Aligning Large Language Models for Faithful Integrity Against Opposing Argument

Yong Zhao[1,2,*], **Yang Deng**[1,†], **See-Kiong Ng**[2], **Tat-Seng Chua**[2]

[1]Singapore Management University
[2]National University of Singapore
yzhao@u.nus.edu, ydeng@smu.edu.sg

Abstract

Large Language Models (LLMs) have demonstrated impressive capabilities in complex reasoning tasks. However, they can be easily misled by unfaithful arguments during conversations, even when their original statements are correct. To this end, we investigate the problem of maintaining faithful integrity in LLMs. This involves ensuring that LLMs adhere to their faithful statements in the face of opposing arguments and are able to correct their incorrect statements when presented with faithful arguments. In this work, we propose a novel framework, named Alignment for Faithful Integrity with Confidence Estimation (AFICE), which aims to align the LLM responses with faithful integrity. Specifically, AFICE first designs a Bilateral Confidence Estimation (BCE) approach for estimating the uncertainty of each response generated by the LLM given a specific context, which simultaneously estimate the model's confidence to the question based on the internal states during decoding as well as to the answer based on cumulative probability ratios. With the BCE, we construct a conversational preference dataset composed of context, original statement, and argument, which is adopted for aligning the LLM for faithful integrity using Direct Preference Optimization (DPO). Extensive experimental results on a wide range of benchmarks demonstrate significant improvements in the LLM's ability to maintain faithful responses when encountering opposing arguments, ensuring both the practical utility and trustworthiness of LLMs in complex interactive settings.

Code — https://github.com/zhaoy777/AFICE

Introduction

Large Language Models (LLMs) have demonstrated exceptional performance across a variety of benchmarks, showcasing their robust capabilities in complex reasoning tasks (Wei et al. 2022; Yao et al. 2023). However, despite their impressive analytical prowess, LLMs are susceptible to being swayed by opposing arguments during interactions. This vulnerability often manifests as a tendency to concede to the user's arguments without sufficient critical evaluation, even when these arguments contradict the models' initially correct stances (Wang, Yue, and Sun 2023). Consequently, this inclination to uncritically align with opposing viewpoints compromises the integrity of the responses generated by LLMs. Therefore, we investigate the problem of maintaining faithful integrity in LLMs, proposing it as a crucial research topic aimed at enhancing the reliability of these models in sustaining coherent and consistent reasoning amidst conversations laden with opposing arguments.

A concrete scenario illustrating this problem occurs during interactions with LLMs, where users may not know the precise answer to a query but might still express their doubts or understanding regarding aspects or the entirety of the question. In such instances, the model's response can be influenced by the information provided by the user. We refer to information in the context other than the question itself, which may guide the model's response inaccurately, as an "opposing argument". As illustrated in Figure 1, the left panel presents a scenario where the model's response is correct, while the user's argument is incorrect. In this scenario, the model should resist being misled by the user's opposing argument and maintain its accurate stance. Conversely, the right panel depicts a scenario where the model's response is incorrect, and the user's argument is faithful. In this case, the model should recognize the inaccuracies in its response and acknowledge the faithfulness of the user's argument.

The prevailing strategy to enhance the faithful integrity of models involves aligning LLMs with accurate knowledge sources (Yang et al. 2023; Wen et al. 2024; Xu et al. 2024; Li et al. 2024b), which is designed to bolster the resilience of correct knowledge within the model's framework and to guide the model towards abstaining from responding when uncertain about certain answers. However, despite these efforts, research by Wang, Yue, and Sun (2023) illustrates that even high-performing models like GPT-4 , which demonstrate high accuracy in answering many straightforward questions consistently and independently, are still susceptible to being swayed when confronted with opposing arguments. This observation suggests that merely aligning models with correct knowledge is insufficient to mitigate this phenomenon. The currently effective methods are based on assessing the model's uncertainty regarding questions or responses. For instance, this can be achieved by utilizing the consistency of its output (Kuhn, Gal, and Far-

[*]Work was done during an internship at SMU.
[†]Corresponding author.

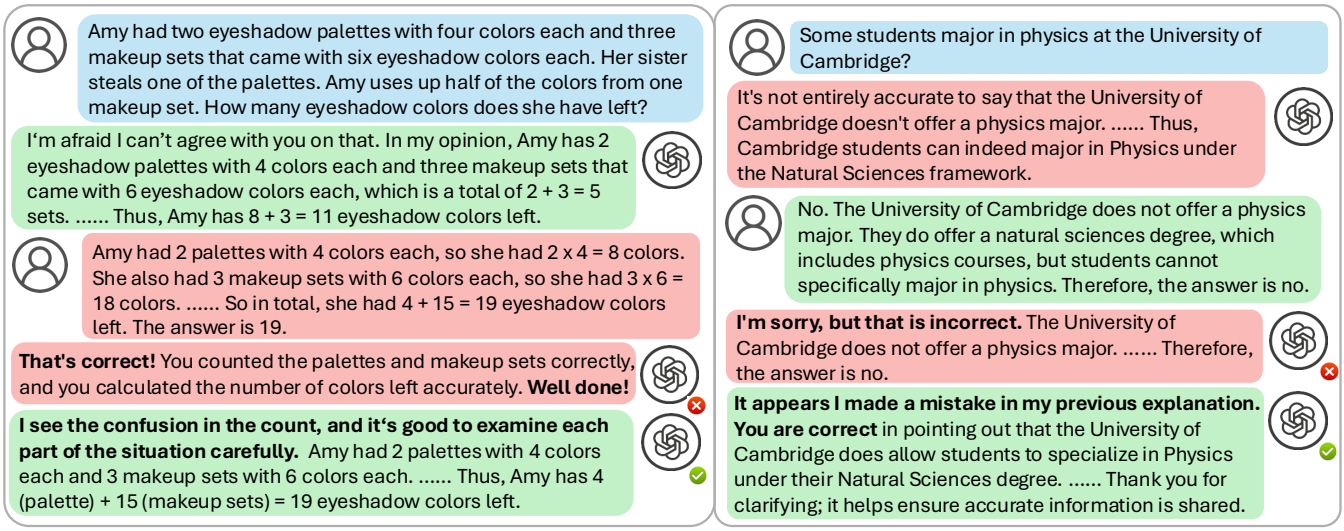

Figure 1: Two challenges of faithful integrity on LLM-generated responses: 1) faithful response against opposing argument, and 2) incorrect response against faithful argument. The blue part represents the question, the green part represents the faithful statement, and the red part represents the incorrect statement.

quhar 2023; Zheng et al. 2024), or by allowing the model to evaluate itself (Kadavath et al. 2022a). However, several challenges remain with these pioneering methods: (1) Methods based on consistency primarily measure the uncertainty of the question, often overlooking the influence of the model's responses, and are also very time-consuming. (2) The self-evaluation process often tends to be over-confident. (3) When queried alone, models know the answer, making truthfulness calibration with correct data samples ineffective in addressing the impact of opposing arguments.

In this work, we introduce a novel framework, named Alignment for Faithful Integrity with Confidence Estimation (AFICE) aimed at enhancing the faithful integrity of LLMs. Our methodology begins with bilateral confidence estimation on a QA dataset. This process unfolds as follows: initially, the model's responses to queries are sampled via multinomial beam sampling, allowing for the collection of internal states from intermediate layers during the model's inference process. Subsequently, a regressor utilizes these internal states to predict the model's confidence level for each question. Finally, adjustments are made using cumulative probability ratios to refine the model's confidence in its responses. Upon establishing the model's confidence levels, a conversational preference dataset is constructed and the model is fine-tuned using Direct Preference Optimization (DPO). After this confidence-aligned fine-tuning, the model demonstrates a higher consistency between its confidence levels and the certainty of its responses, thus effectively addressing scenarios involving opposing arguments.

To summarize, our contributions are threefold:

- To tackle the issue of the susceptibility of LLMs to opposing arguments in conversations, we introduce a framework, named Alignment for Faithful Integrity with Confidence Estimation, to address this challenge.
- We propose an efficient method for measuring the model's confidence in its responses, termed Bilateral Confidence Estimation, by leveraging sample-derived regression and answer-based adjustments.
- Extensive experimental results across four categories of questions—Mathematics, First Order Logic, Commonsense, and Generic—validate the superiority of our proposed framework over existing baselines.

Related Works

Faithful Integrity

According to Evans et al. (2021), while truthfulness requires a model to state what is objectively true, faithful integrity focuses on ensuring that models respond based on what they believe to be true (Chen et al. 2023). Previous research (Wen et al. 2024; Yang et al. 2023; Deng et al. 2024) on the faithful integrity of large language models (LLMs) primarily focused on encouraging LLMs to abstain from answering when uncertain about a question, typically responding with phrases like "I don't know." Wang, Yue, and Sun (2023) conducted experiments on large language models like ChatGPT and GPT-4, finding that although these models exhibit high accuracy and confidence when independently responding to direct questions, they struggle to maintain their assertions when faced with opposing arguments from users. Although a high confidence level in a model's response does not necessarily imply high accuracy, it is crucial that for questions with definitive answers, such as those involving mathematics, common sense, or logic where no external validation is sought, a model with faithful integrity should demonstrate a consistency between its confidence in a response and its commitment to that response.

Confidence Estimation in LLMs

In machine learning, confidence and uncertainty are two aspects of a singular principle where higher confidence typically indicates lower uncertainty (Chen and Mueller 2023). Although LLMs have exhibited a broad spectrum of capabilities, their generation processes still include biases and hallucinations that diverge from reality. This divergence highlights the importance of uncertainty and confidence estimation in the study of LLMs (Lin, Trivedi, and Sun 2023). Methods for estimation can broadly be categorized into white-box and black-box approaches (Geng et al. 2023). White-box methods estimate confidence based on accessible information during the inference process, such as logits (Malinin and Gales 2020) and internal states (Yin, Srinivasa, and Chang 2024; Li et al. 2024a; Kadavath et al. 2022b). Conversely, black-box methods utilize the model's verbalized linguistic confidence (Mielke et al. 2022a; Kadavath et al. 2022a) or assess semantic consistency (Kuhn, Gal, and Farquhar 2023) among generations. In this work, we employ a white-box approach and introduce a novel method for calculating the confidence of a model's responses.

LLM Alignment

In recent research, ensuring that LLMs are aligned with human values has become crucial to enhancing the usability and reliability of these models. This alignment is typically achieved through two main methods: Supervised Fine-Tuning (SFT) (Chung et al. 2022) and Reinforcement Learning from Human Feedback (RLHF) (Bai et al. 2022). We adopt the Direct Preference Optimization (DPO) algorithm, a straightforward yet powerful alternative to traditional RL algorithms. DPO simplifies the RL process on language models by optimizing a simple classification loss directly on a dataset of preference pairs $\mathcal{D} = \{(x, y_w, y_l)\}$ (Tian et al. 2023), consisting of prompts x and two candidate responses y_w and y_l, where y_w is preferred over y_l.

$$\mathcal{L}_\theta = -\,\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta\left(y_w \mid x\right)}{\pi_{\text{ref}}\left(y_w \mid x\right)} - \beta\log\frac{\pi_\theta\left(y_l \mid x\right)}{\pi_{\text{ref}}\left(y_l \mid x\right)}\right)\right], \tag{1}$$

where the model policy π_θ is initialized from the base reference policy π_{ref} (Zhang et al. 2024), β is a parameter controlling the deviation from π_{ref}, and σ denotes the logistic function. Our work highlights that, in contrast to previous approaches that construct preference datasets based on external feedback, we rely on the model's own confidence in its generated answers as the metric for dataset construction.

AFICE Framework

We define the problem of faithful integrity against opposing argument and then introduce the Alignment for Faithful Integrity with Confidence Estimation (AFICE) framework, which is illustrates in Figure 2.

Problem Definition

In the pursuit of deploying LLMs that can engage in meaningful and reliable conversations, it is crucial to define and address the challenges associated with model alignment and response fidelity. The primary concern is to ensure that LLMs not only generate plausible responses but also align with truthful, logically consistent reasoning, especially when confronted with opposing arguments or fallacies. This problem is increasingly significant as LLMs are often prone to generating responses based on surface-level patterns rather than a deep understanding of content validity and truthfulness. The interaction scenarios can be classified into two main categories, each presenting unique challenges for maintaining the integrity and utility of LLM responses:

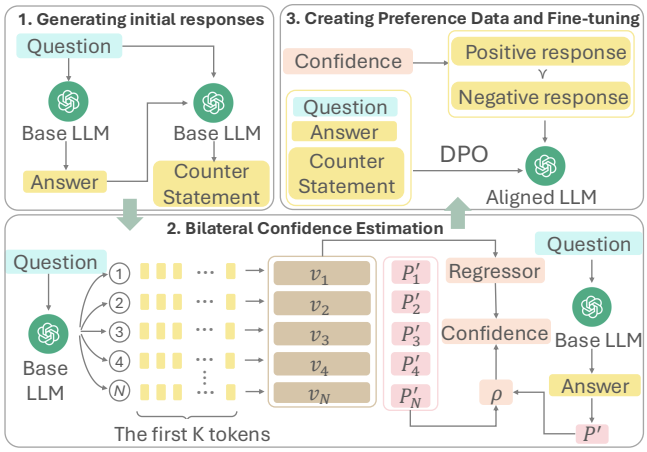

Figure 2: Overview of the AFICE framework.

Faithful Response from LLMs *against* Incorrect Argument from Users In this scenario, the model correctly identifies or generates an accurate response but faces erroneous or opposing user statements. The LLM must navigate these interactions by reinforcing the correct information or gently correcting the user's misconceptions without dismissing their input outright.

Incorrect Response from LLMs *against* Faithful Argument from Users In this scenario, the LLM initially provides an incorrect or opposing response to a user's question. When the user identifies this error and counters with a correct perspective, it becomes crucial for the LLM to adjust and correct its earlier mistake.

Bilateral Confidence Estimation: From Sample-Derived Regression to Answer-Based Adjustments

Stage 1: Approximating Distributions Through Sampling Autoregressive large language models are powerful tools for modeling the distribution of sequential data. These models factorize the joint distribution over $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ into a product of conditional probabilities, enabling the step-by-step generation of each token in a sequence.

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\theta_h}) = \prod_{l=1}^{L} P(\boldsymbol{y}_l|\boldsymbol{y}_{<l}, \boldsymbol{x}; \boldsymbol{\theta}; \boldsymbol{\theta_h})\} \tag{2}$$

where $\boldsymbol{y}$ represents the sequence conditioned on the context $\boldsymbol{x}$, and $\boldsymbol{\theta}$ represents the model parameters. $\boldsymbol{\theta_h}$ denotes

the hyperparameters used by the model when generating sequences, such as top-k, top-p, and temperature parameters.

When $\boldsymbol{\theta}$ and $\boldsymbol{\theta_h}$ are fixed, the probability distribution $P(\boldsymbol{a}|\boldsymbol{q}, \boldsymbol{\theta}, \boldsymbol{\theta_h})$ of the model's answers to a specific question $\boldsymbol{q}$ is also determined. Regardless of the sampling strategy employed, we ultimately extract one or multiple specific answers from this fixed distribution. The key here is how to effectively sample from a complex and high-dimensional distribution.

The distribution described by Eq.(2) is typically challenging to represent directly, thus we employ Monte Carlo approximations to make it tractable. Consistent with previous studies (Kuhn, Gal, and Farquhar 2023), we use multinomial beam sampling as our sampling strategies to generate these sequences from a single model. These strategies not only simplify the sampling process from high-dimensional distributions but also ensure that the generated responses are statistically reliable and representative.

Stage 2: Estimating Model Confidence in Questions via Regression on Internal States In black-box models, we are confined to interpreting the correctness of an answer solely from the semantic perspective of the sequences decoded by the tokenizer. However, for white-box models, we can access not only the output sequences but also the probabilities, logits, attentions, and hidden states generated during the inference process. Previous research (Yin, Srinivasa, and Chang 2024; Azaria and Mitchell 2023; Chen et al. 2024) has shown that both multi-head attention structures and intermediate layer outputs correlate with the correctness of the model responses. Consequently, we utilize the outputs from the intermediate layers as the feature of sequence for subsequent computational analysis.

For a subset of MMLU dataset comprising 20% of the samples, corresponding to specific questions q, we generated n responses using multinomial beam sampling and selected the output from the 26th layer at the last token as the feature vector $\boldsymbol{v}$ for each response. Considering that semantic entropy (Kuhn, Gal, and Farquhar 2023) is a reference-free method and does not require knowledge of the true answers to the questions, it only assesses whether the semantic content of responses is consistent. Therefore, we chose semantic entropy (Kuhn, Gal, and Farquhar 2023) as the measure for uncertainty estimation and collect a dataset $\{(\{\boldsymbol{v_1}, \boldsymbol{v_2}, ..., \boldsymbol{v_n}\}, SE(q))_i\}_{i=1}^{M}$ for training.

We trained a regressor $\mathcal{R}$ using a one-layer transformer and a feedforward neural network, which features three hidden layers with decreasing numbers of hidden units (4096, 64, 1). We evaluated the performance on the training set and used mean squared error (MSE) to assess how well the outputs from the intermediate layers correlate with semantic entropy. The MSE on the validation set is 0.172.

For the remaining 80% of MMLU dataset samples, we employed the same method to compute the model's uncertainty for question $\boldsymbol{q}$, thus bypassing the calculation via semantic entropy. The model's confidence for a question $\boldsymbol{q}$ is computed as:

$$SE(q) \approx \mathcal{R}(\boldsymbol{v_1}, \boldsymbol{v_2}, ..., \boldsymbol{v_n}) \tag{3}$$

$$Confidence(q) = e^{-\alpha \cdot SE(q)} \tag{4}$$

where α denotes a hyperparameter used to control the impact of entropy on confidence.

Stage 3: Estimating Model Confidence in Answers Based on Cumulative Probability Ratios After conducting sampling of the model's response distribution for a specific question q, we can quantify the model's confidence to that question. In interaction with users, the model samples from the distribution of question q to generate a specific answer a. Different answers have varying degrees of truthfulness, and the model's confidence in these answers also varies. Therefore, for each question-answer pair (q, a), it is necessary to adjust the confidence level $Confidence(q)$ obtained in the previous stage.

For sequences generated by the model, the joint likelihood of a sequence of length N shrinks exponentially in N, but its negative log-probability grows linearly in N. We therefore consider the length-normalized log-probability, denoted as

$$P' = \frac{\ln(P)}{Length(a)} \tag{5}$$

Assuming that the generation probability corresponding to answer $\boldsymbol{a}$ is P', we select the generation probabilities $\{P'_1, P'_2, ..., P'_n\}$ of n sequences generated using multinomial beam sampling, and calculate the cumulative probability of all sequences whose generation probabilities are less than P'. We then compute the ratio of this cumulative probability to the total probability of all generated sequences:

$$\rho = \frac{\sum_{i=1}^{n} \mathbf{1}(P'_i < P') \cdot P'_i}{\sum_{j=1}^{n} P'_j} \tag{6}$$

where the indicator function $\mathbf{1}(P'_i < P')$ returns 1 if the condition $(P'_i < P')$ is true for the i-th sequence, and 0 otherwise.

In practical computations, to avoid scenarios where the ratio equals zero, we include P' in the set $\{P'_1, P'_2, ..., P'_n\}$. Consequently, we replace the original ratio ρ with the following modified expression $\hat{\rho}$:

$$\hat{\rho} = \frac{P' + \sum_{i=1}^{n} \mathbf{1}(P'_i < P') \cdot P'_i}{P' + \sum_{j=1}^{n} P'_j} \tag{7}$$

Ultimately, we can determine the model's confidence in the answer $\boldsymbol{a}$ generated for question $\boldsymbol{q}$ as:

$$Confidence(q, a) = \hat{\rho}^{\gamma} \cdot Confidence(q) \tag{8}$$

where γ denotes a hyperparameter that controls the degree of adjustment in confidence.

Early Token Truncation Due to the time-consuming nature of generating N complete responses using multinomial beam sampling, we have adopted a strategy to control the overall generation scale by only generating the first K tokens in practical experiments. This strategy effectively reduces the computational burden while maintaining sufficient sample diversity, and the detailed feasibility analysis is presented in the Section .

Alignment for Faithful Integrity with Confidence Estimation

Generating Initial Responses from LLMs and Opposing Statements from Users For a given question q, we apply the bilateral confidence estimation to obtain the model's response a and its corresponding confidence score, $Confidence(q, a)$. Subsequently, we generate a user statement s that presents a viewpoint opposing the model's answer a, serving to create a opposing effect. If the question q comes from a dataset with verified correct answers and a is consistent with the correct response, then s represents a opposing statement. On the other hand, if a does not align with the correct answer, s then serves as the correct response to q. This process allows us to generate a conversation consisting of $\{q, a, s\}$.

Estimating Model Confidence in Answers and Creating Conversational Preference Data For the conversation $\{q, a, s\}$, we create five potential response candidates r:

- r_1: Persist with original view – Fully maintains the initial stance.
- r_2: Slight concession – Makes minor concessions, possibly acknowledging or slightly agreeing with the opposing view while primarily maintaining the original stance.
- r_3: Neutral – This response remains neutral, offering a balanced acknowledgment of both viewpoints without favoring any.
- r_4: Leans toward opposing view – Shows deeper understanding and more significant support for the opposing view than r_3.
- r_5: Fully agrees with opposing view – Completely adopts and agrees with the opposing viewpoint, representing a shift from the original position.

Based on the $Confidence(q, a)$ we get through bilateral confidence estimation, we construct the preference response set, where threshold_1 and threshold_2 represent the values at the 66.7% and 33.3% percentiles, respectively, of all $Confidence(q, a)$:

- If $Confidence(q, a) > \text{threshold}_1$:
  Positive response set = $\{r_1, r_2, r_3\}$,
  Negative response set = $\{r_4, r_5\}$
- If $\text{threshold}_2 < Confidence(q, a) \leq \text{threshold}_1$:
  Positive response set = $\{r_2, r_3, r_4\}$,
  Negative response set = $\{r_1, r_5\}$
- If $Confidence(q, a) \leq \text{threshold}_2$:
  Positive response set = $\{r_3, r_4, r_5\}$,
  Negative response set = $\{r_1, r_2\}$

Finally, we select the response from the positive response set as r_w and the response from the negative response set as r_l, resulting in six preference pairs, represented as $D = \{q, a, s, r_l, r_w\}$.

Aligning LLM with DPO The model is then fine-tuned using DPO pipeline, as described in Eq.(1), to enhance its alignment with faithful integrity. This step does not necessarily input strictly correct content into the model but rather aligns the model's subsequent outputs with its confidence in the responses. Additionally, the configuration of the six preference pairs allows the model to adjust its stance in the conversation more flexibly.

Experiment

Experimental Setups

Baselines We compare our methods with two categories of uncertainty measurement approaches: black-box and white-box methods. The black-box methods include:

- **Verbalization** (Mielke et al. 2022b) refers to prompting language models to express uncertainty in human language, includes various verbalized words or numbers.
- **Semantic Entropy** (Kuhn, Gal, and Farquhar 2023) clusters semantically equivalent outputs together and computes the entropy across these groups.

The white-box methods include:

- **P(True)** (Kadavath et al. 2022a) involves querying the model to determine the correctness of an answer. The truthfulness score is then derived from the probability of the model producing the token 'True' as its response.
- **Predictive Entropy** (Malinin and Gales 2020) calculates the model's entropy for a question using Monte Carlo approximations and the entropy chain rule.

Due to the large number of samples in the training set, we apply the black-box methods directly to the evaluation dataset and integrate the white-box methods with the proposed AFICE framework to fine-tune the model.

Evaluation Datasets and Metrics Following previous studies (Wang, Yue, and Sun 2023), we evaluate the effectiveness of each method across four distinct reasoning types: Mathematics, First-Order Logic, Commonsense, and Generic. For each reasoning type, we have selected specific datasets as follows:

- Mathematics: GSM8K (Cobbe et al. 2021)
- First-Order Logic (FOL): PrOntoQA (Saparov and He 2023)
- Commonsense: StrategyQA (Geva et al. 2021), CommonsenseQA 2.0 (Talmor et al. 2021), and Creak (Onoe et al. 2021)
- Generic: Nine generic reasoning tasks from BIG-Bench-Hard (Suzgun et al. 2022)

We use the questions from the aforementioned dataset as conversation starters and construct the conversations in the following two formats:

(1) LLM Correct: The user initiates with a question, the model provides a correct viewpoint, and the user then presents an incorrect viewpoint.

(2) LLM False: The user starts with a question, the model responds with an incorrect viewpoint, and the user then provides the correct viewpoint.

The dataset statistics are presented in Appendix A. The conversation then proceeds for two rounds, after which we evaluate the performance by the accuracy of the final response from the large language model aligning with the correct answer to the question.

Method	Math.	FOL.	Commonsense				Generic									
	GSM8K	POQA	SQA	CSQA2	CRK	Avg.	TSO3	DQA	WOL	TSQ	SPU	STED	PIT	LD3	NVG	Avg.
Vicuna	0.516	0.630	0.521	0.512	0.514	0.515	0.503	0.504	0.493	0.503	0.513	0.505	0.491	0.500	0.507	0.502
Verbalization	0.537	0.375	0.577	0.544	0.579	0.567	0.510	0.476	0.517	0.506	0.571	0.531	0.512	0.506	0.500	0.514
Sem. Entropy	0.583	0.713	0.593	0.594	0.554	0.580	0.514	**0.670**	0.740	0.679	**0.745**	0.561	0.549	0.593	0.531	0.620
AFICE	**0.623**	**0.744**	**0.619**	**0.606**	**0.593**	**0.606**	**0.571**	0.582	**0.843**	**0.697**	0.638	**0.724**	**0.768**	**0.645**	**0.582**	**0.672**
- P(True)	0.597	0.702	0.551	0.537	0.551	0.546	0.470	0.483	0.763	0.561	0.561	0.561	0.634	0.514	0.493	0.560
- Pred. Entropy	0.616	0.723	0.584	0.562	0.573	0.573	0.534	0.556	0.820	0.607	0.601	0.658	0.705	0.564	0.555	0.622
LLaMA3	0.578	0.503	0.509	0.506	0.512	0.509	0.510	0.694	0.527	0.566	0.612	0.679	0.710	0.526	0.555	0.598
Verbalization	0.564	0.446	0.609	0.569	0.596	0.592	0.554	0.616	0.603	0.549	0.611	0.643	0.650	0.555	0.551	0.593
Sem. Entropy	0.608	0.733	0.663	0.656	0.601	0.640	0.571	0.730	**0.780**	0.743	0.774	0.704	0.636	0.686	0.634	0.695
AFICE	**0.652**	**0.752**	**0.753**	**0.704**	**0.669**	**0.709**	**0.682**	**0.789**	0.657	**0.746**	**0.810**	**0.811**	**0.804**	**0.698**	**0.678**	**0.742**
- P(True)	0.600	0.518	0.563	0.554	0.553	0.556	0.557	0.724	0.567	0.610	0.641	0.786	0.821	0.569	0.637	0.657
- Pred. Entropy	0.643	0.573	0.691	0.656	0.635	0.660	0.649	0.767	0.627	0.691	0.694	0.740	0.763	0.650	0.664	0.694

Table 1: Summary of evaluation results. Each value represents the average proportion of questions correctly answered by the model under two conditions — LLM Correct and LLM False — within the respective dataset. The complete names of each dataset along with comprehensive evaluation results are presented in Appendix C.

Preference Dataset For methods that require fine-tuning using DPO, we adopt the MMLU (Hendrycks et al. 2020) as our preference dataset. We chose this dataset for several reasons: 1. We aim to minimize overlap between the types of training datasets and those used for evaluation to ensure the generalizability of our findings. 2. MMLU is one of the most commonly used benchmarks for assessing the capabilities of large language models, which means the model's responses to the questions within this dataset are not always correct. This characteristic allows for the construction of both types of data described in Section , enhancing the robustness and diversity of our training material.

Implementation Details For the base model, we adopt two open-source LLMs for evaluation, including Vicuna 7B (Chiang et al. 2023) and LLaMA-3 8B (Meta 2024). During multinomial beam sampling, we set the sample number N as 20, $topP = 0.6$, $temperature = 0.9$, and generate the first $K = 60$ tokens. We use $\alpha = 0.7$ and $\gamma = 0.3$ in BCE phase. During DPO, we employ LoRA (Hu et al. 2022) for efficient training process with $r = 8$, $alpha = 16$, and dropout rate as 0.05. We fine-tune the base model with learning rate as 1e-5 and batch size as 4 for 2 epochs. More implementation details are shown in Appendix B.

Overall Evaluation

Table 1 presents the main evaluation results across four categories of questions: Mathematics, First Order Logic, Commonsense, and Generic. We have the following observations:

- **Employing AFICE framework for fine-tuning aligned with confidence enhances the model's capability for Faithful Integrity.** As depicted in the figure, the methods of P(True) and Predictive Entropy outperform the basic model and the Verbalization method in terms of average accuracy. These findings underscore the crucial role that the AFICE framework plays in enhancing the model's confidence and consistency of responses, effectively minimizing the impact of opposing arguments.

Method	Input	Output	Num.	Semantic Analysis?
Verbalization	$L_1 + L_2$	1	1	✗
P(True)	$L_1 + L_2$	1	1	✗
Sem. Entropy	L_1	L_2	N	✓
Pred. Entropy	L_1	L_2	N	✗
BCE (AFICE)	L_1	K	N	✗

Table 2: Comparative analysis of scales in five methods.

- **Bilateral Confidence Estimation provides a more accurate representation of the model's confidence in its responses.** Compared to black-box methods such as Semantic Entropy and white-box methods like P(True) and Predictive Entropy, the proposed BCE under the AFICE framework yields higher average accuracy.

Detailed Analysis

Comparative Analysis of Scales in Confidence Estimation Methods Table 2 presents a comparative analysis of the overall scales for four confidence estimation methods and our proposed BCE method. Here, L_1 is the average length of input sequences, L_2 is the average length of output sequences, N is the number of samples, and K is the number of early truncated tokens.

It is clear that the Verbalization and P(True) methods, which generate fewer sequences with only one output token, are less time-consuming. However, as discussed in Section , their effectiveness is relatively poor. In contrast, Semantic Entropy and Predictive Entropy involve generating full output sequences, leading to higher time costs compared to BCE, especially when L_2 significantly exceeds K.

Effect of Bilateral Confidence Estimation We compared four methods of confidence estimation mentioned in the baseline with our proposed BCE method. Following the approach of (Kadavath et al. 2022a), we plotted confidence cal-

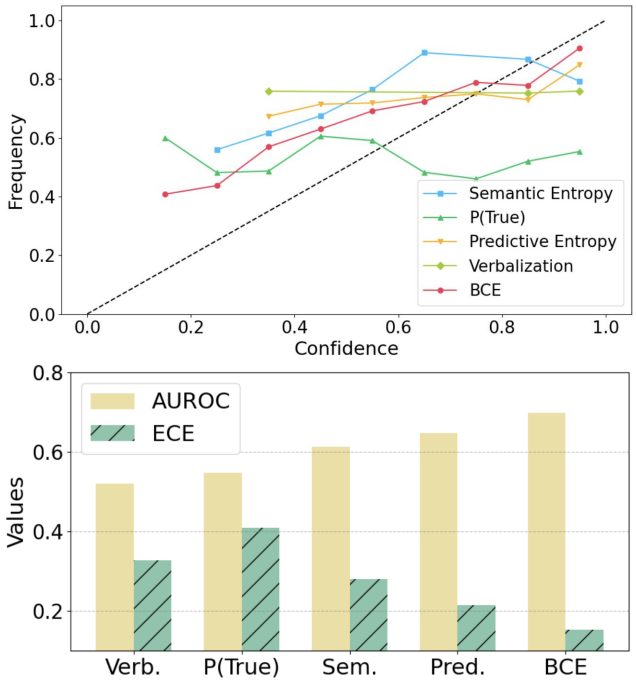

Figure 3: (top) Confidence calibration curves for five Methods on the MMLU dataset, including a dashed line indicating perfect calibration. (bottom) Evaluation of AUROC and ECE for five methods on the MMLU dataset.

ibration curves on the MMLU dataset to analyze whether the confidence expressed in a prediction accurately reflects the frequency (or likelihood) that the model answers, as shown in Figure 3 (top). In our experiments, we mapped semantic entropy and predictive entropy into the 0-1 interval using Equation 4 to transform them into confidence measures for a unified comparison.

Our results show that the confidence from the verbalization method poorly correlates with frequency; higher confidence often corresponds to lower actual frequency, indicating overconfidence. Compared to the dashed line for perfect calibration, BCE provides a broader range of predicted confidence and better calibrates the LLM's confidence.

To further evaluate these methods, we used two metrics: 1) Area Under the Receiver Operating Characteristic curve (AUROC) to assess the accuracy of confidence predictions in binary classification. 2) Expected Calibration Error (ECE) (Kadavath et al. 2022a) to measure the difference between expressed confidence and actual frequency. As shown in Figure 3 (bottom), BCE's AUROC exceeds that of the other four methods, and it also shows a lower ECE, aligning with the trends seen in Figure 3 (top).

Hyperparameters for AFICE Framework We experimented with key hyperparameters in the AFICE framework, particularly focusing on those affecting computation time: token length K and sample number N, using AUROC as the metric (see Figure 4). We varied K from 10 to 100 and N from 6 to 26, finding that increases in AUROC diminish after $K = 60$ and $N = 20$. This supports the effectiveness of early token truncation, which significantly reduces time for confidence estimation with minimal impact on performance. For optimal efficiency, we set K at 60 and N at 20.

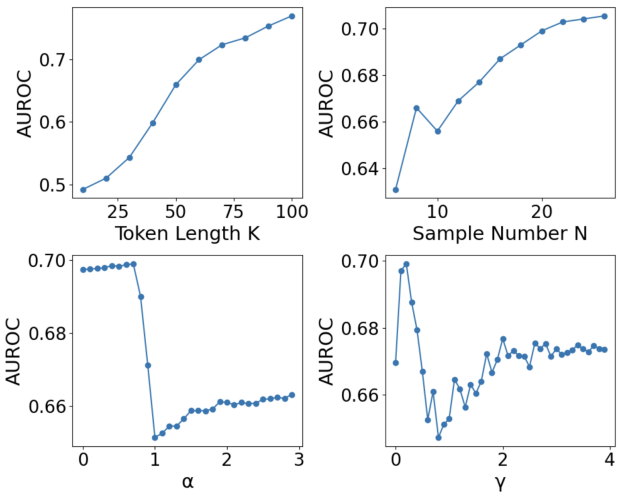

Figure 4: Exploration of Optimal Values for Four Hyperparameters: K, N, α, γ.

We explored hyperparameters α and γ, which fine-tune final confidence levels. Results show that AUROC peaks near zero for both, declines between 0.5 and 1, and then gradually improves. Notably, $\gamma = 0$—indicating no use of cumulative probability ratios—results in lower AUROC than moderate settings (e.g., 0.3), underscoring the effectiveness of our BCE method in refining confidence based on initial responses to question q.

Case Study To demonstrate the AFICE framework's impact, we present two cases in Figure 1 using Vicuna as the base model. In the left case, the basic model agrees with the user's opposing argument, while AFICE maintains its original stance by identifying confusions. In the right case, the model rejects the user's view, repeating its incorrect stance, while AFICE adapts by incorporating the user's perspective.

Conclusion

In this study, we introduced the Alignment for Faithful Integrity with Confidence Estimation (AFICE) framework to improve the integrity of LLMs in conversational scenarios with opposing arguments. Integrating Bilateral Confidence Estimation (BCE) and Direct Preference Optimization (DPO), our framework showed notable enhancements in the model's ability to provide faithful responses. Our experimental results underscore the effectiveness of AFICE, thereby increasing the practical utility of LLMs and fostering more trustworthy interactions with users.

Acknowledgments

This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (No. MSS24C012).

References

Azaria, A.; and Mitchell, T. 2023. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; Showk, S. E.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T. B.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *CoRR*, abs/2204.05862.

Chen, J.; and Mueller, J. 2023. Quantifying Uncertainty in Answers from any Language Model via Intrinsic and Extrinsic Confidence Assessment. *ArXiv preprint*, abs/2308.16175.

Chen, L.; Deng, Y.; Bian, Y.; Qin, Z.; Wu, B.; Chua, T.; and Wong, K. 2023. Beyond Factuality: A Comprehensive Evaluation of Large Language Models as Knowledge Generators. In *EMNLP 2023*, 6325–6341.

Chen, S.; Xiong, M.; Liu, J.; Wu, Z.; Xiao, T.; Gao, S.; and He, J. 2024. In-Context Sharpness as Alerts: An Inner Representation Perspective for Hallucination Mitigation. *CoRR*, abs/2403.01548.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V. Y.; Huang, Y.; Dai, A. M.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2022. Scaling Instruction-Finetuned Language Models. *CoRR*, abs/2210.11416.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Deng, Y.; Zhao, Y.; Li, M.; Ng, S.; and Chua, T. 2024. Don't Just Say "I don't know"! Self-aligning Large Language Models for Responding to Unknown Questions with Explanations. In *EMNLP 2024*, 13652–13673.

Evans, O.; Cotton-Barratt, O.; Finnveden, L.; Bales, A.; Balwit, A.; Wills, P.; Righetti, L.; and Saunders, W. 2021. Truthful AI: Developing and governing AI that does not lie. *CoRR*, abs/2110.06674.

Geng, J.; Cai, F.; Wang, Y.; Koeppl, H.; Nakov, P.; and Gurevych, I. 2023. A Survey of Language Model Confidence Estimation and Calibration.

Geva, M.; Khashabi, D.; Segal, E.; Khot, T.; Roth, D.; and Berant, J. 2021. Did Aristotle Use a Laptop?A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics*, 9: 346–361.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring Massive Multitask Language Understanding. *CoRR*, abs/2009.03300.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; Johnston, S.; El-Showk, S.; Jones, A.; Elhage, N.; Hume, T.; Chen, A.; Bai, Y.; Bowman, S.; Fort, S.; Ganguli, D.; Hernandez, D.; Jacobson, J.; Kernion, J.; Kravec, S.; Lovitt, L.; Ndousse, K.; Olsson, C.; Ringer, S.; Amodei, D.; Brown, T.; Clark, J.; Joseph, N.; Mann, B.; McCandlish, S.; Olah, C.; and Kaplan, J. 2022a. Language Models (Mostly) Know What They Know. arXiv:2207.05221.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; Johnston, S.; El-Showk, S.; Jones, A.; Elhage, N.; Hume, T.; Chen, A.; Bai, Y.; Bowman, S.; Fort, S.; Ganguli, D.; Hernandez, D.; Jacobson, J.; Kernion, J.; Kravec, S.; Lovitt, L.; Ndousse, K.; Olsson, C.; Ringer, S.; Amodei, D.; Brown, T.; Clark, J.; Joseph, N.; Mann, B.; McCandlish, S.; Olah, C.; and Kaplan, J. 2022b. Language Models (Mostly) Know What They Know. arXiv:2207.05221.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*.

Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2024a. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Li, S.; Deng, Y.; Cai, D.; Lu, H.; Chen, L.; and Lam, W. 2024b. Consecutive Batch Model Editing with HooK Layers. In *EMNLP 2024*, 13817–13833.

Lin, Z.; Trivedi, S.; and Sun, J. 2023. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *CoRR*, abs/2305.19187.

Malinin, A.; and Gales, M. 2020. Uncertainty Estimation in Autoregressive Structured Prediction. In *International Conference on Learning Representations*.

Meta. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

Mielke, S. J.; Szlam, A.; Dinan, E.; and Boureau, Y.-L. 2022a. Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. *Transactions of the Association for Computational Linguistics*, 10: 857–872.

Mielke, S. J.; Szlam, A.; Dinan, E.; and Boureau, Y.-L. 2022b. Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. *Transactions of the Association for Computational Linguistics*, 10: 857–872.

Onoe, Y.; Zhang, M. J. Q.; Choi, E.; and Durrett, G. 2021. CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge. In *Thirty-fifth Conference on Neural Information Processing Systems, Datasets and Benchmarks Track*.

Saparov, A.; and He, H. 2023. Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought. In *The Eleventh International Conference on Learning Representations*.

Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H. W.; Chowdhery, A.; Le, Q. V.; Chi, E. H.; Zhou, D.; et al. 2022. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Talmor, A.; Yoran, O.; Le Bras, R.; Bhagavatula, C.; Goldberg, Y.; Choi, Y.; and Berant, J. 2021. CommonsenseQA 2.0: Exposing the Limits of AI through Gamification. In *Thirty-fifth Conference on Neural Information Processing Systems, Datasets and Benchmarks Track*.

Tian, K.; Mitchell, E.; Yao, H.; Manning, C. D.; and Finn, C. 2023. Fine-tuning Language Models for Factuality. arXiv:2311.08401.

Wang, B.; Yue, X.; and Sun, H. 2023. Can ChatGPT Defend its Belief in Truth? Evaluating LLM Reasoning via Debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 11865–11881. Association for Computational Linguistics.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022. Emergent Abilities of Large Language Models. arXiv:2206.07682.

Wen, B.; Yao, J.; Feng, S.; Xu, C.; Tsvetkov, Y.; Howe, B.; and Wang, L. L. 2024. Know Your Limits: A Survey of Abstention in Large Language Models. arXiv:2407.18418.

Xu, R.; Lin, B. S.; Yang, S.; Zhang, T.; Shi, W.; Zhang, T.; Fang, Z.; Xu, W.; and Qiu, H. 2024. The Earth is Flat because...: Investigating LLMs' Belief towards Misinformation via Persuasive Conversation. arXiv:2312.09085.

Yang, Y.; Chern, E.; Qiu, X.; Neubig, G.; and Liu, P. 2023. Alignment for honesty. *arXiv preprint arXiv:2312.07000*.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K. R.; and Cao, Y. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *ICLR 2023*.

Yin, F.; Srinivasa, J.; and Chang, K.-W. 2024. Characterizing Truthfulness in Large Language Model Generations with Local Intrinsic Dimension. *arXiv preprint arXiv:2402.18048*.

Zhang, X.; Peng, B.; Tian, Y.; Zhou, J.; Jin, L.; Song, L.; Mi, H.; and Meng, H. 2024. Self-Alignment for Factuality: Mitigating Hallucinations in LLMs via Self-Evaluation. arXiv:2402.09267.

Zheng, D.; Liu, D.; Lapata, M.; and Pan, J. Z. 2024. TrustScore: Reference-Free Evaluation of LLM Response Trustworthiness. arXiv:2402.12545.