT_EX and BibT_EX style files provided roughly follow the American Psychological Association format. If your own bib file is named `custom.bib`, then placing the following before any appendices in your L^AT_EX file will generate the references section for you:

```
\bibliography{custom}
```

You can obtain the complete ACL Anthology as a BibT_EX file from <https://aclweb.org/anthology/anthology.bib.gz>. To include both the Anthology and your own `.bib` file, use the following instead of the above.

```
\bibliography{anthology,custom}
```

Please see Section D for information on preparing BibT_EX files.

C.3 Equations

An example equation is shown below:

$$A = \pi r^2 \quad (9)$$

Labels for equation numbers, sections, subsections, figures and tables are all defined with the `\label{label}` command and cross references to them are made with the `\ref{label}` command.

This an example cross-reference to Equation 9.

C.4 Appendices

Use `\appendix` before any appendix section to switch the section numbering over to letters. See Appendix A for an example.

D BibT_EX Files

Unicode cannot be used in BibT_EX entries, and some ways of typing special characters can disrupt BibT_EX's alphabetization. The recommended way of typing special characters is shown in Table 5.

Output	natbib command	ACL only command
(Gusfield, 1997)	\citep	
Gusfield, 1997	\citealp	
Gusfield (1997)	\citet	
(1997)	\citeyearpar	
Gusfield's (1997)		\citeposs

Table 6: Citation commands supported by the style file. The style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.