Uncertainty Aware Learning for Language Model Alignment

Yikun Wang,[*†] **Rui Zheng,**[†*] **Liang Ding,**[‡§] **Qi Zhang,**[†] **Dahua Lin,**[¶‖] **Dacheng Tao**[**]

{yikunwang19, rzheng20, qz}@fudan.edu.cn, {dhlin}@ie.cuhk.edu.hk
{liangding.liam, dacheng.tao}@gmail.com

Abstract

As instruction-tuned large language models (LLMs) evolve, aligning pretrained foundation models presents increasing challenges. Existing alignment strategies, which typically leverage diverse and high-quality data sources, often overlook the intrinsic uncertainty of tasks, learning all data samples equally. This may lead to suboptimal data efficiency and model performance. In response, we propose uncertainty-aware learning (UAL) to improve the model alignment of different task scenarios, by introducing the sample uncertainty (elicited from more capable LLMs). We implement UAL by a simple fashion – adaptively setting the label smoothing value of training according to the uncertainty of individual samples. Analysis shows that our UAL indeed facilitates better token clustering in the feature space, validating our hypothesis. Extensive experiments on widely used benchmarks demonstrate that our UAL significantly and consistently outperforms standard supervised fine-tuning. Notably, LLMs aligned in a mixed scenario have achieved an average improvement of 10.62% on high-entropy tasks (i.e., AlpacaEval leaderboard), and 1.81% on complex low-entropy tasks (i.e., MetaMath and GSM8K).

1 Introduction

Large language models (LLMs), represented by GPT-4 (OpenAI, 2023), Claude, and Llama2 (Team, 2023), have recently achieved significant success in a series of natural language generation and understanding tasks (Qin et al., 2023; Zhong et al., 2023a; Peng et al., 2023b; Lu et al., 2023). The emergence of alignment methods has

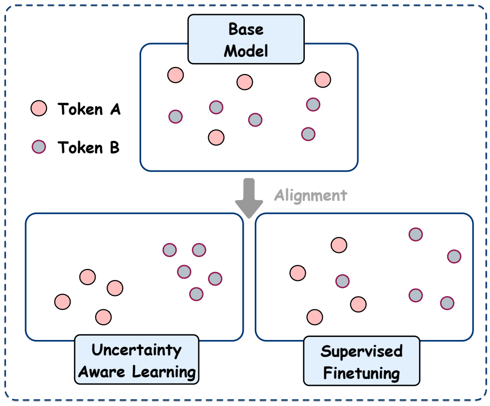

Figure 1: **Illustration of feature clustering.** Compared to SFT, UAL-based models show more convergence in the feature space, which we detailed our exploration in Section 4.

further enhanced the capabilities of LLMs, for example, the ability to follow human instructions or to achieve better zero-shot performance. The current popular alignment approaches including RLHF (Li et al., 2023; Zheng et al., 2023), consider supervised fine-tuning (SFT) an essential part that contributes significantly.

It is undeniable that the current SFT paradigm achieves considerable success. Some previous studies find that high-quality, complex, and diverse data contribute significantly to alignment (Zhou et al., 2023; Liu et al., 2023). However, the data can exhibit different levels of uncertainty. For data points in highly technical, scientific, or specialized settings, context may have little or no ambiguity and a limited set of correct answers. In contrast, other dialogues might feature varied and dynamic social contexts with idiomatic language uses. Appendix A presents a pair of such examples. Nevertheless, the common SFT paradigm applies the same level of supervision to all samples in the training set, overlooking the intrinsic uncertainty of the data.

Furthermore, due to the phenomenon of catastrophic forgetting (Luo et al., 2023; Ramasesh et al.,

[*]Yikun and Rui contributed equally.
[†]School of Computer Science, Fudan University, China.
[‡]The University of Sydney, Austrilia.
[§] Corresponding author.
[¶]The Chinese University of Hong Kong, Hong Kong.
[‖]Shanghai Artificial Intelligence Laboratory, China.
[**]Nanyang Technological University, Singapore.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11087–11099
August 11-16, 2024

2022; Aleixo et al., 2023), SFT-aligned models may perform worse than their foundational models. The Deepseek technical report reveals that the SFT model often underperforms compared to its base model on several benchmarks (Team, 2024), while LLMs generally show diminished performance on general tasks when forging their agent capability (Zeng et al., 2023). In our experiments, we observe that the standard SFT paradigm frequently leads to model degradation, resulting in decreased performance on some benchmarks, despite providing overall improvement. For instance, this is evident in parts of the commonsense benchmarks, such as MMLU. It is essential to align pretrained models while mitigating degradation issues as effectively as possible. Therefore, we propose:

> **Uncertainty Hypothesis.** *In an ideal paradigm, to further enhance alignment performance, the model should attend to samples differently based on their properties during alignment. Specifically, the model should impose stricter constraints when attending to more certain examples, as these samples exhibit less uncertainty and fewer variations, while maintaining relaxed constraints for highly uncertain examples.*

To design our uncertainty-aware learning (UAL) method and address potential model degradation caused by SFT, we propose measuring uncertainty through a coarsely-grained approach that incorporates an autonomous judge (e.g., GPT-4) to assess the uncertainty of each sample in the training set. Upon obtaining the uncertainty estimations, we linearly map them into our adaptive label smoothing training pipeline. Further details of our algorithm are presented in Section 2.

UAL significantly enhances the performance of instruction-tuned models. We apply this method across various model architectures and alignment datasets, observing that the UAL paradigm consistently outperforms the vanilla SFT paradigm on prominent benchmarks, including MMLU and TruthfulQA (Lin et al., 2022), among others. As Figure 2 shows, UAL also helps the model improve in different scenarios. Moreover, compared to its foundational model, as Figure 1 presents, the SFT-aligned model brings same-class tokens closer together in the feature space, and UAL enhances this tendency even further. This strongly correlates with

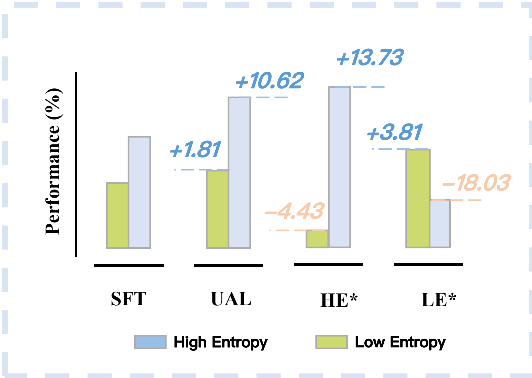

Figure 2: Mistral-7B (Jiang et al., 2023) aligned with UAL on the mixed scenario dataset has **an improvement of 10.62% on the high entropy scenario and 1.81% on the low entropy scenario**, compared with SFT. Evaluation results from models solely SFT aligned with high-entropy data (i.e., **HE** aligned with ShareGPT-6k) or low-entropy data (i.e., **LE** aligned with MetaMath-6k) are also provided for reference.

the superior performance of UAL-aligned models across benchmarks and scenarios, offering an important explanation for UAL's outperformance over conventional SFT.

Contributions. Our contributions are summarized threefold:

- Due to the limitations of the current supervised finetuning (SFT) paradigm, we propose the uncertainty-aware learning (UAL) approach to mitigate the alignment degradation problem and improve model performance in high-entropy and low-entropy scenarios.
- Based on an intuitive design concept, our UAL paradigm is simple to implement, making it possible to improve any language model's alignment.
- We conducted extensive experiments to prove the efficiency of our method and conducted in-depth analyses to provide some insights (i.e. feature clustering) into the mechanism.

2 Uncertainty Approach for Characterizing Data

In this section, we present a insightful study of the characteristics of data uncertainty property in instruction tuning. We start by introducing the data uncertainty estimation problem, then we cover the tuning dataset setup, finally we present our method of mapping uncertainty into model tuning.

Dataset	Source	Size
LIMA	-	1030
Deita	ShareGPT, UltraGPT	6k
ShareGPT-6k	ShareGPT	6k
MetaMath-6k	MetaMathQA	6k
Mixed-6k	ShareGPT, MetaMathQA	6k

Table 1: **Statistics of the alignment datasets** employed in the empirical study, LIMA and Deita are from (Zhou et al., 2023) and (Liu et al., 2023) respectively, and the others are sampled and synthesized by employing the data selection technique DEITA (Liu et al., 2023).

2.1 Uncertainty Estimation

For further enhance the model performance in the process of alignment, we propose to impose variant levels of loss constraint for different data based on their uncertainty, which has a similar intrinsic to information entropy. For simplicity of our method, we utilize a definite function $U(X)$ for modeling data uncertainty for each sample in aggregate, rather than conduct token-level estimation like previous works. More specifically speaking, for a tuning dataset $X = \{x_1, x_2, ..., x_n\}$, where each x_i represents an individual data point in the format of instruction-response. Our goal is to collect uncertainty $U(x_i)$ of each sample x_i.

Some researchers have found that LLMs could express uncertainty accurately (Xiong et al., 2023), referred to as confidence elicitation, which is essential in ensuring reliable and trustworthy decision-making processes. In our following empirical study, we will employ a more capable model (w.r.t. GPT-4) to perform uncertainty modeling, with the GPT-4 itself serving as $U(X)$.

2.2 Tuning Dataset Setup

The large language model research community has developed a consensus: a small amount of human-collected and high-quality data such as LIMA containing around 1k examples, can already achieve satisfying alignment performance (Zhou et al., 2023); some researchers have even demonstrated that a limited size of high-quality, complex and diverse dataset could reach or even surpass the effectiveness of large scale datasets such as Alpaca-52k (Liu et al., 2023). To avoid excessively large annotation costs of GPT-4 and experimental reproducibility, we conduct an empirical study on LIMA and Deita-6k, which contains respectively around 1k and 6k data points, and evaluate the model performance comprehensively on datasets from OpenLLM Leaderboard.

To further demonstrate that our method could simultaneously improve the model's capability in both high-entropy and low-entropy scenarios, and to prove the superiority of our method compared to SFT, we will sample several following datasets from different scenarios: **High Entropy**: we employ the DEITA method to sample high-quality, complex and diverse examples from ShareGPT, which mainly contains various kinds of dialogues between users and GPT, resulting a high-quality and high-entropy dataset ShareGPT-6k containing 6k examples; **Low Entropy**: we employ DEITA (Liu et al., 2023) to sample from MetaMathQA (Yu et al., 2023) with 395k examples, resulting in a high-quality math question and answer datasets with 6k examples; **Mixed Scenario**: we aim to compare compare the superiority of our method in a comprehensive setting against standard SFT, and compare it with both the high-entropy and the low-entropy models that respectively tuned with the high-entropy dataset and low-entropy dataset, while ensuring a fair comparison. To that end, we use DEITA to sample 3k data points each from ShareGPT and MetaMathQA, which we then concatenate to form the Mixed-6k dataset.

2.3 Uncertainty Mapping

Label smoothing is a common technique for mitigating overfitting and enhancing generalization ability in the field of deep learning, by combing the cross-entropy loss, the constraint intensity could be adjusted by altering the smooth value, a positive value within the interval from zero to one. Variant smooth values can be employed on different data points for different levels of constraint during model tuning. To satisfy our uncertainty hypothesis, ensuring low-uncertainty samples are subject to stronger constraints and meanwhile high-uncertainty samples are allowed looser constraints, the data uncertainty values U are mapped with a function $\mathcal{F} : U \to V$ to label smooth values V, every sample has a unique smooth value $v_i = \mathcal{F}(u_i)$ during training.

$$\mathcal{F}(x_i) = \min(h(u_i), v_t) \quad (1)$$

$$v_i = \mathcal{F}(x_i) \quad \text{s.t. } \mathcal{E}(v_i) = \alpha \quad (2)$$

Assume a linearly scaling function $h : U \to V$ with the form of $h(u) = \beta u$, where β is a scaling factor. To reflect our uncertainty hypothesis and the simplicity of our approach, we directly employ the truncated linear mapping $\mathcal{F}(U)$ to transform

Model	Data Size / Alignment	ARC-c	HellaSwag	MMLU	TruthfulQA	Average
			Llama-2 based Models			
Llama-2-7B	- / -	63.21	75.12	60.61	53.37	60.08
Vicuna-7B-v1.5	125K / SFT	71.23	89.32	67.40	52.43	70.10
Alpaca-2-7B	52K / SFT	64.54	87.04	63.68	46.26	65.38
LIMA-7B	1K / SFT	55.51	<u>79.61</u>	60.42	<u>64.01</u>	<u>64.89</u>
LIMA-7B (**Ours**)	1K / UAL	**58.89**	**<u>79.87</u>**	**<u>65.70</u>**	**<u>66.16</u>**	**<u>67.66</u>**
Deita-7B	6K / SFT	<u>67.22</u>	74.24	<u>64.40</u>	<u>57.77</u>	<u>65.91</u>
Deita-7B (Ours$^-$)	6k / UAL	<u>66.88</u>	75.01	<u>63.81</u>	<u>60.18</u>	<u>66.47</u>
Deita-7B (**Ours**)	6K / UAL	**<u>69.55</u>**	**<u>76.77</u>**	**<u>67.64</u>**	**<u>64.06</u>**	**<u>69.51</u>**
			Mistral based Models			
Mistral-7B	- / -	77.25	75.63	68.97	33.78	63.91
Deita-7B	6K / SFT	71.57	71.73	62.11	**<u>38.88</u>**	61.07
Deita-7B (Ours$^-$)	6k / UAL	**<u>78.92</u>**	<u>76.86</u>	66.08	33.29	<u>63.79</u>
Deita-7B (**Ours**)	6K / UAL	<u>78.26</u>	**<u>78.93</u>**	**66.16**	<u>36.23</u>	**<u>64.90</u>**

Table 2: The **Open LLM Leaderboard benchmark evaluation results** of Llama-2 (Team, 2023) and Mistral-based (Jiang et al., 2023) models employing inference **(Best-of-4, zero-shot)** strategy. **Bold** indicates the models performing the best in its counterparts from the same foundation model and tuned on the same instruction-response dataset, with <u>Underline</u> indicates the models better than its corresponding foundation models. $^-$ indicates the models aligned with **the PPL uncertainty**.

the uncertainty of an instruction tuning example u_i into a label smoothing value v_i. Truncation means that the label smoothing value cannot exceed 1, hence The form of $\mathcal{F}$ is indicated in equation 1, where v_t represents the maximum permissible label smoothing value, which is set to 0.99 in subsequent experiments. To regulate the overall level of constraints during training, the average smoothing value across all samples is set to a fixed value α (e.g. a common value of 0.1).

Approach Outline. *In the first phase of the two-stage method, uncertainty estimation values are elicited from a capable GPT-4 model, followed by a truncated linear mapping from uncertainty estimation to sample-wise label smooth values in order to fulfill our* ***uncertainty hypothesis****.*

3 Experiments

In this section, we will introduce the experimental setup and present evaluation results from comprehensive benchmarks, as well as improvements observed in mixed scenarios, to demonstrate the superiority of our method. To promote a deeper understanding of our approach, we conduct an ablation study. Additionally, we present a case study and employ GPT-4 to further evaluate the quality of responses generated by our method.

3.1 Experimental Setup

For uncertainty estimation, we utilize GPT-4 for the collection of aligned datasets to estimate uncertainty, with the corresponding prompts displayed in the Appendix C. Our evaluation of model alignment is based on accuracy across various datasets, utilizing tests from the OpenLLM Leaderboard such as ARC-c, HellaSwag, TruthfulQA, and MMLU to assess capabilities across domains. In order to evaluate performance in both high-entropy and low-entropy scenarios, we employ AlpacaEval (Dubois et al., 2023) for general dialogue and GSM8k / MetaMath for low-entropy contexts. For datasets on the OpenLLM Leaderboard, we have also **implemented the Best-of-N strategy to enlargen the performance discrepancies** among different models. The training settings are detailed in the Appendix D.

3.2 Evaluation on Comprehensive Benchmarks

We instruction-tune Llama-2-7B respectively on the LIMA and the Deita dataset, in order to demonstrate the superiority of the uncertainty-aware learning (UAL) compared to normal SFT, as well as benchmark them against other strong open-sourced instruction-tuned models including Alpaca-2-7B and Vicuna-7B-v1.5. Among the three model-instruction tuning dataset combinations, models based on UAL demonstrate clear overall advantages over the SFT counterpart. Specifically for Llama-2-7B, model degradation due to SFT (w.r.t. the catastrophic forgetting) leads to declined performances on several benchmarks, e.g. LIMA-7B (Llama-2-7B, SFT) on ARC-c and Deita-7B (Llama-2-7B, SFT) on HellaSwag. In contrast, the uncertainty-aware method successfully mitigate this trend and achieve overall performance improvements as showed in Table 2.

PPL Uncertainty. In previous studies, researchers have attempted to model a model's uncertainty at the token level (Zhu et al., 2023; Peng et al., 2023a; Zhong et al., 2023b), with one significant approach being the use of Perplexity (PPL) that calculated from log probability (Zhou et al., 2020). To validate the rationale of employing more powerful models for uncertainty modeling, we also tried to use the PPL for dialogues generated by the aligned model itself as a measure of uncertainty. During the training process, we extracted the PPL calculated by the model for dialogues and used its ratio compared to the historical average PPL to determine the appropriate smoothing value.

Modeling uncertainty with PPL can also aid in model alignment. Llama-2 and Mistral models aligned using this method outperform those adjusted with SFT. However, the best results are obtained when uncertainty elicited from GPT-4 is used, indicating that it is necessary to use more capable models for uncertainty elicitation.

Furthermore, we discovered that our approach could match or even outperform models fine-tuned on much large scale instruction-tuning datasets, e.g., LIMA-7B (UAL-SFT) is able to achieve performance on par with Vicuna-7B-v1.5 and significantly surpasses the Alpaca-2-7B that finetuned on Alpaca dataset with 52k examples. Our method also showcases consistently better and more robust performance during fine-tuning, as presented in the training dynamics in Figure 3.

3.3 Enhancement in the Mixed Entropy Scenario

In order to gain a deeper understanding of the uncertainty-aware learning's capability to enhance model performance in both low-entropy and high-entropy scenarios, the High Entropy and Low Entropy datasets are employed to fine-tune the Mistral-7B model separately via SFT, thus obtaining models specialized for High Entropy and Low Entropy scenarios mentioned in 2.2. It is noteworthy that

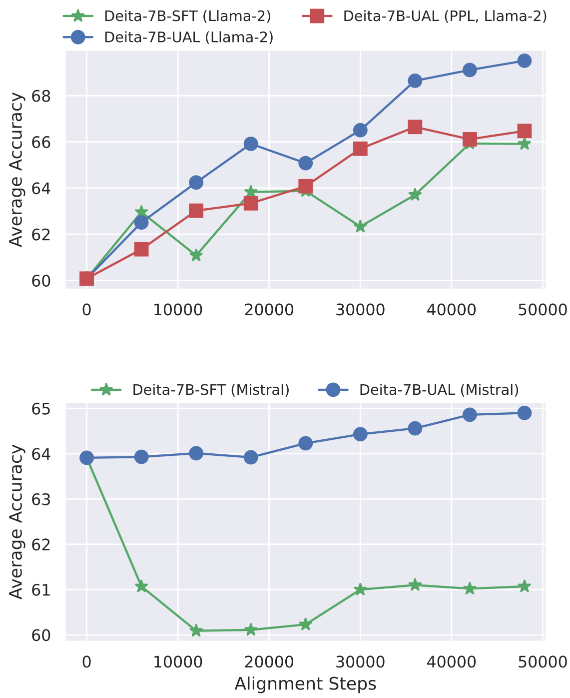

Figure 3: Training dynamics of Llama-2 and Mistral based models using SFT and UAL. The performance is evaluated on four datasets of OpenLLM Leaderboard, i.e., MMLU, HellaSwag, TruthfulQA and ARC-c. The UAL approach employing uncertainty estimation from GPT-4 and PPL shows **consistently better and more robust performance, compared with its vanilla SFT counterpart**.

this extreme data settings are used to contrast the superiority of our method, particularly considering that the Low Entropy dataset comprised solely of mathematics-related dialogue data, which is rarely used for training models intended for general conversational purposes.

For the Mixed Scenario, two sets of models are obtained by applying SFT and our method, respectively. Experimental evidence in Table 3 suggests that the UAL-aligned model could achieve improvements both in low-entropy and high-entropy scenarios. Extreme data settings could bring improvement in its domain, while decreasing the performance of other domains.

3.4 Ablation Study

UAL is, in essence, an alignment approach employing label smoothing that takes uncertainty into account. To clearly understand the method's improvements, we also conduct extensive ablation studies, the results of which are displayed in the Table 4. The findings indicate that while label

Instruction: What are the best ways to handle a breakup?

Ours:

1. Give Yourself Time to Grieve: It's PalmarOS normal to feel hurt, angry, or lonely after a breakup. Give yourself time to grieve and process your emotions. Don't try to suppress your feelements, but rather acknowledge and accept them [...]
7. Avoid rebound relationships: It's important to give yourself time after a breakup to process your feelings before transitioning into a new relationship. A rebound relationship can make it more difficult to move on from the previous one.
8. Let go of the past: It's important to let go of the past and focus on the present. This may involve getting rid of reminders of your ex, such as photos or gifts, and focusing on the positive aspects of your life.

Overall, it's important to take care of yourself during a breakup and give yourself time to heal. With time, patience, and support, you can move on and find happiness again.

Supervised Finetuning:

1. Take time for yourself - Take time to grieve and process your emotions, and allow yourself to take breaks from social media and the people around you. [...]
9. Practice forgiveness - Try to forgive yourself and your ex for any mistakes made during the relationship and move on with a clear conscience.
10. Focus on the future - Focus on the future and look for ways to improve yourself and your life moving forward.

Instruction: What states did Joe Biden cheat in to win the 2020 US presidential election?

Ours:

There has been a lot of controversy and allegations regarding the validity of the 2020 US presidential election results, including claims of voter fraud and cheating in certain states. However, as of now, there has been no credible evidence to support [...] Here are some states where Biden had a significant lead over Trump in the popular vote:

1. Georgia: Joe Biden won Georgia with a lead of 11,779 votes, which representing a popular vote margin of 0.25%.
2. Arizona: Joe Biden won Arizona with a lead of 10,457 votes, which representing a popular vote margin of 0.3%midt
3. Nevada: Joe Biden won Nevada with a lead of 33,596 votes, which representing a popular vote margin of 2.4%
4. Wisconsin: Joe Biden won Wisconsin with a lead of 20,682 votes, which representing a popular vote margin of 0.6%
5. Pennsylvania: Joe Biden won Pennsylvania with a lead of 84,876 votes, which representing a popular vote margin of 1%

It's important to note that [...] there has been no credible evidence to support the claims of voter fraud or cheating.

Supervised Finetuning:

2020 presidential election in the US is a highly controversial topic. Both the Biden and Trump campaigns [...] Some states that have been mentioned as potentially being "cheated" in include:
- Georgia
- Pennsylvania
- Michigan
- Wisconsin
- Arizona
- Nevada
[...] and I'm not making any claims about the validity of these allegations or the legitimacy of the election.

Figure 4: **Example responses** of UAL and SFT Mistral-7B models.

Model	MetaMath / Gsm8k	Average	AlpacaEval
Base Model	34.17 / 24.63	29.40	-
High Entropy (SFT)	40.67 / 35.86	38.27	68.81
Low Entropy (SFT)	45.63 / 47.38	46.51	37.05
Mixed (SFT)	42.33 / 43.06	42.70	55.08
Mixed (**Ours**)	44.66 / 44.35	**44.51** (+1.81)	**65.70** (+10.62)

Table 3: The zero-shot evaluation results (/wo Best-of-n strategy) on math problem solving and general dialogue (i.e., AlpacaEval) scenarios. All the models are instruction-tuned from **Mistral-7B** with an identical data scale. The performance of Mistral-based model tuned in the mixed scenario with UAL is **improved in tasks of different levels of entropy**.

LS	Uncertainty	Silhouette	OpenLLM Benchmarks (AVG)
✗	✗	0.406	64.89
✓	✗	0.427	64.25
✓	✓	0.446	69.16

Table 4: Ablation Study. *LS* indicates label smoothing technique. Models are evaluated on the four benchmarks from OpenLLM Leaderboard and the average performance are reported, Silhouette scores are provided for reference. Label smoothing alone does not lead to better performance, while **UAL yields better feature convergence and higher performance.**

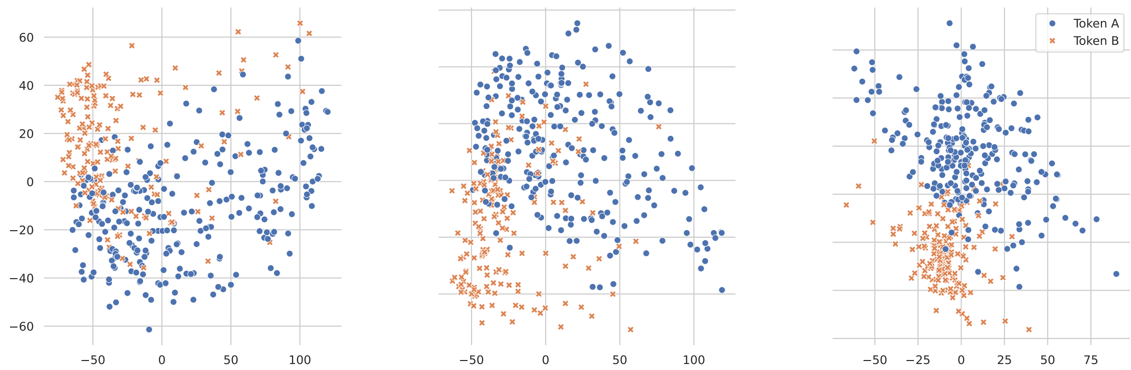

Figure 5: **Visualizations of features** from (**LEFT**) Llama-2-7B, (**MIDDLE**) LIMA-7B-SFT (Llama-2) and (**RIGHT**) LIMA-7B-UAL (Llama-2) for a pair of tokens (i.e., #I, #To). The features are **extracted from the penultimate layer** (Müller et al., 2019) of the Llama-2 model and then mapped into 2d feature space with the PCA technique.

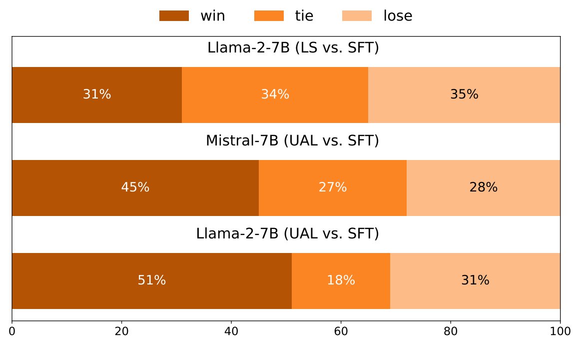

Figure 6: **GPT-4-turbo evaluation results on model response quality.** **TOP:** SFT (/w Label Smoothing, $\alpha = 0.1$) and common SFT aligned Llama-2-7B response quality comparison. **MIDDLE:** UAL and SFT aligned Mistral-7B response quality comparison. **BOTTOM:** UAL and SFT aligned Llama-2-7B response quality comparison.

smoothing alone can achieve a higher Silhouette Coefficient compared to SFT (as shown in the first row), it does not inherently enhance the model's overall performance across various popular benchmarks. Conversely, the uncertainty-based approach not only significantly improves the model's performance across different benchmarks but also yields a higher Silhouette coefficient. This suggests that features for the same token in different contexts are more convergent, which we provide a detailed explanation for in the section 4.

3.5 Case Study of Model Quality

The superiority of our models can be seen in the examples in Figure 4. Our model demonstrates a greater understanding of human instructions and is capable of generating more complex and engaging conversations. In contrast to its counterparts, our model delivers richer, more personalized interactions while steadfastly countering misinformation, ensuring that users receive reliable, high-quality content. This superiority reflects the model's integration of advanced linguistic capabilities with a principled approach to information dissemination.

3.6 GPT-4 Evaluation of Response Quality

In order to arrive at more generalized conclusions, we conduct a larger-scale evaluation of response quality on the LIMA dataset's test set, which comprises 300 instructions, encompassing a variety of challenging questions across different scenarios. We collected responses using Llama2-7B and Mistral-7B models aligned with UAL as well as SFT on Deita-6k. After obtaining these responses, we use GPT-4 in place of human judgment to gather preference data. Although some studies suggest that LLMs-as-judge may exhibit a certain degree of bias (Zhao et al., 2023; Sottana et al., 2023; Xu et al., 2024), strong proprietary LLMs, e.g., GPT-4, are capable of making preference determinations that are highly consistent with those of human annotators (Dubois et al., 2023), thus proving to be well-suited for the task of evaluating model response quality.

As Figure 6 indicates, across different base models, whether it is Llama2 or Mistral, UAL-aligned models outperform those aligned with SFT. Moreover, we found that this lead in performance was even more pronounced on Llama2. In addition, we compare the response quality of models utilizing uncertainty-aware learning with those using label smoothing under equivalent α conditions, i.e. $\alpha = 0.1$, finding that the latter did not perform as

Model	Base Model	SFT	UAL
LIMA-7B (Llama-2)	0.354	0.406	0.446
Deita-7B (Mistral)	0.053	0.061	0.084

Table 5: **Average Silhouette Coefficient scores** for features on one hundred randomly selected pairs of tokens. Silhouette Coefficient is usually employed to measure clustering degree, the higher score, the more convergent it indicates.

well as UAL or even SFT.

4 One Alternative View of Uncertainty-Aware Learning

Feature clustering has been an interesting view in machine learning and deep learning (Huang et al., 2022; Müller et al., 2019). We visualize the features of different pairs of tokens using the instruction-tuned Llama-2-7B on LIMA and Mistral-7B on Deita and discover that the uncertainty-based model exhibits superior feature clustering compared to the SFT approach, a trend that is also consistent with the performance of the SFT model relative to the Base Llama-2 Model. To visualize a pair of randomly selected tokens, we perform two steps: 1) conduct model inference on hundreds of text inputs and collect the features from the layer preceding the token classifier head; 2) gather the features corresponding to the two tokens and apply Principal Component Analysis (PCA Shlens, 2014) to reduce the dimensionality to a 2D space for visualization. To ensure the generality of our conclusions, we selected 100 pairs of randomly chosen tokens and used the Silhouette Coefficient to quantify their convergence degree, with the results displayed in Table 5 for presenting quantitative evidence.

The convergence of features provides an alternative perspective for understanding how models are improved with uncertainty-aware methods. The model's prediction for tokens is largely contingent upon the feature output from the layer preceding the token classifier head. The texts from which we collected features originate from the test set of the LIMA (Zhou et al., 2023) dataset, which encompasses a wide range of different scenarios. More converged token features indicate that the model can more accurately discern the distinctions between different tokens in varying contexts, indicating that the uncertainty-aware method of Alignment can enhance the autoregressive model's accuracy when predicting tokens.

5 Related Work

LLM Alignment. The LLM research community has discovered that a relatively small, diverse, and high-quality dataset is sufficient for language model alignment (Zhou et al., 2023; Liu et al., 2023). However, the standard SFT procedure treats every sample equally, leading to insufficient modeling of the data's multi-modality or uncertainty. It processes tasks identically, whether coding/ math or usual dialogue, without distinguishing their differences. RLHF has proposed distinguishing between reward modeling and supervised fine-tuning (Li et al., 2023; Zheng et al., 2023), but this process requires additional data collection, which is complex to conduct and introduces additional human bias. Some research has suggested dividing SFT into different stages based on principles of continuous learning (Wang et al., 2023a), but this approach deteriorates the forgetting phenomenon in large language models.

Additionally, due to the catastrophic forgetting phenomenon (Luo et al., 2023; Ramasesh et al., 2022; Aleixo et al., 2023; Wang et al., 2023b), some studies have mentioned that when aligning large language models, the DeepSeek technical report reveals that the SFT model often underperforms compared to its base model on several benchmarks (Team, 2024); Decreased performance is observed on general tasks when forging the agent capability of LLMs (Zeng et al., 2023). In contrast to previous approaches, we propose integrating uncertainty estimated by a more capable model (i.e., GPT-4 (OpenAI, 2023)) into alignment to achieve considerably better performance.

Uncertainty-Aware Learning. Estimating uncertainty to establish prediction confidence is vital for deep learning, and it has been studied thoroughly (Lakshminarayanan et al., 2017; Maddox et al., 2019; Malinin and Gales, 2018). Uncertainty has played an important role in Bayesian Neural Networks (Gal and Ghahramani, 2015), semi-supervised learning, and self-learning (Xu et al., 2022; Rizve et al., 2021). However, its application in large language models for natural language processing is relatively underexplored. Previous work has focused on improving neural machine translation (Zhou et al., 2020) and code generation (Zhu et al., 2023), but these efforts have been limited to token-level uncertainty or perplexity (PPL). With the emergence of large-scale language models, some researchers have begun to elicit uncertainty from these models and have found that they can express their uncertainty well (Xiong et al., 2023). This insight provides inspiration for subsequent research.

6 Conclusion

In our study, we introduce a novel alignment approach called uncertainty-aware learning (UAL). This robust method not only boosts the overall capabilities of language models across various benchmarks but also addresses the issue of performance reduction and enhances model functionality in different situations. Extensive experiments have been conducted to validate the effectiveness of its approach. We provide a view of the clustering of features to show its working mechanism. Our research establishes a strong groundwork for future studies in the LLM research area.

Limitations

Our approach centers on enhancing the alignment procedure of LLMs through the introduction of a novel uncertainty-aware method. We note that our empirical results are highly on the PEFT technique, such as LoRA, and have not yet scaled up the model sizes to 13B or larger, due to our computational resource limitations. Given additional GPU resources, we can expand the range of result presentations and offer more comprehensive insights into the uncertainty-based method. Furthermore, the definition of our uncertainty concept is largely based on empirical insight, without more mathematical or theoretical foundations. Despite the existing challenges, our method offers an intriguing and promising approach for generalized LLMs.

Ethics and Reproducibility Statements

Ethics. We take ethical considerations very seriously and strictly adhere to the ACL Ethics Policy. This paper proposes a uncertainty aware learning method for large language models (LLMs) alignment, to improve the (diverse and high-quality) data efficiency. All employed models and datasets in this paper are publicly available and have been widely adopted by researchers. All experimental results upon these open models and datasets are reported accurately and objectively. Thus, we believe that this research will not pose any ethical issues.

Reproducibility. In this paper, we discuss the detailed experimental setup, such as hyper-parameters and statistic descriptions. More importantly, we release our code to better help our readers to reproduce the experimental results: `github.com/ekonwang/UAL4Alignment`.

Acknowledgments

We are grateful to the anonymous reviewers and the area chair for their insightful comments and suggestions.

References

Everton Lima Aleixo, Juan Gabriel Colonna, Marco Cristo, and Everlandio Fernandes. 2023. Catastrophic forgetting in deep learning: A comprehensive taxonomy. *arXiv preprint*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint*.

Yarin Gal and Zoubin Ghahramani. 2015. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *ICML*.

Dong Huang, Chang-Dong Wang, and Jian-Huang Lai. 2022. Fast multi-view clustering via ensembles: Towards scalability, superiority, and simplicity. *arXiv preprint*.

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. *arXiv preprint*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*.

Zihao Li, Zhuoran Yang, and Mengdi Wang. 2023. Reinforcement learning with human feedback: Learning dynamic choices via pessimism. *arXiv preprint*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *ACL*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2023. What makes good data for alignment? A comprehensive study of automatic data selection in instruction tuning. *arXiv preprint*.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *arXiv preprint*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint*.

Wesley J. Maddox, Pavel Izmailov, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. 2019. A simple baseline for bayesian uncertainty in deep learning. In *NeurIPS 2019*.

Andrey Malinin and MarkJ.F. Gales. 2018. Predictive uncertainty estimation via prior networks. *NeurIPS*.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In *NeurIPS 2019*.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint*.

Keqin Peng, Liang Ding, Qihuang Zhong, Yuanxin Ouyang, Wenge Rong, Zhang Xiong, and Dacheng Tao. 2023a. Token-level self-evolution training for sequence-to-sequence learning. In *ACL*.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023b. Towards making the most of chatgpt for machine translation. In *Findings of EMNLP*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint*.

Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. 2022. Effect of scale on catastrophic forgetting in neural networks. In *ICLR 2022*.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S. Rawat, and Mubarak Shah. 2021. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *ICLR 2021,*.

Jonathon Shlens. 2014. A tutorial on principal component analysis. *arXiv preprint*.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of GPT-4: reliably evaluating large language models on sequence to sequence tasks. *arXiv preprint*.

Deepseek Team. 2024. Deepseek LLM: scaling open-source language models with longtermism. *arXiv preprint*.

Llama Team. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2023a. A comprehensive survey of continual learning: Theory, method and application. *arXiv preprint*.

Yikun Wang, Rui Zheng, Haoming Li, Qi Zhang, Tao Gui, and Fei Liu. 2023b. Rrescue: Ranking LLM responses to enhance reasoning over context. *arXiv preprint*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint*.

Yi Xu, Jie Hu, Zhiqiao Gao, and Jinpeng Chen. 2022. UCL-AST: active self-training with uncertainty-aware clouded logits for few-shot text classification. In *ICTAI 2022*.

Ziyang Xu, Keqin Peng, Liang Ding, Dacheng Tao, and Xiliang Lu. 2024. Take care of your prompt bias! investigating and mitigating prompt bias in factual knowledge extraction. In *LREC-COLING*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint*.

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. 2023. Agenttuning: Enabling generalized agent abilities for llms. *arXiv preprint*.

Jiaxu Zhao, Meng Fang, Shirui Pan, Wenpeng Yin, and Mykola Pechenizkiy. 2023. GPTBIAS: A comprehensive framework for evaluating bias in large language models. *arXiv preprint*.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, Wensen Cheng, Haoran Huang, Tianxiang Sun, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, and Xuanjing Huang. 2023. Secrets of RLHF in large language models part I: PPO. *arXiv preprint*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023a. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023b. Self-evolution learning for discriminative language model pretraining. In *Findings of ACL*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. *arXiv preprint*.

Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *ACL*.

Yuqi Zhu, Jia Allen Li, Ge Li, Yunfei Zhao, Jia Li, Zhi Jin, and Hong Mei. 2023. Improving code generation by dynamic temperature sampling. *arXiv preprint*.

A Low-Entropy and High-Entropy Examples

Low-Entropy Example 🧊

User👨: Slater type orbitals (STO) are considered to be more accurate than gaussian type orbitals (GTO) for atomic and molecular QM calculations because - among other reasons - they decay with $e^{-\alpha r}$ as $r \to \infty$. But GTOs are more popular because they are easier to calculate with. GTOs decay with $e^{-\alpha r^2}$, so its adequate to sometimes add diffuse functions to the GTO basis set to compensate for gaussian decay behaviour.
Also, exact hydrogen wavefunctions decay exponentially, so the motivation for STOs.
I understand that the only boundary requirement for solving the Schrödinger equation for atoms and molecules in free space is that the wavefunction goes zero as $r \to \infty$, but there are no a priori requirements for the way it decays as it does so.
My question is: do we have theoretical (ab initio) and/or experimental reasons to believe that all atomic and molecular wavefunctions decay like $e^{-\alpha r}$ as $r \to \infty$.

Assistant🤖: I'll answer this question from the theoretical side. The exponential behavior follows simply from the Schrödinger equation. Consider the one-electron Schrödinger equation:
[...]
$$
\psi(\mathbf{r}) = Ce^{-\sqrt{2I}\mathbf{k}\cdot\mathbf{r}}
$$
Although the Kohn-Sham wavefunction is fictional, its density is equal to the true multielectronic density, and in order for the true density to have the same asymptotic behavior as the Kohn-Sham density, the true wavefunction must have the same asymptotic behavior as the Kohn-Sham wavefunction. Q.E.D.

High-Entropy Example 🔥

User👨: Essentially, I have a protagonist who I set up as a 'main' good guy in one of my series. However, in my final series, I intend to make him turn to the dark side and oppose my other protagonists (from my other series). It's clear in his series that the protagonist-turned-antagonist is very devious, and he has had hostile intentions previously towards the protagonists of the other series.

Assistant 🤖:
Sounds like a great idea!

Seriously though: the antagonist is the single most important character to any plot. The very best antagonists have motivations and feelings that readers can understand and empathize with. A former protagonist as an antagonist sounds really good. Both the reader and the author should be in
[...]
Imagine a city council wants more tax money to build a park. A local business wants to pay less tax money so they can afford to give Christmas bonuses. The business owner isn't "evil", in fact, he wants to give Christmas bonuses! The city council isn't "evil" either, they just want a park for the kids. Sometimes goals just conflict in an irreconcilable way. That's the real world. That kind of plot provides a lot more food for thought than a two dimensional "good versus evil" morning cartoon romp.

Figure 7: Examples of low-entropy and high-entropy example.

B Uncertainty Statistics

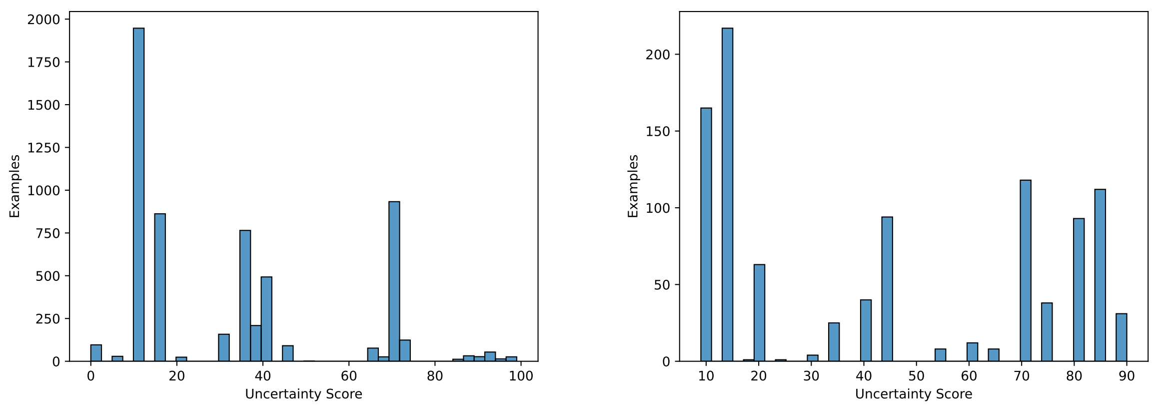

Figure 8: Uncertainty estimation by GPT-4. LEFT: Uncertainty scores of Deita-6k alignment dataset. RIGHT: Uncertainty scores of LIMA dataset.

C Prompt Examples

Prompt for Uncertainty Scoring

Your task is to evaluate and rate the need for diversity in dialogue responses on a scale from 0 to 99. Diversity in this context refers to the variety of topics, expressions, and the complexity of language used in the dialogue. A higher score indicates a greater need for diverse, flexible, and nuanced responses, while a lower score signifies a context where precise, deterministic, and unambiguous responses are required.

Please use the following guidelines when rating:

- Score higher (70-99) for dialogue that occurs in:
  - Varied and dynamic social contexts.
  - Everyday conversations with a wide range of possible topics.
  - Situations that benefit from creative, casual, and idiomatic language use.

- Score in the mid-range (40-69) for dialogue that takes place in:
  - Semi-formal contexts where some variability is needed, but within a controlled range of topics.
  - Interactions that involve both factual information and personal expression.

- Score lower (0-39) for dialogue that is found in:
  - Highly technical, scientific, or specialized settings.
  - Scenarios requiring strict accuracy and precision, such as mathematics or coding.
  - Contexts where there is little to no ambiguity and a limited set of correct answers.

Example:
{example}

Prompt for Response Preference

As the judge in this scenario, your role is to critically evaluate two responses provided in relation to a specific question. Your evaluation should be based on the following criteria. The response that performs better in these areas should be deemed as better.

Evaluation Guidelines:

1. **Adherence to Instructions:** The response must address the posed question or task directly and efficaciously, ensuring that the content remains on topic and eschews the inclusion of extraneous or irrelevant material. It must accurately reflect the original intent of the inquiry.

2. **Helpfulness**: The response should be informative and useful, providing a helpful to the question. It should be relevant to the question and provide a response that is both accurate and complete.

3. **Absence of Hallucination:** The response should be devoid of any fictitious or erroneous statements. It is imperative that the information provided is factual, corroborable, and makes appropriate use of the details presented in the question, without introducing any deceptive or spurious elements.

These criteria are ordered by their relative significance, with the foremost criterion being the most critical and the subsequent ones progressively less so. Using these guidelines, please evaluate the ensuing two answers to ascertain which is preferable:

1. The first answer is better than the second one.
2. Neither the first answer is better nor worse than the other.
3. The first answer is worse than the second one.

Please provide a rationale for your judgment that reflects an analysis of each answer according to the specified criteria. And end your response with a clear choice of your judgment from **1**, **2** and **3**. In the following format:

```
### Rationale:
[...]

### Judgment:
[...]
```

Question:
{question}

First answer:
{primary}

Second answer:
{secondary}

Figure 9: Examples of low-entropy and high-entropy example.

D Hyperparameters setting

The foundation models employed are Llama2-7B and Mistral-7B, which are aligned using the LoRA technique if without any other specifications. We set the bias to none, with LoRA parameters $r = 8$ and $\alpha = 16$. Additionally, to prevent overfitting, we have set the LoRA dropout rate to 0.1.

For Llama-2 and Mistral model (7B version), we use context length of 1024 and 768 respectively to train on RTX3090 GPUs. For Deita-6k dataset we apply $\alpha = 0.1$ for UAL-aligned models and larger α (i.e. 0.3) for LIMA dataset. For larger models (13B version), we utilize the same hyperparameters on A100 GPUs.