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 4 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 5 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 5 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 4: 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 5: Data samples from TriviaQA (Joshi et al., 2017)

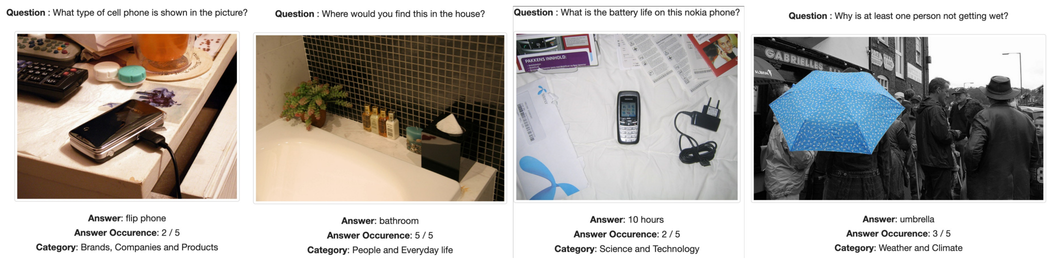


Figure 5: Data samples from OK-VQA (Marino et al., 2019) across different knowledge categories.

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 6) 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 6: 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 6. 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 9.

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

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 7: 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 achieves 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 8: 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 7 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 8 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-

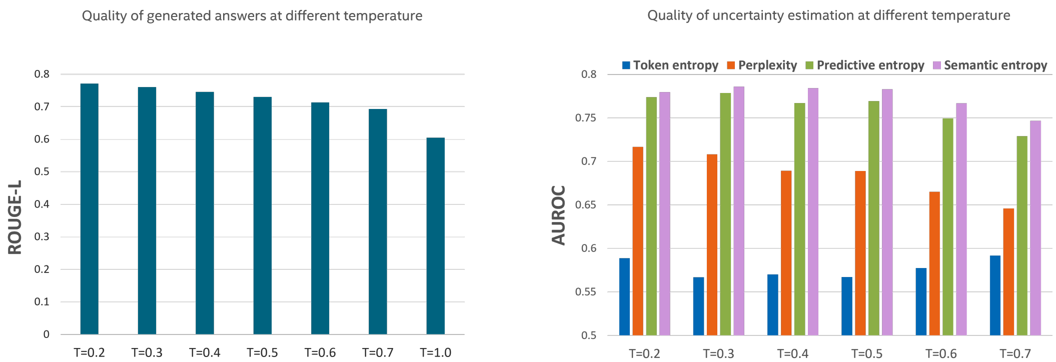


Figure 6: Ablation study: Effect of temperature value on the quality of generated text and the quality of uncertainty estimates evaluated with AUROC for hallucination detection. The study was performed on pre-trained Llama-2-7B model with CoQA dataset. Based on this study, we selected temperature T=0.3 as it results in optimal AUROC and ROUGE-L scores.

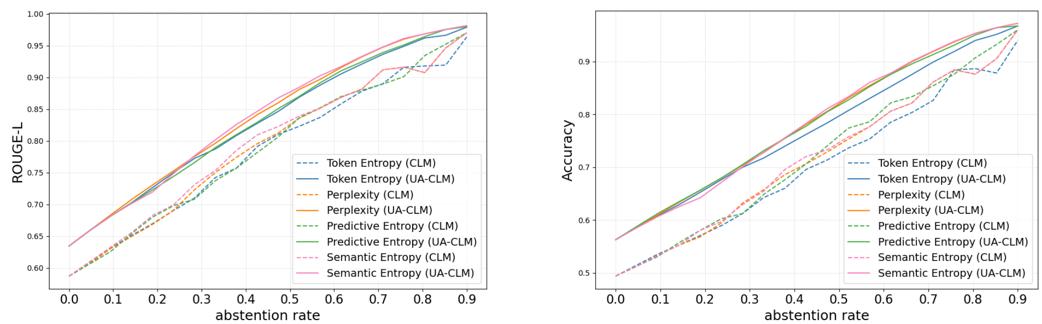


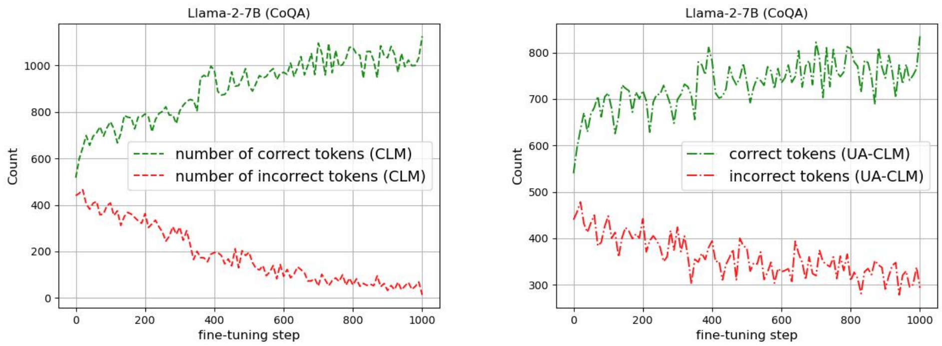
Figure 7: Selective generation (Llama-2-7B/TriviaQA)

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 9: Ablation study: Effect of different loss functions during fine-tuning. Exact match is used as accuracy metric in computing AUARC.

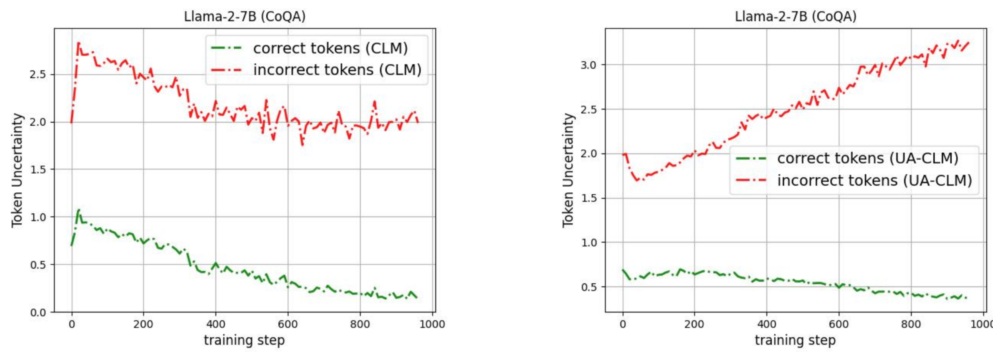
Dataset	Model	Fine-tuning Loss	AUROC (Hallucination/Confabulation detection)				AUARC (Area under rejection accuracy curve)			
			Token Entropy	Perplexity	Predictive Entropy	Semantic Entropy	Token Entropy	Perplexity	Predictive Entropy	Semantic Entropy
OKVQA	Llava-1.5-7b	$\mathcal{L}_{\text{CLM}}$	0.5504	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 7 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.



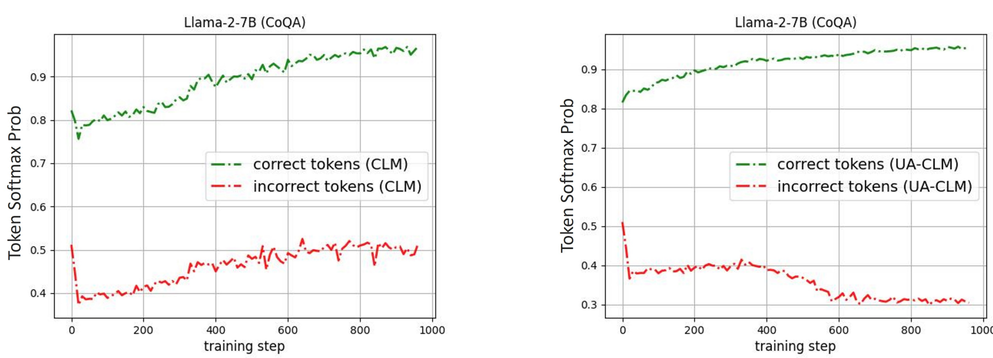
(a) CLM (b) UA-CLM

Figure 8: 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.



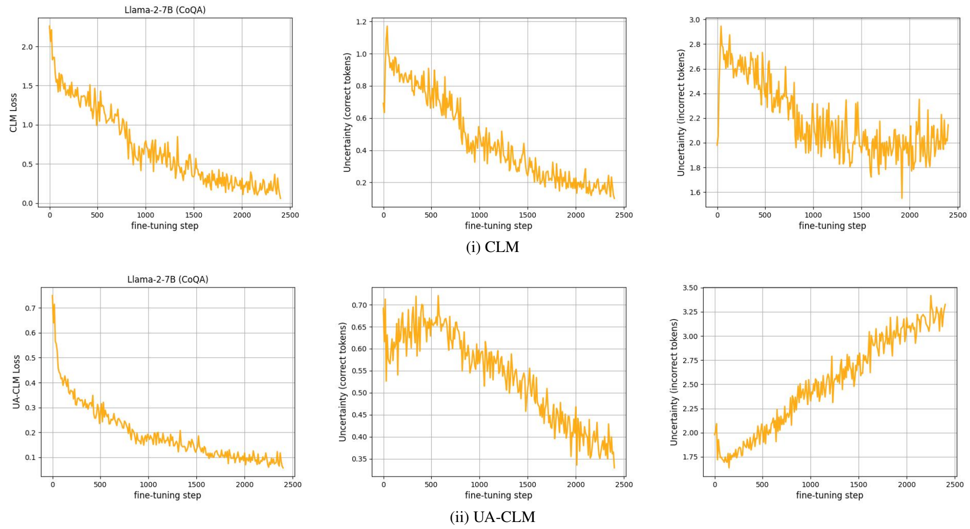
(a) CLM (b) UA-CLM

Figure 9: 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

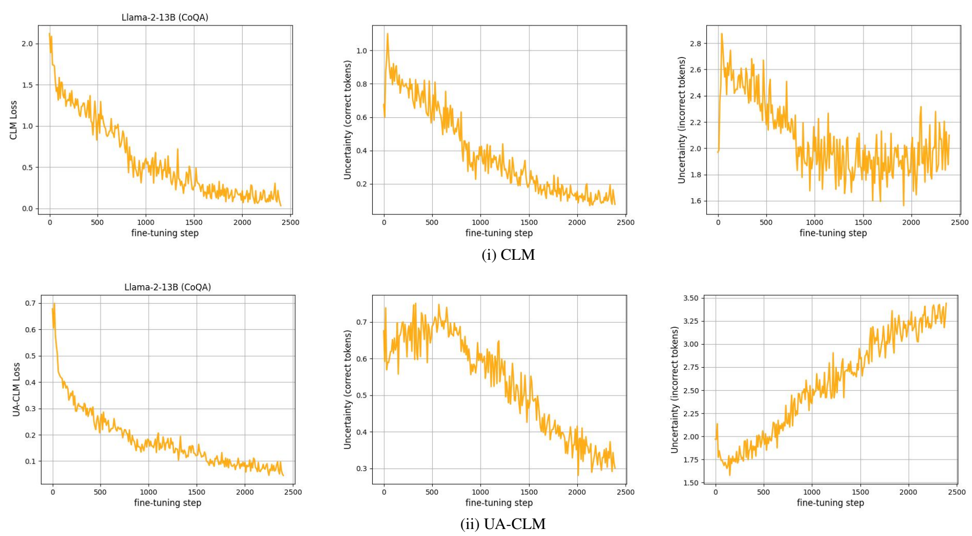
Figure 10: Analysis of Token Softmax Probability associated with Correct and Incorrect tokens during fine-tuning with CLM and UA-CLM. A well-calibrated model should assign high probability to correct tokens and lower probability to incorrect tokens. With standard CLM loss, probabilities for both correct 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



(i) CLM

(ii) UA-CLM

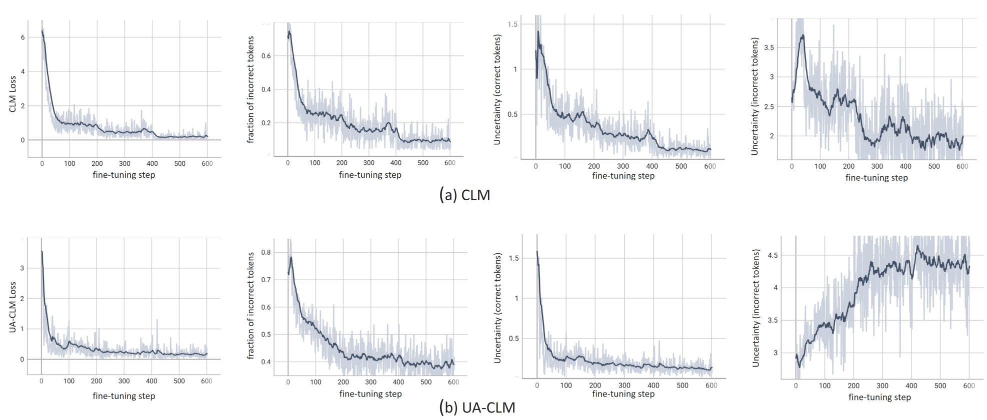
Figure 11: Llama-2-7B: Loss convergence and uncertainty values associated with correct and incorrect tokens.



(i) CLM

(ii) UA-CLM

Figure 12: Llama-2-13B: Loss convergence and uncertainty values for correct and incorrect tokens.



(a) CLM

(b) UA-CLM

Figure 13: Llava-1.5: Loss convergence and uncertainty values associated with correct and incorrect tokens.

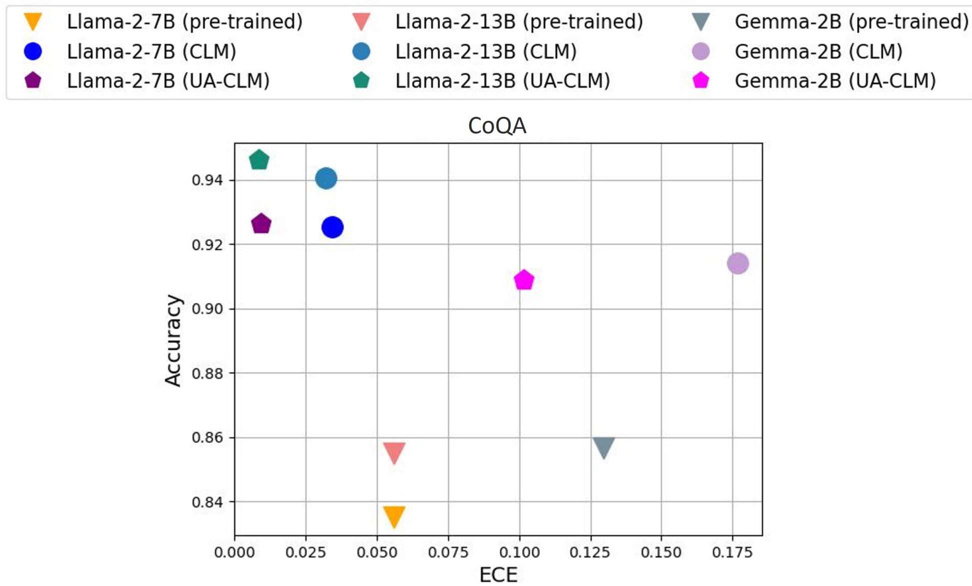


Figure 14: Accuracy versus Expected Calibration Error (ECE) comparison between UA-CLM, CLM, and pre-trained baseline across different LLM architectures on CoQA dataset. The ideal model should have high accuracy and low expected calibration error, indicating accurate predictions with well-calibrated uncertainty 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.

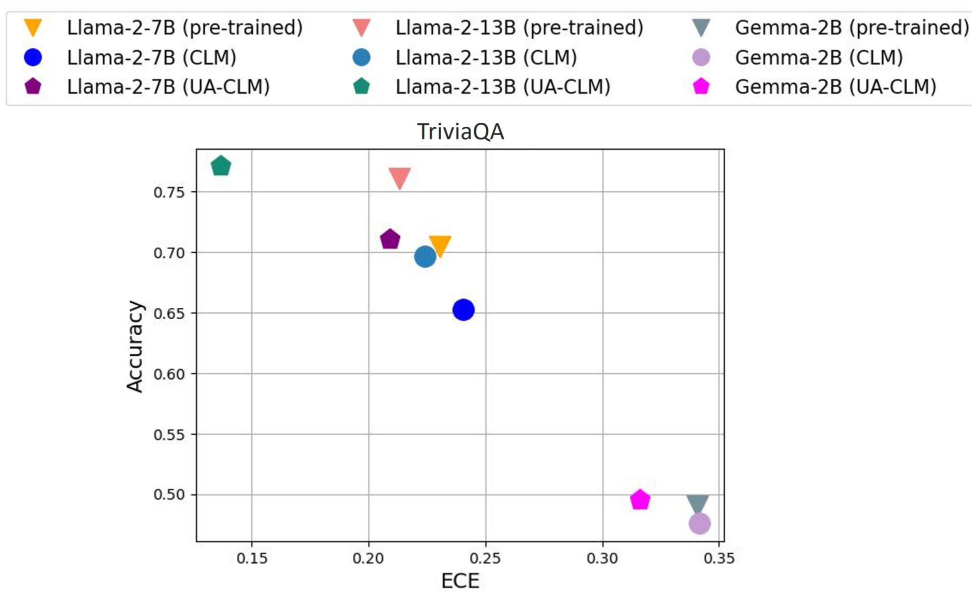


Figure 15: Accuracy versus Expected Calibration Error (ECE) comparison between UA-CLM, CLM, and pre-trained baseline across different LLM architectures on TriviaQA dataset. The ideal model should have high accuracy and low expected calibration error, indicating accurate predictions with well-calibrated 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.