ENHANCING TRUST IN LARGE LANGUAGE MODELS WITH UNCERTAINTY-AWARE FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have revolutionized the field of natural language processing with their impressive reasoning and question-answering capabilities. However, these models are sometimes prone to generating credible-sounding but incorrect information, a phenomenon known as LLM hallucinations. Reliable uncertainty estimation in LLMs is essential for fostering trust in their generated responses and serves as a critical tool for the detection and prevention of erroneous or hallucinated outputs. To achieve reliable and well-calibrated uncertainty quantification in open-ended and free-form natural language generation, we propose an uncertainty-aware fine-tuning approach for LLMs. This approach enhances the model's ability to provide reliable uncertainty estimates without compromising accuracy, thereby guiding them to produce more trustworthy responses. We introduce a novel uncertainty-aware causal language modeling loss function, grounded in the principles of decision theory. Through rigorous evaluation on multiple free-form question-answering datasets and models, we demonstrate that our uncertainty-aware fine-tuning approach yields better calibrated uncertainty estimates in natural language generation tasks than fine-tuning with the standard causal language modeling loss. Furthermore, the experimental results show that the proposed method significantly improves the model's ability to detect hallucinations and identify out-of-domain prompts.

1 INTRODUCTION

Large Language Models (LLMs) have shown remarkable success in various natural language processing tasks (Touvron et al., 2023; Gemma et al., 2024; Achiam et al., 2023) and are increasingly becoming ubiquitous in a variety of domains for their decision-making and reasoning abilities (Eigner & Händler, 2024). However, their real-world deployment, particularly in high-stakes and safety-critical applications, is hindered by challenges such as hallucinations and out-of-domain prompts, which can lead to the generation of erroneous or nonsensical outputs. Hallucinations, often described as plausible-sounding but incorrect or unfaithful model generations (Ji et al., 2023), present a crucial challenge in developing trustworthy systems especially in critical domains such as medical (Ahmad et al., 2023) and legal (Magesh et al., 2024). The ability to recognize out-of-domain prompts and to acknowledge the limits of a model's knowledge base paves the way for building safe AI systems (Amodei et al., 2016).

Uncertainty quantification (UQ) in LLMs plays a pivotal role in understanding what the model knows and does not know, which is an active area of research for free-form natural language generation (NLG) (Kadavath et al., 2022; Kuhn et al., 2023; Lin et al., 2024). UQ methods has emerged as a step towards determining the trustworthiness of responses generated by LLMs (Fadeeva et al., 2023; Plaut et al., 2024; Kadavath et al., 2022). Uncertainty estimation techniques such as semantic entropy (Kuhn et al., 2023) have shown to be effective indicators in detecting 'confabulations' (Farquhar et al., 2024), a subcategory of hallucinations characterized by the generation of arbitrary and incorrect responses.

The calibration of uncertainty estimates is crucial for the reliability of LLMs; a well-calibrated model should correlate low uncertainty with accurate responses and high uncertainty with likely incorrect responses. However, recent studies (Xiong et al., 2024; Yang et al., 2024) have revealed that LLM predictions are often poorly calibrated, leading to overconfidence in incorrect outputs.

This problem is more pronounced in fine-tuned language models (Kong et al., 2020; Liu et al., 2024b). Unlike the pre-training phase, where models are exposed to vast amounts of unlabeled data, fine-tuning involves limited labeled data. Consequently, the immense capacity of LLMs can lead to overfitting on this limited data, producing overconfident predictions (Kong et al., 2020). This presents a substantial challenge, as the model may produce unreliable uncertainty metrics that are influenced by the model's miscalibrated token confidence. Moreover, evaluating calibration in NLG is particularly challenging due to the variable lengths of generated text compared to reference sentences. Traditional calibration error metrics (Naeini et al., 2015; Nixon et al., 2019) assume a fixed number of outcomes or classes, which aligns well with classification tasks but not with the open-ended nature of NLG, where the number of tokens in generated text and reference sentences can differ. To address this, we exploit the inverse correlation between uncertainty quantification and the quality of generated text, allowing us to perform calibration analysis in NLG settings.

In this work, we introduce an uncertainty-aware fine-tuning approach for LLMs, which is grounded in decision theory and tailored for free-form natural language generation. Our approach is orthogonal to existing uncertainty quantification methods and is driven by the goal of enhancing the reliability of uncertainty metrics through uncertainty-aware fine-tuning. Specifically, we achieve this with an optimization objective that encourages the model to learn to associate high uncertainty with incorrectly generated tokens and low uncertainty with correctly generated tokens, while maximizing accuracy within the framework of causal language modeling. We show that fine-tuning with our calibration objective enhances the reliability of uncertainty quantification in LLMs. The calibrated uncertainty estimates serve as a crucial tool for enhancing the trustworthiness of generated responses, represents a substantial step towards identifying confabulations, and improving decision-making capabilities through selective generation (Ren et al., 2022).

Our contributions are as follows:

- We propose uncertainty-aware causal language modeling (UA-CLM) loss function, designed for fine-tuning LLMs to produce well-calibrated uncertainty estimates in free-form natural language generation.
- We conduct uncertainty calibration analysis in free-form NLG settings, employing an innovative methodology that leverages the inverse correlation between uncertainty quantification and the quality of the generated text.
- We perform a comprehensive empirical evaluation across four key aspects: hallucination detection; selective generation; out-of-prompt detection; and calibration analysis, demonstrating that UA-CLM significantly enhances the quality of uncertainty estimates in LLMs in open-ended and free-form question-answering tasks. Notably, these enhancements in uncertainty calibration are achieved without compromising the accuracy when compared to standard CLM.
- In addition, we have also applied our proposed UA-CLM methodology to a large vision language model (LVLM), demonstrating its efficacy in the open-ended visual question-answering task. This extension of our work shows the versatility of UA-CLM in handling multimodal inputs and complex tasks beyond text-based question-answering, showcasing it's utilty to wider range of applications.

2 Background and Related works

2.1 Uncertainty estimation in natural language generation

We refer to Abdar et al. (2021); Gawlikowski et al. (2023) for surveys on uncertainty quantification (UQ) in deep learning. In machine learning models, predictive uncertainty is composed of two primary sources: epistemic uncertainty, associated with model's lack of knowledge and aleatoric uncertainty - the inherent noise in the data or observation. As LLMs continue to evolve rapidly, there is an increasing interest in enhancing our understanding of the uncertainty associated with LLM responses for developing trustworthy and reliable systems (Fadeeva et al., 2023). UQ methods from deep learning can be effectively applied to structured natural language processing tasks, such as text classification (Xiao & Wang, 2019) and multiple-choice question answering (Kumar et al., 2023). However, the application of these methods to free-form natural language generation presents distinct challenges.

The landscape of UQ in free-form NLG is a dynamic field of research, where methodologies are generally bifurcated into white-box (Fomicheva et al., 2020; Kuhn et al., 2023) and black-box approaches (Lin et al., 2024; Xiong et al., 2024). White-box methods necessitate access to the model's logits, or likelihood scores, or other internals of LLMs. Black-box methods rely solely on the analysis of the text sequences generated by the LLMs. Recent work (Tian et al., 2023) explores prompting techniques in reinforcement learning from human feedback (RLHF) language models to explicitly elicit verbalized response confidence. In contrast, another line of research has explored unsupervised methods for UQ from the models by utilizing confidence (Plaut et al., 2024), perplexity (Fomicheva et al., 2020), token entropy (Malinin & Gales, 2021) and semantic entropy (Kuhn et al., 2023) to quantify uncertainty in LLM responses. Previous research (Minderer et al., 2021) has shown that confidence or entropy measures can be susceptible to poor calibration and may not fully reflect a model's underlying uncertainties. Our work focuses on improving these uncertainty metrics in white-box settings by better calibrating the language models with uncertainty-aware finetuning.

2.2 Model Calibration

Calibration is important in applications where decision-making relies on not just the model's predictions, but also on the trustworthiness of its uncertainty scores. It is important to capture well-calibrated uncertainty estimates of a model for creating reliable and trustworthy systems. Model calibration is a well-explored area of research in deep learning, with a variety of strategies proposed to enhance the calibration of deep neural networks for classification and regression tasks. These strategies include post-hoc rescaling techniques (Guo et al., 2017; Kull et al., 2017), which adjust the model's predictions to better align with true event likelihoods; data augmentation, which enriches the training dataset to promote generalization (Thulasidasan et al., 2019; Hendrycks et al., 2020); and probabilistic modeling approaches that integrate uncertainty directly into the model's architecture (Blundell et al., 2015; Lakshminarayanan et al., 2017). Other line of works utilize explicit calibration loss functions during training to directly optimize for calibration (Kumar et al., 2018; Krishnan & Tickoo, 2020; Mukhoti et al., 2020; Karandikar et al., 2021), that has resulted in better calibrated models.

Calibration of LLMs for natural language processing tasks is an ongoing area of research. Prior works have largely focused on refining LLMs for structured tasks like text classification (Kong et al., 2020) or multiple-choice question answering (Desai & Durrett, 2020; Jiang et al., 2021). Studies like the one by Xiong et al. (2024) have highlighted that despite the impressive performance of foundational LLMs on a wide array of tasks, these models often exhibit poor calibration, particularly exhibiting overconfidence in their predictions. As LLMs are increasingly used in natural language generation tasks, new calibration techniques (Geng et al., 2024) are emerging to enhance the reliability of the generated text. More recently, Liu et al. (2024b) introduced a calibration technique for LLMs that trains a single linear layer over the model's last hidden layer representations to predict a bias term, which is then used to adjust the model's logits and alter the generation confidence for short-form and long-form responses. Band et al. (2024) propose a training objective for linguistic calibration, utilizing reinforcement learning to optimize and calibrate long-form text generations. Kapoor et al. (2024) proposed a calibration tuning method designed for LLMs in multiple-choice question-answering settings. Prior work by Liu et al. (2024b) has shown that standard fine-tuning of LLMs can lead to poorer calibration. The calibration of LLMs for free-form text generation, as well as uncertainty-aware fine-tuning, represents a significant open area of research. Our work addresses this gap by developing an uncertainty-aware fine-tuning method of LLMs, fine-tuning less than 1% of the model parameters, to achieve well-calibrated models for free-form natural language generation.

2.3 Fine-tuning Large Language Models

With the emergence of foundation models, fine-tuning have become a common practice in the field of natural language processing, enabling the adaptation of general-purpose pre-trained models to specialized tasks and domains. As fine-tuning a LLM with billions of parameters can be resource-intensive, parameter-efficient fine-tuning (Mangrulkar et al., 2022) strategies have been proposed.These parameter-efficient fine-tuning techniques also mitigate catastrophic forgetting more effectively in comparison to full fine-tuning (Wang et al., 2022). One approach is to update only a subset of the model's parameters, such as adapter modules (Houlsby et al., 2019) or

Low-Rank Adaptation (LoRA) (Hu et al., 2022), where only a small set of parameters are updated while the pre-trained weights are frozen. Another strategy is prompt-based fine-tuning (Liu et al., 2023), where models are conditioned on task-specific prompts to guide the text generation without model parameter updates. We leverage LoRA strategy to illustrate our proposed uncertainty-aware finetuning in this work.

3 UNCERTAINTY-AWARE CAUSAL LANGUAGE MODELING

Motivated by the need to overcome the challenges of uncertainty miscalibration (Xiong et al., 2024) in Large Language Models (LLMs) and the increasing trend of fine-tuning pre-trained foundational models for domain-specific adaptation — where fine-tuned LLMs often exhibit overconfidence in their predictions (Kong et al., 2020) — we propose a novel uncertainty calibration fine-tuning approach for natural language generation settings. We introduce a novel uncertainty-aware causal language modeling loss based on the principles of decision theory (Murphy, 2012). Our fine-tuning approach emphasizes increasing the uncertainty for wrong token predictions, while optimizing for accuracy and certainty for correct token predictions. Decision theory offers a mathematical and theoretical framework that guides to achieve optimal predictions by employing a task-specific utility function. Within the decision theory framework, our task is to generate natural language text accompanied by reliable uncertainty estimates. The utility function in this scenario is represented by the uncertainty-aware optimization objective function that is aimed at producing well-calibrated uncertainty estimates for causal language modeling. We design a differentiable loss function that incentivizes the model to yield low uncertainty when it generates correct tokens, and encourages the model to exhibit high uncertainty when it is at risk of predicting the next token incorrectly.

In causal language modeling, the goal is to predict the next token in a sequence given the previous tokens. Given a sequence of tokens $[w_1, w_2, \ldots, w_T]$, where T is the length of the sequence and each token w_i is an element from a fixed vocabulary of size V, the model aims to learn the conditional probability distribution $P_\theta(w_i|w_{0:i\text{-}1})$ for each token w_i given the preceding set of tokens $w_{0:i\text{-}1}$; where, θ represents the parameters of the LLM. The loss function for standard causal language modeling (CLM) is typically the negative log-likelihood as defined in Equation 1 below.

$$\mathcal{L}^{\theta}_{\text{CLM}} := -\frac{1}{T}\sum_{i=0}^{T} \log P_\theta(w_i|w_{0:i-1}) \tag{1}$$

Desideratum: The desired and ideal outcome in causal language modeling is to achieve a state where every correctly generated token is assigned low predictive uncertainty, and high predictive probability, reflecting the model's high confidence in its accuracy. Conversely, for every token that is generated incorrectly, the model should assign high uncertainty, and low predictive probability, denoting low confidence in these instances. This ensures that the model's confidence levels and uncertainty estimates are perfectly calibrated with the actual correctness of its predictions.

We define the uncertainty-aware causal language modeling (UA-CLM) loss in Equation 2 based on the above desideratum. The loss function captures the trade-off between predictive accuracy and uncertainty calibration, and is composed of two terms: one that deals with incorrectly generated tokens and one that deals with correctly generated tokens.

$$\begin{aligned}\mathcal{L}^{\theta}_{\text{UA-CLM}} := &\underbrace{-\frac{1}{|\widetilde{C}|}\sum_{i\,\in\,\widetilde{C}} P_\theta(w_i|w_{0:i\text{-}1}) \log\left(\tanh\left(H_i\right)\right)}_{\text{Utility function for incorrect tokens}} \\ &\underbrace{-\frac{1}{|C|}\sum_{i\,\in\,C} \left(1 - P_\theta(w_i|w_{0:i\text{-}1})\right) \log\left(1 - \tanh\left(H_i\right)\right)}_{\text{Utility function for correct tokens}}\end{aligned} \tag{2}$$

where,

$$H_i := -\sum_{j=1}^{V} P_\theta(w_i^j|w_{0:i\text{-}1}) \log P_\theta(w_i^j|w_{0:i\text{-}1}) \tag{3}$$

Here, H_i is the entropy of the probability distribution of the ith token w_i given the previous tokens $w_{0:i-1}$ in the sequence, $\overline{w_i}$ is the ground-truth reference token, $C := \{i \mid w_i = \overline{w_i}\}$ is a set of indices corresponding to correctly predicted tokens, $\widetilde{C} := \{i \mid w_i \neq \overline{w_i}\}$ is a set of indices corresponding to incorrectly predicted tokens, V is the size of the vocabulary, w_i^j is the jth token in the vocabulary, and $P_\theta(w_i^j|w_{0:i-1})$ is the predicted probability of the jth token in the vocabulary. The hyperbolic tangent function is employed to scale the token entropy values, ensuring that $\tanh(H_i)$ lies in the interval [0, 1].

The loss function in Equation 2 offers theoretical guarantees as an optimization objective and satisfies the desideratum, converging to a perfect value of zero when all correctly generated tokens have a predictive probability of 1 (indicating high confidence) and scaled predictive entropy of 0 (implying low uncertainty), while all incorrectly generated tokens have a predictive probability of 0 (representing low confidence) and scaled predictive entropy of 1 (implying high uncertainty). The differentiable utility functions in the UA-CLM loss as shown in Equation 2 steers the predictive probabilities and uncertainty estimates to align with the accuracy of subsequent token predictions in autoregressive models. When the uncertainty estimates in the predicted tokens are misaligned, the loss increases, thereby directing the stochastic gradient computations to drive the loss towards minimization. This loss reduces when the uncertainty estimates conform to the desideratum, enabling the model to produce well-calibrated uncertainties while maximizing accuracy.

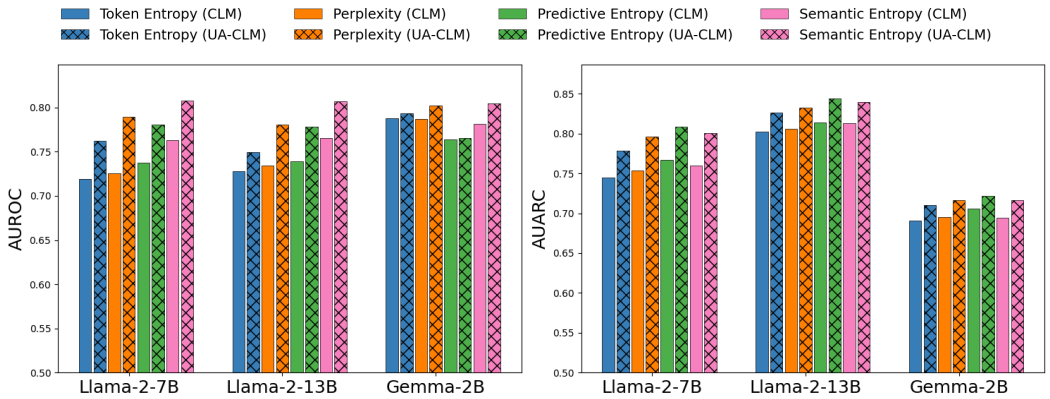
Algorithm 1 Uncertainty-aware fine-tuning in LLMs

Input: Pre-trained LLM $\mathbb{M}$ with parameters ϕ, LoRA parameters θ, learning rate η, epochs E, training data $\mathcal{D}$
Output: Uncertainty calibrated fine-tuned LLM $\mathbb{M}'$
Initialize LoRA parameters θ, and freeze ϕ
for $epoch = 1$ to E **do**
  for each batch $B \subseteq \mathcal{D}$ **do**
    Compute forward pass to get $P_{\phi+\theta}(w_i|w_{0:i\text{-}1})$ and token uncertainty estimate H_i
    Compute UA-CLM loss $\mathcal{L}^{\phi+\theta}_{\text{UA-CLM}}$ ▷ Equation 2
    Compute gradients of loss function w.r.t. θ, $\nabla_\theta \mathcal{L}^{\phi+\theta}_{\text{UA-CLM}}$
    Update LoRA parameters: $\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L}^{\phi+\theta}_{\text{UA-CLM}}$
  end for
end for
$\theta^* \leftarrow \theta$
$\mathbb{M}' \leftarrow \mathbb{M}$ with updated LoRA parameters θ^*
return $\mathbb{M}'$

Our proposed UA-CLM loss is designed to be agnostic to various parameter-efficient fine-tuning methods in LLMs; however, in this paper, we leverage and illustrate its application through Low-Rank Adaptation (LoRA) (Hu et al., 2022) as described in Algorithm 1. By incorporating uncertainty directly into the loss function, the model can not only learn to improve accuracy but also to understand a meaningful representation of uncertainty in its predictions. This dual emphasis on accuracy and uncertainty in the optimization objective ensures that the model's uncertainty estimates are closely aligned with the actual predictive accuracy of the generated tokens, leading to improved uncertainty calibration in natural language generation.

4 Experiments and Results

We perform extensive empirical evaluation to compare our proposed uncertainty-aware causal language modeling (UA-CLM) fine-tuning method to the standard causal language modeling (CLM) fine-tuning method on open-ended, sentence-length and free-form natural language generation tasks. Our comprehensive evaluation rigorously assesses the quality of uncertainty estimates and the quality of the generated text. This includes an analysis of broadly four aspects: hallucination detection, uncertainty-guided selective generation, out-of-domain prompt detection, and calibration analysis based on the inverse correlation between the uncertainty estimates and the quality of generated text.



(a) Hallucination detection (b) Selective generation

Figure 1: The proposed Uncertainty-aware Causal Language Modeling (UA-CLM) outperforms standard Causal Language Modeling (CLM) in all four UQ metrics across various models. The performance is evaluated using AUROC for hallucination detection and AUARC for selective generation based on four distinct UQ metrics.

4.1 EXPERIMENTAL SETTINGS

Datasets We utilize free-form question-answering (QA) datasets to evaluate the proposed methods on LLMs: CoQA (Reddy et al., 2019), an open-book conversational QA dataset; TriviaQA (Joshi et al., 2017), a reading comprehension QA dataset. These datasets are frequently utilized benchmarks for evaluating uncertainty quantification (UQ) in LLMs for natural language generation, as evidenced by prior works from Kuhn et al. (2023), Lin et al. (2024), and Farquhar et al. (2024). We employ the OK-VQA dataset (Marino et al., 2019), an open-ended visual question-answering (VQA) dataset to extend our evaluation to large vision language models (LVLMs), thereby providing a comprehensive analysis of our approach across diverse open-ended free-form QA tasks. Additionally, we use BioASQ (Krithara et al., 2023), a biomedical question-answering dataset for the evaluation of out-of-domain prompt detection. In our experiments, we utilize the development split of the CoQA dataset, which contains approximately 8,000 question-answer pairs, the validation split of TriviaQA with around 10,000 question-answer pairs, and the validation split of OK-VQA, comprising roughly 5,000 question-answer pairs along with their corresponding images. For each dataset, we allocate 20% of the data for fine-tuning purposes, while the remaining 80% serve as the test sets for evaluation. We use a standard prompt across all datasets, more details on the datasets and the prompt are provided in Appendix A.1.1 and A.1.2, respectively.

Models We use the Llama-2 models with 7B and 13B parameters (Touvron et al., 2023) and the Gemma model with 2B parameters (Gemma et al., 2024) for the free-form QA experiments. Additionally, we utilize the LLaVA-1.5 model with 7B parameters (Liu et al., 2024a) for the open-ended visual question-answering task.

Fine-tuning We perform parameter-efficient fine-tuning using the Low-rank Adaptation (LoRA) framework (Hu et al., 2022). The models undergo fine-tuning as described in Section 3 for the uncertainty-aware causal language modeling method. For comparative purposes, we also fine-tune models using the standard cross-entropy loss to evaluate against the standard causal language modeling fine-tuning method. For all our experiments, the models are fine-tuned for a concise duration of 3 epochs, utilizing only 20% of the data split. The optimization is carried out using the AdamW optimizer (Loshchilov & Hutter, 2019), with an initial learning rate of $1e$-4, a weight decay of 0.001, and a warm-up ratio of 0.03. We follow the same setup for both CLM and UA-CLM methods for a fair comparison. During the fine-tuning process, only the LoRA parameters are updated, while all other model parameters remain frozen. We provide more details on the hyperparameters and implementation in Appendix A.1.3, to facilitate the reproducibility of the results.

Table 1: Evaluation of uncertainty quantification: Comparative analysis of the proposed Uncertainty-aware Causal Language Modeling (UA-CLM) and standard Causal Language Modeling (CLM) methods (the best values are in **bold**). The comparison spans different datasets and models, with quality of uncertainty quantification evaluated using the Area Under the Receiver Operating Characteristic (AUROC) and the Area Under the Accuracy-Rejection Curve (AUARC) based on four different uncertainty metrics.

Dataset	Model	Finetuning Method	AUROC ↑ (Hallucination detection)				AUARC ↑ (Area under accuracy-rejection curve)			
			Token Entropy	Perplexity	Predictive Entropy	Semantic Entropy	Token Entropy	Perplexity	Predictive Entropy	Semantic Entropy
CoQA	Llama-2-7B	CLM	0.6252	0.6320	0.6635	0.6889	0.8230	0.8290	0.8516	0.8405
		UA-CLM	**0.6955**	**0.7398**	**0.7413**	**0.7741**	**0.8246**	**0.8477**	**0.8743**	**0.8571**
	Llama-2-13B	CLM	0.6302	0.6348	0.6815	0.6910	**0.8586**	0.8594	0.8741	0.8711
		UA-CLM	**0.6701**	**0.7255**	**0.7363**	**0.7694**	0.8571	**0.8596**	**0.8847**	**0.8722**
	Gemma-2B	CLM	0.7723	0.7606	0.7295	0.7618	0.8239	0.8259	0.8447	0.8281
		UA-CLM	**0.7780**	**0.7837**	**0.7358**	**0.7871**	**0.8350**	**0.8367**	**0.8521**	**0.8380**
TriviaQA	Llama-2-7B	CLM	0.8135	0.8192	0.8108	0.8371	0.6667	0.6779	0.6824	0.6798
		UA-CLM	**0.8293**	**0.8393**	**0.8197**	**0.8423**	**0.7324**	**0.7442**	**0.7430**	**0.7434**
	Llama-2-13B	CLM	0.8264	0.8333	0.7971	0.8407	0.7464	0.7526	0.7532	0.7556
		UA-CLM	**0.8297**	**0.8352**	**0.8033**	**0.8447**	**0.7960**	**0.8059**	**0.8040**	**0.8069**
	Gemma-2B	CLM	0.8030	0.8138	**0.7989**	0.8018	0.5582	0.5639	0.5677	0.5599
		UA-CLM	**0.8085**	**0.8211**	0.7960	**0.8228**	**0.5858**	**0.5963**	**0.5910**	**0.5956**
OK-VQA	LLaVA-1.5-7B	CLM	0.5504	0.5419	0.5455	0.5370	0.5809	0.5781	0.5790	0.5747
		UA-CLM	**0.6001**	**0.5984**	**0.6106**	**0.6638**	**0.5989**	**0.5965**	**0.6012**	**0.6265**

Table 2: Generated text quality evaluation: Comparative analysis of Causal Language Modeling (CLM) and Uncertainty-aware Causal Language Modeling (UA-CLM) fine-tuning methods. The results in the table indicate that UA-CLM achievies similar or better ROUGE-L scores and accuracy achieved by standard CLM across a range of models and datasets.

Finetuning Method	Llama-2-7B (CoQA)		Llama-2-7B (TriviaQA)		Llama-2-13B (CoQA)		Llama-2-13B (TriviaQA)		Gemma-2B (CoQA)		Gemma-2B (TriviaQA)	
	ROUGE-L	Accuracy	ROUGE-L	Accuracy	ROUGE-L	Accuracy	ROUGE-L	Accuracy	ROUGE-L	Accuracy	ROUGE-L	Accuracy
CLM	0.8886	0.9253	0.5867	0.6385	0.9106	0.9406	0.6588	0.6967	0.8654	0.9143	0.4349	0.4759
UA-CLM	0.8882	0.9264	0.6342	0.6754	0.9118	0.9461	0.7277	0.7710	0.8632	0.9088	0.4563	0.4959

UQ metrics To assess the uncertainty quantification from CLM and UA-CLM in the context of free-form text generation, we employ four widely used metrics as the baselines: mean token entropy (Fomicheva et al., 2020), perplexity (Fadeeva et al., 2023), predictive entropy (Malinin & Gales, 2021) and semantic entropy (Kuhn et al., 2023). The predictive entropy and semantic entropy are estimated by generating 5 stochastic sequences from the model, each obtained through temperature sampling with a temperature setting of T=0.3.

4.2 Evaluation and Results

Hallucination detection We evaluate the performance of detecting confabulations (hallucinations) in the generated text using uncertainty estimates. Hallucination detection is a binary classification task to distinguish between correct and hallucinated (incorrect) responses based on the uncertainty estimate. We use Area Under the Receiver Operating Characteristic (AUROC) (Davis & Goadrich, 2006) to evaluate the quality of uncertainty quantification in terms of the model's capability to detect hallucinations, following the methodology used in (Farquhar et al., 2024). A higher AUROC indicates that the model is more effective at identifying correct responses and flagging hallucinations. The bar plot depicted in Figure 1(a) provides a visual representation of the AUROC performance for both the CLM and UA-CLM methods. It shows the improvement in various uncertainty quantification metrics across three distinct LLMs, underscoring the enhanced reliability of uncertainty estimates in the generated text achieved by UA-CLM. This plot consolidates the AUROC performance metrics from the CoQA and TriviaQA datasets, with specific numbers for each dataset and each LLM provided in Table 1. This table also includes the results for the OK-VQA dataset when evaluated with LVLM, offering a comprehensive study across different datasets, models, and

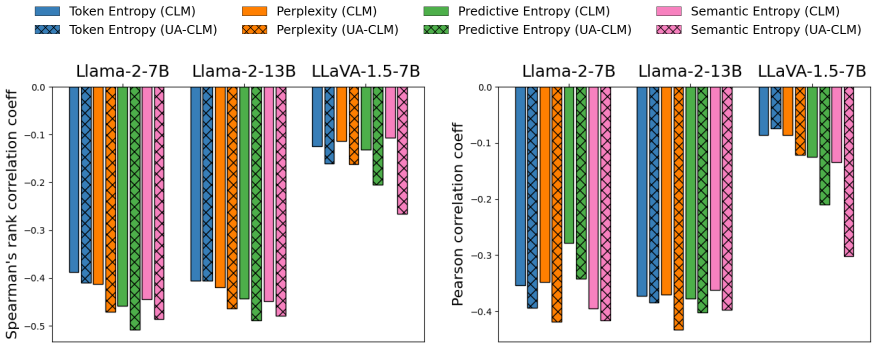


Figure 2: Uncertainty calibration analysis: Spearman's rank correlation coefficient and Pearson correlation coefficient between uncertainty estimates and generated text quality (ROUGE-L) scores for free-form open-ended question answering. Stronger negative correlation is desired for well-calibrated uncertainty quantification.

uncertainty quantification metrics. We observe that the UA-CLM method exhibits a significant improvement in hallucination detection performance of up to 17.1% on QA tasks, and upto 23.6% on VQA task, over the standard CLM method.

Notably, these enhancements with UA-CLM are achieved without compromising quality of the generated text or the overall accuracy as presented in Table 2. We refer to Appendix A.2 for details of text quality metrics, and additional evaluation metrics can be found in Appendix A.4.

Uncertainty-guided selective generation The ability of a large language model to decide when to generate a response and when to abstain from providing one, based on its uncertainty estimates is crucial for building trustworthy and reliable generative AI models. This capability enables models to recognize and communicate their limitations. Selective generation (Ren et al., 2022) plays an important role in scenarios where providing an incorrect response could have negative consequences, such as in medical diagnosis, legal advice, or safety-critical information systems. We adopt the methodology proposed by Farquhar et al. (2024) and utilize the Area Under the Accuracy-Rejection Curve (AUARC) (Nadeem et al., 2009) to evaluate both the performance of the model and the quality of uncertainty estimates in the context of selective generation. AUARC serves as a valuable metric for evaluating the quality of a model's uncertainty estimates and its decision-making ability regarding when to make predictions and when to abstain due to high uncertainty. The bar plots presented in Figure 1(b), along with the numerical data provided in Table 1 shows a significant improvement in the AUARC scores achieved by the UA-CLM. This notable enhancement in AUARC indicates the uncertainty estimates with UA-CLM's can lead to better informed downstream decision-making.

Correlation between uncertainty estimates and generated text quality Calibration serves as a mechanism for ensuring the quality of uncertainty estimates. To assess the calibration of uncertainty in free-form text generation, metrics such as Expected Calibration Error (ECE) (Naeini et al., 2015) is not feasible due to the potential mismatch in the number of tokens between the generated text and the ground truth reference. Hence we employ Spearman's rank correlation coefficient (Zwillinger & Kokoska, 1999) and Pearson correlation coefficient (Benesty et al., 2009) to evaluate the reliability of uncertainty estimates and examine how well these estimates align with the quality of the generated text, this is a novel methodology for calibration analysis to circumvent the challenges posed by varying token lengths in the generative outputs. We analyze the negative correlation between uncertainty estimates and ROUGE-L (Lin & Och, 2004) scores, a widely recognized metric for gauging generated text quality. A model with well-calibrated uncertainty estimates should demonstrate a strong negative correlation between uncertainty estimates and the generated text quality. As depicted in Figure 2, our findings from the CoQA, TriviaQA and OK-VQA datasets reveal a consistent negative correlation: the uncertainty estimates increase as the ROUGE-L scores tend to decrease, and vice-versa. The results indicate that fine-tuning with UA-CLM show a more pronounced inverse correlation between uncertainty and text quality compared to standard fine-tuning methods.

Table 3: Out-of-domain detection: Evaluation with Biomedical question answering (BioASQ) as out-of-domain dataset on Llama-2-7B finetuned with CoQA dataset. The table shows the comparison of CLM and UA-CLM with AUROC and AUPR scores for out-of-domain detection using different uncertainty metrics.

Method	AUROC ↑ (Out-of-domain detection)				AUPR ↑ (Out-of-domain detection)			
	Token Entropy	Perplexity	Predictive Entropy	Semantic Entropy	Token Entropy	Perplexity	Predictive Entropy	Semantic Entropy
CLM	0.7828	0.7661	0.7500	0.7902	0.8124	0.7928	0.8570	0.8422
UA-CLM	**0.9025**	**0.8931**	**0.7763**	**0.9061**	**0.9235**	**0.9117**	**0.8958**	**0.9250**

This demonstrates the effectiveness of uncertainty-aware fine-tuning to obtain better calibrated uncertainty in free-form text generation tasks.

Out-of-domain detection In our experiments, we assess the model's capability to identify whether a given prompt is out-of-domain, referring to a question or input that falls outside the scope of model's trained knowledge base. To quantify this ability, we employ two widely recognized metrics: the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision-Recall curve (AUPR) (Saito & Rehmsmeier, 2015). The Biomedical question-answering (BioASQ) (Krithara et al., 2023) dataset serves as our out-of-domain dataset, while the Conversational Question Answering (CoQA) dataset is used to represent in-domain data. We leverage uncertainty metrics as a means to detect out-of-domain prompts effectively. The results, as detailed in Table 3, demonstrate that our UA-CLM significantly outperforms the standard CLM in out-of-domain detection tasks. This performance is consistently observed across all four uncertainty metrics employed in the study, with up to 16.5% improvement in AUROC and up to 15% improvement in AUPR scores.

5 Discussion

We proposed a novel fine-tuning approach to improve uncertainty calibration in Large Language Models (LLMs) devised for natural language generation. Our method incorporates a differentiable uncertainty-aware causal language modeling loss, which is grounded in the principles of decision theory. This loss function is designed to enhance the model's ability to provide well-calibrated uncertainty estimates, a crucial aspect of trustworthy AI models.

Our extensive empirical evaluations on open-ended and free-form question-answering tasks has shown that the uncertainty-aware causal language modeling approach yield better-calibrated uncertainty quantification, which in turn significantly enhances the model's ability to detect hallucinations, identify out-of-domain prompts, and selective generation decisions. We also introduced a novel application of correlation analysis to the evaluation of sentence-level uncertainty calibration in free-form text generation, accounting for the varying sentence lengths between generated responses and ground-truth references.

Limitations and Future work: Currently, the proposed method is tailored to white-box model settings, where the model internals are accessible for calibration fine-tuning. However, there is a potential to extend uncertainty-aware fine-tuning to black-box models by calibrating an auxiliary model, or prompt tuning for calibrated uncertainty quantification. Additionally, the focus of this work has been on calibrating token-level uncertainty, which sets the stage for the exploration of calibrating sentence-level uncertainty. We plan to explore these two avenues in our future work. We hope this work opens new avenues for the research community to enhance the uncertainty calibration in LLMs for free-form natural language generation.

In conclusion, this work contributes towards the broader goal of developing trustworthy LLMs. The ability to recognize out-of-domain prompts and to acknowledge the limits of a model's knowledge base through reliable uncertainty quantification paves the way for reducing hallucinations and enhancing decision-making in AI systems.

REFERENCES

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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. Creating trustworthy llms: Dealing with hallucinations in healthcare ai. *arXiv preprint arXiv:2311.01463*, 2023.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. Linguistic calibration of long-form generations. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=rJVjQSQ8ye`.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pp. 37–40. Springer, 2009.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 295–302, 2020.

Eva Eigner and Thorsten Händler. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385*, 2024.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, et al. Lm-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 446–461, 2023.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 539–555, 2020.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

Gemma, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

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

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

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

Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. Calibration-tuning: Teaching large language models to know what they don't know. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pp. 1–14, 2024.

Archit Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael C Mozer, and Becca Roelofs. Soft calibration objectives for neural networks. *Advances in Neural Information Processing Systems*, 34:29768–29779, 2021.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. Calibrated language model fine-tuning for in-and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1326–1340, 2020.

Ranganath Krishnan and Omesh Tickoo. Improving model calibration with accuracy versus uncertainty optimization. *Advances in Neural Information Processing Systems*, 33:18237–18248, 2020.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10 (1):170, 2023.

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

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, pp. 623–631. PMLR, 2017.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pp. 2805–2814. PMLR, 2018.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04)*, pp. 605–612, 2004.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

Xin Liu, Muhammad Khalifa, and Lu Wang. Litcab: Lightweight language model calibration over short-and long-form responses. In *The Twelfth International Conference on Learning Representations*, 2024b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. Hallucination-free? assessing the reliability of leading ai legal research tools. *arXiv preprint arXiv:2405.20362*, 2024.

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

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`, 2022.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694, 2021.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299, 2020.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In *Machine Learning in Systems Biology*, pp. 65–81. PMLR, 2009.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR workshops*, volume 2, 2019.

Benjamin Plaut, Khanh Nguyen, and Tu Trinh. Softmax probabilities (mostly) predict large language model correctness on multiple-choice q&a. *arXiv preprint arXiv:2402.13213*, 2024.

Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*, 2022.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan, and Jianfeng Gao. Adamix: Mixture-of-adaptations for parameter-efficient model tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5744–5760, 2022.

Yijun Xiao and William Yang Wang. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 7322–7329, 2019.

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

Haoyan Yang, Yixuan Wang, Xingyin Xu, Hanyuan Zhang, and Yirong Bian. Can we trust llms? mitigate overconfidence bias in llms through knowledge transfer. *arXiv preprint arXiv:2405.16856*, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

Daniel Zwillinger and Stephen Kokoska. *CRC standard probability and statistics tables and formulae*. Crc Press, 1999.

ENHANCING TRUST IN LARGE LANGUAGE MODELS WITH UNCERTAINTY-AWARE FINE-TUNING

A APPENDIX

A.1 EXPERIMENTAL DETAILS

A.1.1 DATASETS

CoQA Conversational Question Answering (CoQA) (Reddy et al., 2019) dataset was developed to evaluate models' ability to respond to natural, dialogue-based questions, with free-form text answers supported by highlighted evidence from the passage. The full dataset comprises of 127k question-answer pairs derived from 8k conversations based on text passages across 7 distinct domains. For all our experiments, we utilize the development subset of CoQA, which consists of 8k question-answer pairs. Figure 3 shows the color-coded co-reference chains in CoQA as illustrated in the (Reddy et al., 2019).

TriviaQA TriviaQA (Joshi et al., 2017) is a reading comprehension dataset consisting of over 650k question-answer-evidence triplets. It includes 95,000 question-answer pairs authored by trivia enthusiasts, along with an average of six independently gathered evidence documents per question, providing high-quality distant supervision for answering the questions. In our experiment, we used the validation split of the dataset with around 10,000 question-answer pairs. Table 4 shows some of the samples from the dataset.

OK-VQA Outside Knowledge-Visual Question Answering benchmarks (Marino et al., 2019) consists of visual queries where the image content alone is not sufficient to answer the questions. Thus, it requires models to incorporate external knowledge to generate accurate answers. The dataset consists of 14k questions across 10 knowledge categories. In our experiment, we used the validation split of the dataset with around 5k question-answer pairs. Figure 4 shows a few samples from the dataset across different knowledge categories.

The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Q_1: What are the candidates **running** for?
A_1: Governor
R_1: The Virginia governor's race

Q_2: **Where**?
A_2: Virginia
R_2: The Virginia governor's race

Q_3: Who is the democratic candidate?
A_3: **Terry McAuliffe**
R_3: Democrat Terry McAuliffe

Q_4: Who is **his** opponent?
A_4: **Ken Cuccinelli**
R_4 Republican Ken Cuccinelli

Q_5: What party does **he** belong to?
A_5: Republican
R_5: Republican Ken Cuccinelli

Q_6: Which of **them** is winning?
A_6: Terry McAuliffe
R_6: Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May

Figure 3: Sample from CoQA (Reddy et al., 2019) illustrating the co-reference chain of conversational questions.

Question	Answer
Miami Beach in Florida borders which ocean?	Atlantic
What was the occupation of Lovely Rita according to the song by the Beatles	Traffic Warden
Who was Poopdeck Pappys most famous son?	Popeye
The Nazi regime was Germany's Third Reich; which was the first Reich?	HOLY ROMAN EMPIRE

Table 4: Data samples from TriviaQA (Joshi et al., 2017)

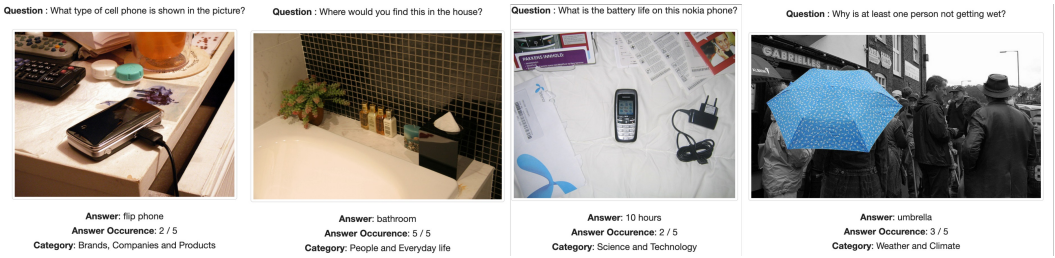


Figure 4: Data samples from OK-VQA (Marino et al., 2019) across different knowledge categories.

BioASQ The BioASQ (Krithara et al., 2023) challenge, conducted every year, focuses on techniques in large-scale biomedical semantic indexing and question answering (QA). For our experiments, we utilize Task B (Table 5) from the eleventh edition of the BioASQ challenge (BioASQ 2023), which includes biomedical questions in English and their corresponding gold standard answers. We consider *exact answers* as gold answers where available; otherwise, we refer to the *ideal answers* field in the dataset.

Question	Answer
Which amino acid in implicated in the Blue diaper syndrome?	tryptophan
What are the outcomes of ubiquitination?	Protein degradation, Degradation of proteins
What causes Serpentine Supravenous Hyperpigmentation?	5-fluorouracil, docetaxel
What are positive cell-cycle regulators that can cause cancer when mutated called?	Proto-oncogenes

Table 5: Data samples from BioASQ (Krithara et al., 2023)

A.1.2 Open-Book QA Prompt

Prompt:

> Answer the following question as briefly as possible.
> Context: [Provided context paragraph]
> Question: [Associated Question]
> Answer:

A.1.3 Finetuning hyperparameters and Implementation

We fine-tune our models for all experiments for 3 epochs using LoRA (Hu et al., 2022) with AdamW optmizer (Loshchilov & Hutter, 2019). We use an initial learning rate of $1e$-4, weight decay of 0.001 and a warm up ratio of 0.03. In our experiments we used Low-Rank Adaptation (LoRA) to efficiently fine-tune pre-trained LLMs and LVLMs for the causal language modeling task. For LLMs, we set the LoRA rank as 32, alpha parameter as 64 and a dropout of 0.1. LoRA was applied specifically to the following modules: *q_proj*, *k_proj*, *v_proj*, *up_proj*, and *down_proj*. In addition to LoRA, we applied 4-bit normalized float (*nf4*) quantization to the model's parameters and utilized *FP16* precision during fine-tuning to reduce the computational overhead.

For inference, we utilized *FP16* precision and the default greedy decoding provided by Hugging Face with temperature value T=0.3. The predictive entropy and semantic entropy are estimated by generating 5 stochastic sequences from the model, each obtained through temperature sampling with a temperature setting of T=0.3.This temperature was chosen to obtain optimal uncertainty estimates balanced with high quality generated text, based on the ablation study shown in Figure 5. Our source

code was implemented using Pytorch [1] framework and the models from Hugging Face [2] library. We will make the source code available to the community for reproducing the results.

For our LVLM model, LLaVA-1.5 (Liu et al., 2024a), we configured LoRA with a rank of 8, an alpha value of 8, and applied a 0.1 dropout rate to mitigate overfitting on the small OK-VQA training subset. In addition to the proposed UA-CLM loss, we experimented with a combined loss function that anneals the CLM loss with our UA-CLM loss. This approach allows the model to learn to answer OK-VQA queries using the context provided in the early stages of training, without uncertainty calibration. As training progresses, we shift our focus toward calibrating the model's uncertainty. By this stage, the model has already learned to answer visual question-answering prompts, allowing us to refine its performance on questions it is likely to answer correctly or incorrectly, based on insights gained during the initial training phases. Specifically, we assign a higher weight to the CLM loss in the early stages of training, gradually increasing the weight of the UA-CLM loss after 20% of the training is completed as shown in Equation 4. Our ablation results for this experiment are presented in Table 8.

$$\mathcal{L} = \mathcal{L}_{\text{CLM}} + \beta \cdot \mathcal{L}_{\text{UA-CLM}} \quad \text{where } \beta = \begin{cases} 0.2 & \text{if steps} \leq 0.2 \cdot \text{total_steps} \\ 0.8 & \text{if steps} > 0.2 \cdot \text{total_steps} \end{cases} \tag{4}$$

A.2 Text generation quality metrics

- **ROUGE-L (Lin & Och, 2004):** Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a widely-used evaluation metric for assessing the quality of text generated based on n-gram matching. We use the Rouge-L variant which uses the longest common subsequence between the generated answer and the ground truth answer.
- **Exact Match (EM):** Exact Match (EM) metric is a stringent evaluation criterion used to assess the performance of models on tasks such as question answering (QA), where a generated response is compared to a reference answer. It is a widely used metric for open-book QA, this metric evaluates a model's ability to extract the precise text span from the context to answer a question.
- **Accuracy:** The generated answer is considered as accurate if it achieves Rouge-L$(y, \hat{y}) > 0.3$, for a given reference answer y and a model generation $\hat{y}$. We follow this criterion for quantifying accuracy in free-form text generation based on the findings from (Kuhn et al., 2023) that demonstrated this criterion closely matches the human evaluation accuracy on COQA and TriviaQA datasets, both of which are utilized in our experiments.
- **BERTScore (Zhang et al., 2020):** BERTScore utilizes word embeddings to compute a similarity score between the tokens in the prediction and ground truth and has shown to well correlate with human judgement. We report Precision, Recall and F1 BERTScores for all our experiments.

A.3 Uncertainty estimation metrics

We assess uncertainty in natural language predictions by utilizing the Area Under the Receiver Operating Characteristic (AUROC) scores, calculated between correct and incorrect predictions across the following metrics:

- **Predictive Entropy** Fomicheva et al. (2020): This is a widely used measure for uncertainty estimation and is defined as the entropy of the model's output probability distribution from stochastic generated responses. Formally, for a specific instance x, the predictive entropy, denoted as $P_E(x)$, is defined as the conditional entropy of the output random variable Y, with realization y, given x (Kuhn et al., 2023): $P_E(x) = H(Y|x) = -\int p(y|x) \ln p(y|x) dy$
- **Semantic Entropy** (Kuhn et al., 2023): Defined as entropy of output distributions in semantic event-space rather than traditional token event-space and has been shown to be a good indicator in detecting confabulation in language models.

[1] https://pytorch.org/
[2] https://huggingface.co/

Table 6: Evaluation of generated text quality metrics: Comparative analysis of Causal Language Modeling (CLM) and Uncertainty-aware Causal Language Modeling (UA-CLM) fine-tuning methods. The results in the table indicate that UA-CLM achievies similar or better generated text quality metrics than standard CLM across a range of models and datasets.

Dataset	Model	Finetuning Method	Rouge-L	Exact Match	Accuracy	BERT Score (Precision)	BERT Score (Recall)	BERT Score (F1)
CoQA	Llama-2-7b	CLM	0.8886	0.8071	0.9253	0.9633	0.9598	0.9604
		UA-CLM	0.8882	0.8027	0.9264	0.9671	0.9644	0.9648
	Llama-2-13b	CLM	0.9106	0.8434	0.9406	0.9678	0.9639	0.9650
		UA-CLM	0.9118	0.8204	0.9461	0.9732	0.9698	0.9705
	Gemma-2b	CLM	0.8654	0.7606	0.9143	0.962	0.9548	0.9570
		UA-CLM	0.8632	0.7632	0.9088	0.9627	0.9554	0.9578
TriviaQA	Llama-2-7b	CLM	0.5867	0.4939	0.6385	0.8743	0.8785	0.8754
		UA-CLM	0.6342	0.5627	0.6754	0.8951	0.8883	0.8910
	Llama-2-13b	CLM	0.6588	0.5883	0.6967	0.9026	0.8989	0.9001
		UA-CLM	0.7277	0.6445	0.7710	0.9204	0.9164	0.9177
	Gemma-2b	CLM	0.4349	0.3674	0.4759	0.8375	0.8349	0.8355
		UA-CLM	0.4563	0.3915	0.4959	0.8404	0.8382	0.8387
OK-VQA	Llava-1.5-7b	CLM	0.5569	0.5099	0.5891	0.8897	0.8864	0.8877
		UA-CLM	0.5354	0.4950	0.5643	0.8841	0.8820	0.8827

Table 7: Uncertainty calibration analysis: The results show UA-CLM have more pronounced negative correlation between the uncertainty estimates and the generated text quality (ROUGE-L) than standard Causal Language Modeling CLM, indicating enhanced reliability in uncertainty quantification with UA-CLM.

Dataset	Model	Finetuning Method	Spearman's rank correlation coefficient ↓				Pearson correlation coefficient ↓			
			Token Entropy	Perplexity	Predictive Entropy	Semantic Entropy	Token Entropy	Perplexity	Predictive Entropy	Semantic Entropy
CoQA	Llama-2-7b	CLM	-0.2130	-0.2379	-0.3398	-0.2898	-0.2029	-0.2109	-0.2710	-0.2881
		UA-CLM	**-0.2479**	**-0.3401**	**-0.4334**	**-0.3742**	**-0.3414**	**-0.3414**	**-0.3414**	**-0.3414**
	Llama-2-13b	CLM	-0.2325	-0.2523	-0.3253	-0.3004	-0.2302	-0.2495	-0.3001	-0.2636
		UA-CLM	**-0.2398**	**-0.3280**	**-0.4170**	**-0.3717**	**-0.2335**	**-0.3244**	**-0.3269**	**-0.3481**
	Gemma-2b	CLM	-0.3639	-0.3629	-0.4335	-0.3756	-0.3860	-0.3713	-0.3483	-0.3399
		UA-CLM	**-0.3676**	**-0.4063**	**-0.4476**	**-0.4127**	**-0.4033**	**-0.4019**	**-0.3517**	**-0.3530**
TriviaQA	Llama-2-7b	CLM	-0.5627	-0.5863	-0.5765	-0.5994	-0.5047	-0.4854	-0.2864	-0.5020
		UA-CLM	**-0.5713**	**-0.6011**	**-0.5822**	-0.5980	**-0.5385**	**-0.5326**	**-0.3382**	-0.4916
	Llama-2-13b	CLM	-0.5711	-0.5845	-0.5522	-0.5959	-0.5155	-0.4915	-0.4548	-0.4612
		UA-CLM	**-0.5725**	**-0.5862**	**-0.5607**	-0.5854	**-0.5362**	**-0.5407**	**-0.4786**	-0.4479
	Gemma-2b	CLM	-0.5636	-0.5772	-0.5609	-0.5537	-0.5020	-0.4534	-0.4494	-0.4514
		UA-CLM	-0.5623	**-0.5913**	-0.5457	**-0.5928**	**-0.5164**	**-0.5010**	**-0.4534**	**-0.4947**
OK-VQA	Llava-1.5-7b	CLM	-0.1253	-0.1132	-0.1320	-0.1062	-0.0862	-0.0861	-0.1256	-0.1340
		UA-CLM	**-0.1606**	**-0.1619**	**-0.2050**	**-0.2660**	-0.0748	**-0.1214**	**-0.2100**	**-0.3020**

- **Perplexity** Fomicheva et al. (2020): A standard metric to assess the quality of model and is defined as the inverse probability of the generated text: Perplexity $= \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log_2 p(w_i|w_1,\ldots,w_{i-1})\right)$

A.4 Additional Results

The results in the Table 6 presents a detailed quantitative evaluation of various text generation quality metrics across various models, datasets, and uncertainty quantification (UQ) metrics. It compares standard Causal Language Modeling (CLM) with our Uncertainty-Aware Causal Language Modeling (UA-CLM).

The results in Table 7 presents quantitative data with the values of Spearman's rank correlation coefficient and Pearson correlation coefficient across different models, datasets, and uncertainty quantification (UQ) metrics, with a specific focus on comparing standard Causal Language Modeling (CLM) and our Uncertainty-Aware Causal Language Modeling (UA-CLM). The data reveals that UA-CLM exhibits a stronger inverse correlation between UQ metrics and ROUGE-L scores, indi-

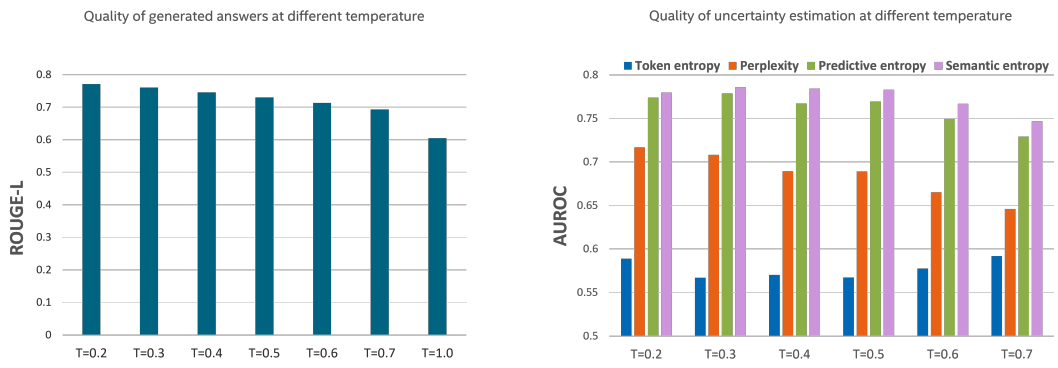


Figure 5: Ablation study: Effect of temperature value on the quality of generated text and the quality of uncertainty estimates evaluated with AUROC for hallucination detection. The study was performed on pre-trained Llama-2-7B model with CoQA dataset. Based on this study, we selected temperature T=0.3 as it results in optimal AUROC and ROUGE-L scores.

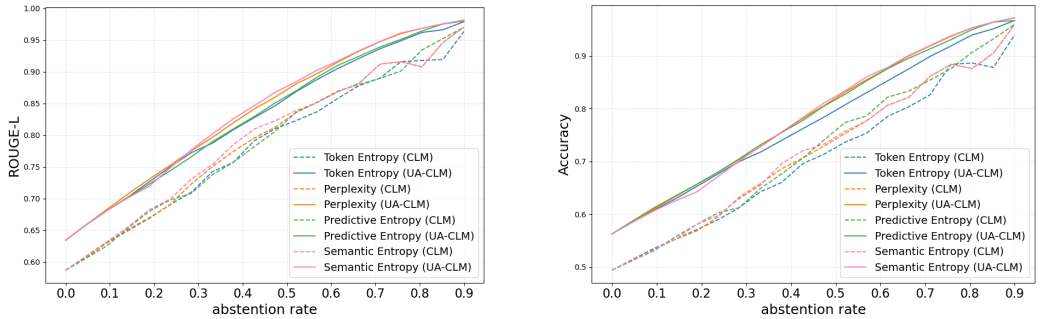


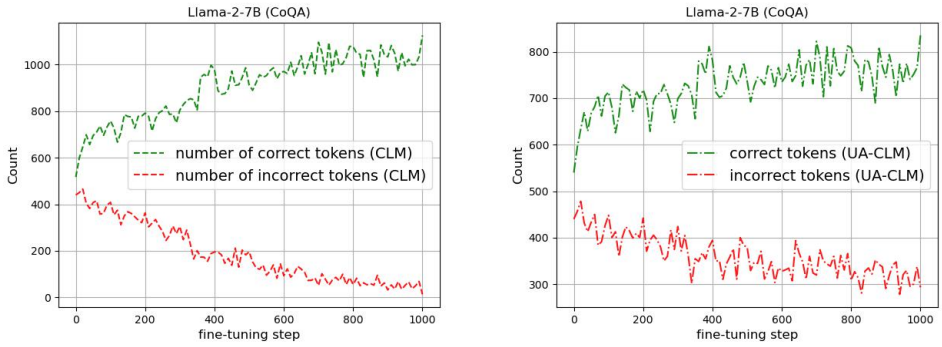
Figure 6: Selective generation (Llama-2-7B/TriviaQA)

cating better reliability of uncertainty estimates. This enhanced inverse relationship suggests that UA-CLM is more adept at associating higher uncertainty with low quality text generation quality and vice versa, which is a key indicator of better uncertainty calibration.

Table 8: Ablation study: Effect of different loss functions during fine-tuning

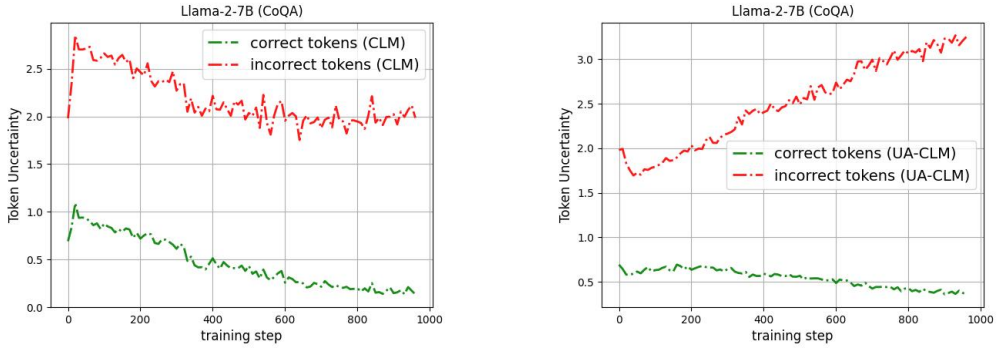
Dataset	Model	Fine-tuning Loss	AUROC (Hallucination/Confabulation detection)				AUARC (Area under rejection accuracy curve)			
			Token Entropy	Perplexity	Predictive Entropy	Semantic Entropy	Token Entropy	Perplexity	Predictive Entropy	Semantic Entropy
OKVQA	Llava-1.5-7b	$\mathcal{L}_{\text{CLM}}$	0.5504	0.5419	0.5455	0.537	0.5809	0.5781	0.579	0.5747
		$\mathcal{L}_{\text{UA-CLM}}$	0.5839	0.6032	0.5701	0.6727	0.5657	0.5771	0.5601	0.6028
		$\mathcal{L}_{\text{CLM}} + \beta * \mathcal{L}_{\text{UA-CLM}}$	0.6001	0.5984	0.6106	0.6638	0.5989	0.5965	0.6012	0.6265
CoQA	Llama-2-7b	$\mathcal{L}_{\text{CLM}}$	0.6252	0.632	0.6635	0.6889	0.823	0.829	0.8516	0.8405
		$\mathcal{L}_{\text{UA-CLM}}$	0.6955	0.7398	0.7413	0.7741	0.8246	0.8477	0.8743	0.8571
		$\mathcal{L}_{\text{CLM}} + \beta * \mathcal{L}_{\text{UA-CLM}}$	0.6101	0.6183	0.6978	0.7252	0.8153	0.8153	0.8614	0.8455
TriviaQA	Llama-2-13b	$\mathcal{L}_{\text{CLM}}$	0.8264	0.8333	0.7971	0.8407	0.7464	0.7526	0.7532	0.7556
		$\mathcal{L}_{\text{UA-CLM}}$	0.8297	0.8352	0.8033	0.8447	0.7960	0.8059	0.804	0.8069
		$\mathcal{L}_{\text{CLM}} + \beta * \mathcal{L}_{\text{UA-CLM}}$	0.8340	0.8263	0.8049	0.8307	0.7666	0.7692	0.7673	0.7693

Figure 6 shows results on selective generation, based on varying levels of abstaining from providing generated response informed by uncertainty estimates. We plotted both ROUGE-L scores and accuracy as functions of the abstention rate, showing how the models perform as they increasingly withhold responses in situations of high uncertainty. The plots clearly shows that the UA-CLM outperforms CLM across all the four uncertainty metrics.



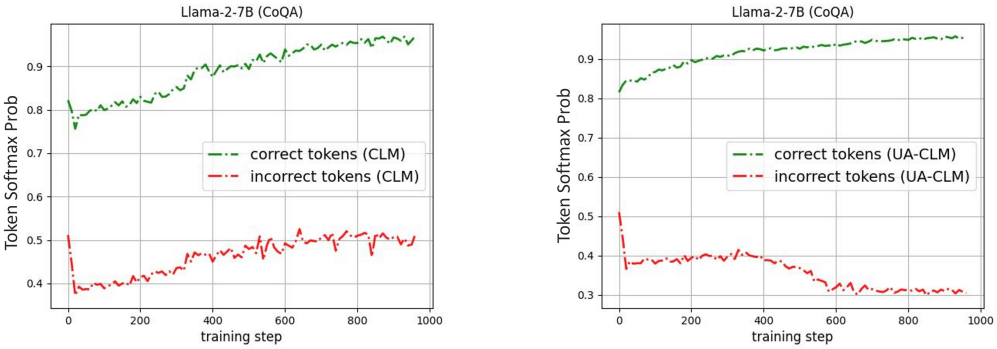
(a) CLM (b) UA-CLM

Figure 7: Analysis of Correct and Incorrect Token Counts in mini-batch during fine-tuning with CLM and UA-CLM. Both CLM and UA-CLM show increase in correct tokens and a decrease in incorrect tokens as fine-tuning progresses.



(a) CLM (b) UA-CLM

Figure 8: Analysis of Token Uncertainty associated with Correct and Incorrect tokens in the mini-batch during fine-tuning with CLM and UA-CLM. A well-calibrated model should provide low uncertainty for correct tokens and higher uncertainty for incorrect tokens. With standard CLM Loss, uncertainty for both correct and incorrect tokens decreases, indicating overconfidence even on incorrect tokens. In contract, with UA-CLM, the uncertainty for incorrect tokens increases and the decreasing uncertainty on correct tokens, supporting that the fine-tuning with UA-CLM improves the reliability of uncertainty estimates.



(a) CLM (b) UA-CLM

Figure 9: Analysis of Token Softmax Probability associated with Correct and Incorrect tokens during fine-tuning with CLM and UA-CLM. A well-calibrated model should assign high probability to correct tokens and lower probability to incorrect tokens. With standard CLM loss, probabilities for both correct and incorrect tokens increase as fine-tuning progress, indicating overconfidence. In contrast, UA-CLM fine-tuning results in higher probabilities for correct tokens and lower probabilities for incorrect tokens, enhancing the reliability of token probability scores

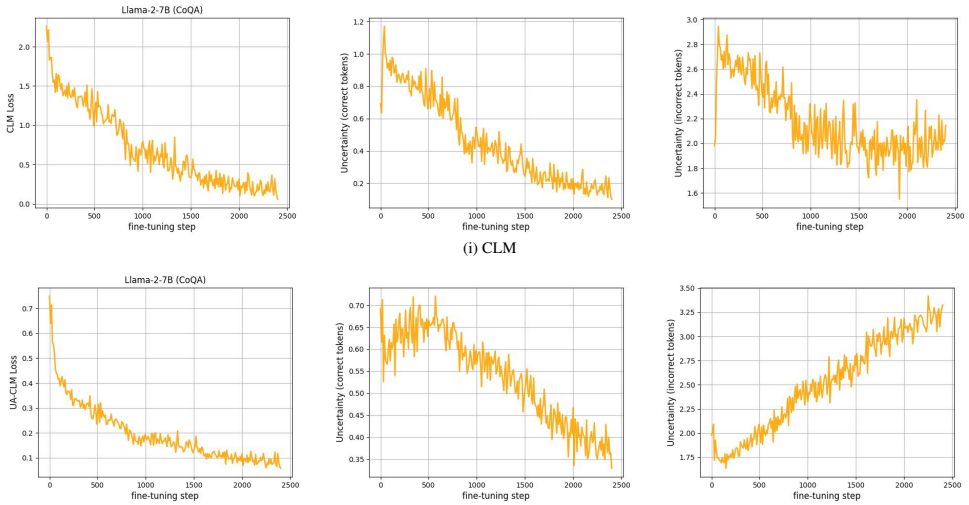


Figure 10: Llama-2-7B: Loss convergence and uncertainty values associated with correct and incorrect tokens.

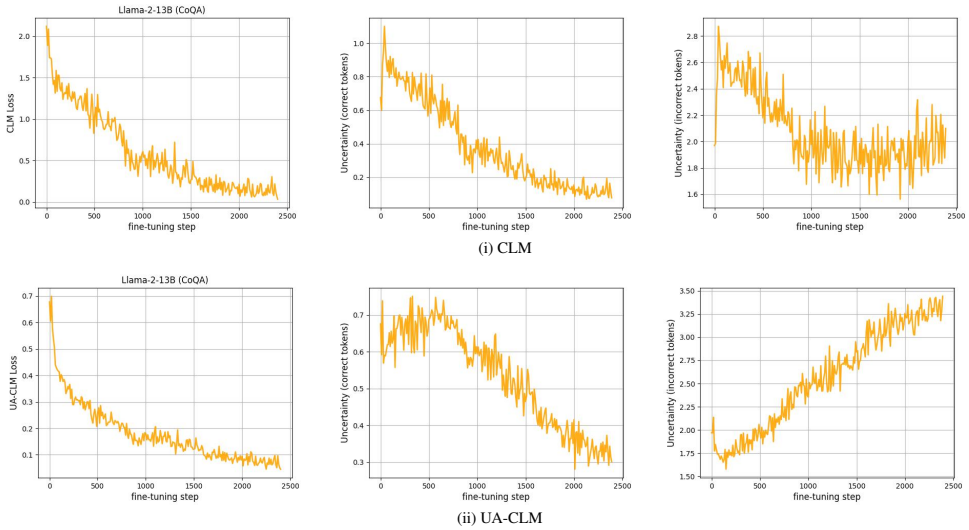


Figure 11: Llama-2-13B: Loss convergence and uncertainty values for correct and incorrect tokens.

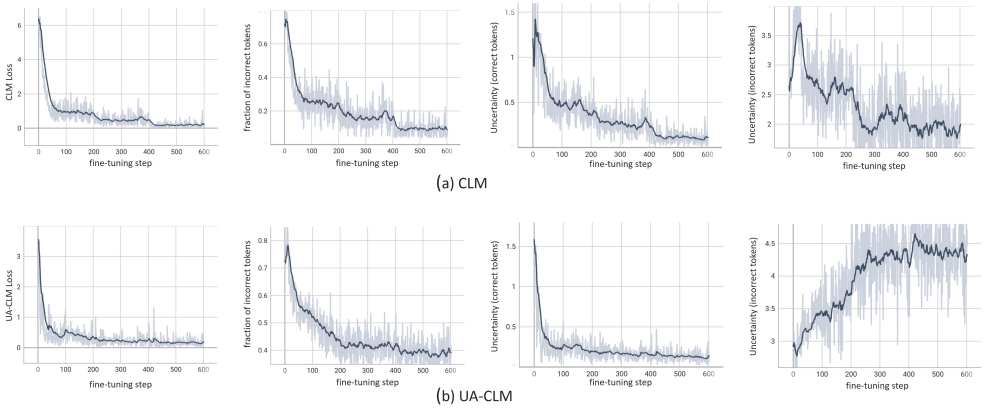


Figure 12: Llava-1.5: Loss convergence and uncertainty values associated with correct and incorrect tokens.

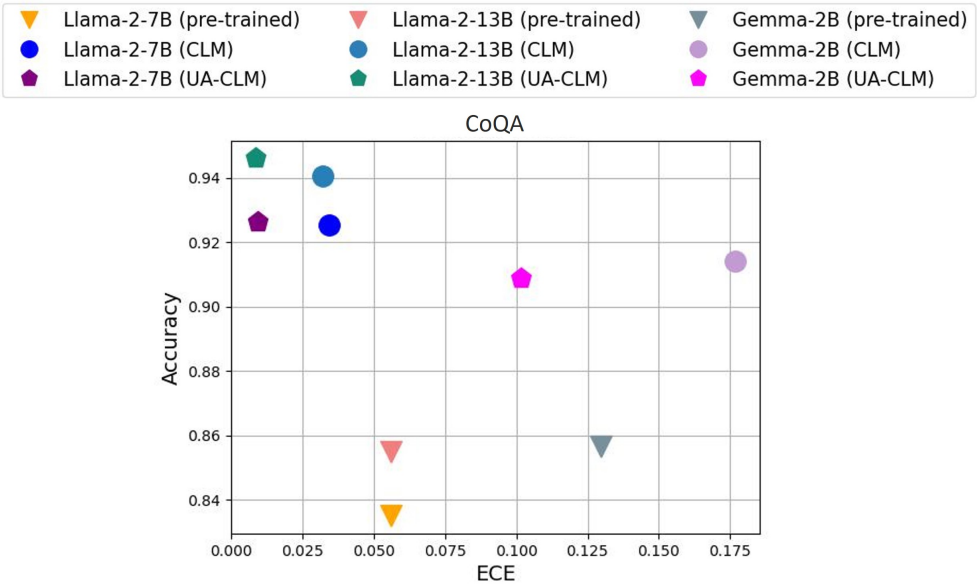


Figure 13: Accuracy versus Expected Calibration Error (ECE) comparison between UA-CLM, CLM, and pre-trained baseline across different LLM architectures on CoQA dataset. The ideal model should have high accuracy and low expected calibration error, indicating accurate predictions with well-calibrated uncertainty quantification (top-left of the Accuracy vs ECE plot). When evaluating three different model architectures, we observe that the accuracy of models with CLM and UA-CLM remains within a similar range and better than the pre-trained baseline. While, the ECE of models fine-tuned with UA-CLM shows significant improvement compared to both the pre-trained baseline and CLM fine-tuning.

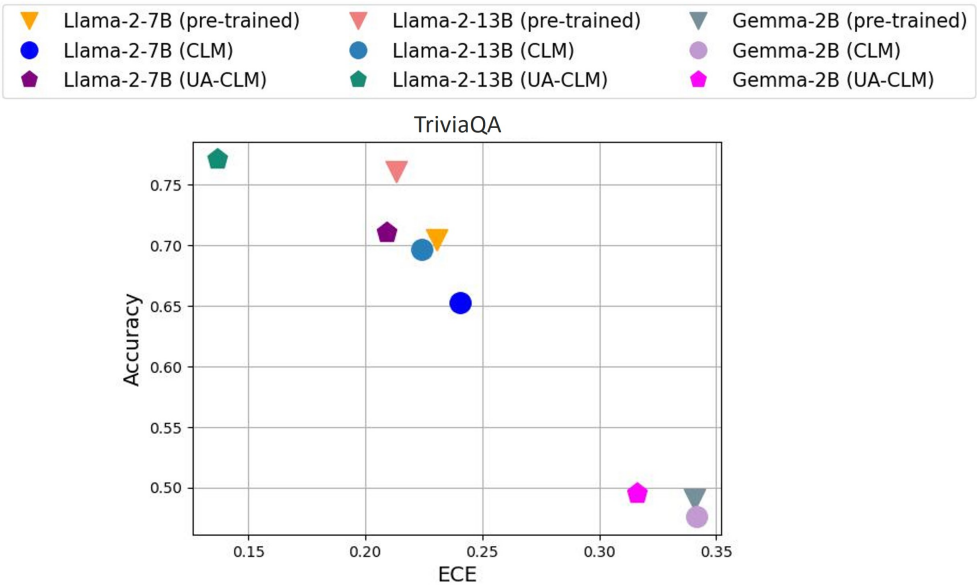


Figure 14: Accuracy versus Expected Calibration Error (ECE) comparison between UA-CLM, CLM, and pre-trained baseline across different LLM architectures on TriviaQA dataset. The ideal model should have high accuracy and low expected calibration error, indicating accurate predictions with well-calibrated uncertainty quantification (top-left of the Accuracy vs ECE plot). When evaluating three different model architectures, we observe that the both accuracy and ECE of the models fine-tuned with UA-CLM shows significant improvement compared to both the pre-trained baseline and CLM fine-tuning.