Confidence Calibration in Large Language Models for Uncertainty Quantification: Affecting Calibration with Conditional Weight Updates

Sophia Somers, Edward Kim
Drexel University
ss5837@drexel.edu, ek826@drexel.edu

Abstract

In any medical applications of Large Language Models (LLMs), it is critical to have accurate uncertainty quantification, as well as control over the over- and under-confidence of the model. Current fine-tuning (FT) methods lack this control, partly because they fail to account for the fact that repeated exposure to a fact does not make it more correct. We propose a revised FT method that updates model weights only when the model does not sufficiently "know" an answer.

We fine-tuned Meta's Llama-3.2, 1B parameter model on the MMLU multiple-choice dataset using traditional FT methods for a Control Model and Conditional Update FT for an Experimental Model.

The tuned models showed different results, with the Control showing greater overconfidence and the Experimental Model showing greater under-confidence as compared to the Base Model. Additionally, the Experimental Model showed a more even distribution of confidence scores, which is advantageous for post-calibration.

This method for affecting confidence calibration while fine-tuning LLMs may potentially help in the broader challenge of creating reliable and trustworthy LLMs.

Introduction

Large Language Models are widely used for tasks like text generation, chatbots and question-answering—in such roles, it is crucial that models are trustworthy, especially in the high-stakes field of clinical applications. Trust can only be achieved when the true statistical accuracy of an LLM response can be known. Furthermore, ideal model 'personality' may vary by topic: topics with grey areas like health-related ethics, rare disease identification, and clinical decision support may require more tentative responses from a model; even topics that are universally agreed upon such as arithmetic require answers with a confidence that is consistent with the true accuracy of the model.

To achieve this ideal of trust, we must first achieve confidence calibration—when an LLM's confidence in its own response consistently matches the accuracy of the response. One way to visualize this relationship is through a calibration curve, showing the accuracy across confidence intervals, with perfect calibration when the curve follows the accuracy = confidence line.

While LLM base models tend to have relatively good calibration, the calibration has been shown to deteriorate as the model is fine-tuned and aligned. The GPT-4 Technical report from OpenAI demonstrated this issue with the calibration curve being almost perfect for the base GPT-4 model when evaluated on a subset of the MMLU dataset, but "post-training hurt calibration significantly" (Achiam et al. 2023). Later papers showed this confidence calibration deterioration in other models and methods of FT, with a tendency for overconfidence. Zhu et al. (2023) demonstrated that Pythia models, LLaMA models, FLAN-T5 and OPT trained on the PILE, T-REx or MMLU datasets became consistently overconfident across model sizes and training dynamics. Currently, post-calibration methods exist to correct the calibration curve after the training is complete, such as Histogram binning (Zadrozny & Elkan 2001), temperature scaling (Guo et al. 2017) and conformal prediction (Kumar et al. 2023). However, these methods do not themselves fix any of the calibration deterioration that occurs during FT, only correcting it after the fact. During the FT process, certain methods, such as label smoothing, reduce a model's tendency to become overconfident (Szegedy et al. 2016). Label smoothing depends on the idea that we never know facts with 100% certainty, encouraging the model "to be less confident on all samples by smoothing the true conditional being learned" (Huang et al. 2025). However, no methods have been discovered to fully control or correct confidence calibration.

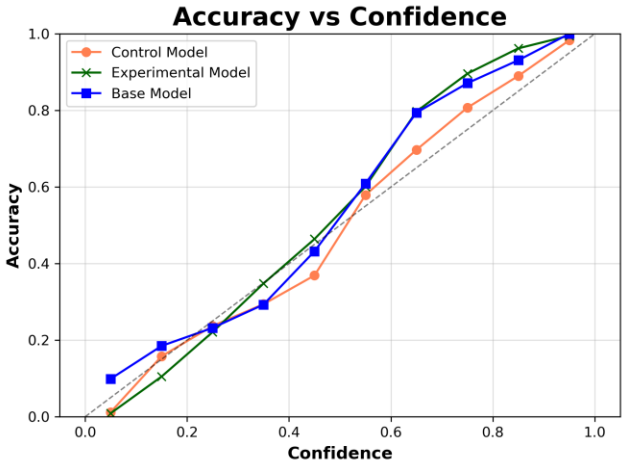

Figure 2: Calibration curves for the Llama-3.2 1B Base model evaluated on the MMLU dataset; Experimental model and Control model, evaluated on 22k questions from MMLU dataset

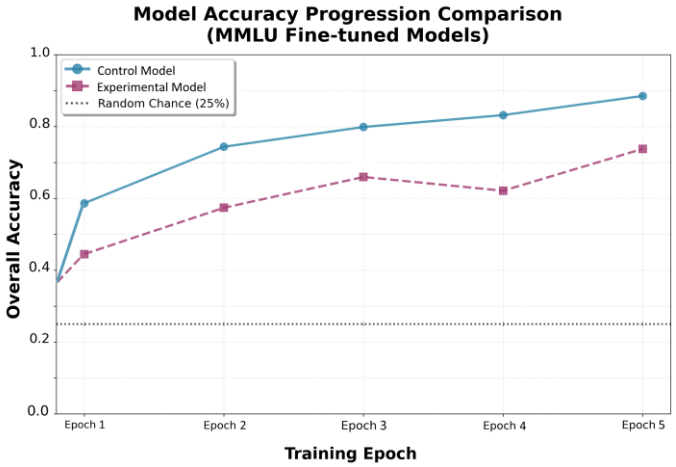

Figure 2: Control and Experimental model accuracies on 22k questions from the MMLU dataset across 5 training epochs

Methods

The ideal fine-tuning process must allow us to control the resulting confidence calibration, in addition to improving response quality in a given task. If sufficiently good calibration cannot be achieved, then the process should at least result in a model which produces an even distribution of confidences, allowing us to easily apply post-calibration.

Therefore, we propose a revised method of fine-tuning which affects the calibration of the tuned model. We take inspiration from the idea that reading a fact multiple times does not make it more correct—an idea not fully represented within the context of LLM fine tuning. In our method, we used the cross-entropy loss of the correct answer compared to the model's predicted probability distribution across tokens. Updates to the weights were only performed when the cross-entropy loss exceeded a threshold parameter t, meaning the weights of the model were not updated if it already 'knew' the answer.

Cross-entropy loss was chosen to determine weight updates because it tells us how well the model has learned a given question. The cross-entropy loss is low ($\cong 0$) when the model is confidently correct, and high ($\gg 1$) when it is confidently incorrect. A confidently correct answer need not be changed, so the weights need not be updated. Oppositely, a high cross-entropy loss would indicate a model placed its confidence in incorrect answers and the weights must be updated. Cross-entropy loss ranges from $[0,\infty)$, with a score of 1.39 indicating an even distribution across four answer choices. The initial experimental threshold t was chosen to be less than 1.39 so that weight updates are triggered unless the correct multiple choice answer is selected with sufficient confidence.

We fine-tuned Meta's Llama-3.2, 1B parameter model on 22k random questions from the MMLU dataset using parameter-efficient fine-tuning (PEFT) methods (t=0) for the Control Model and PEFT with Conditional Updates Fine-Tuning (t=1) for the Experimental Model. A Low-Rank Adaptation was used with rank = 2, and the models were trained for 5 epochs, with the following hyperparameters:

LoRA alpha = 4
LoRA dropout = 0.07
Batch size = 2
Optimizer = AdamW with a learning rate of 1e-4

The MMLU dataset contains multiple choice questions used frequently in LLM benchmarking, such as in the GPT-4 technical report for a similar confidence calibration evaluation. The model was prompted with a question followed by answer options labeled A through D; then, one token was generated. The single-token responses allowed us to easily understand the model's confidence in each answer choice letter by looking at the SoftMax scores over the four letters.

Results

When an evaluation of the confidence calibration was run on the 22k MMLU questions, three major results were examined.

(1) As epochs progressed, the tuned models showed different confidence calibrations, with the Control showing greater overconfidence and the Experimental Model showing greater under-confidence as compared to the Base Llama Model (Fig. 1). The calibration curve of the Control model always either matched the base model or was lower, meaning that the *confidence increased faster than the accuracy*.

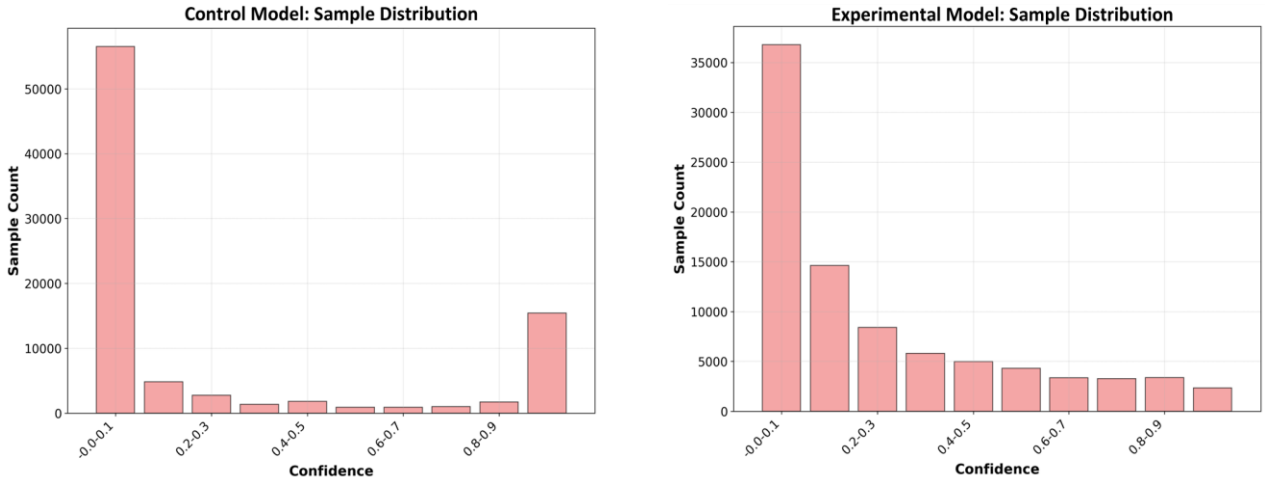

Figure 3: Distribution of Control and Experimental Model confidence scores across all answer choices A through D (including wrong answers the model correctly knew were wrong) when evaluated on 22k questions from MMLU

For moderate and high confidences, the Experimental model lies at or above both the base and Control model, meaning that the Experimental model tended to become less confident.

(2) Additionally, the Experimental model showed a more even distribution of confidence scores, which is advantageous for post-calibration of the model. Methods like histogram binning and conformal prediction rely on a distribution of samples across bins in order to predict the true accuracy based on the confidence score given by the model. Meanwhile, the Control model's confidence scores for each answer choice drifted apart towards 1 and 0, making the distribution of confidence scores steadily more polarized with each epoch (Fig. 3). An interesting consequence of Conditional Update FT as performed on the Experimental model is that data points in calibration evaluation tend to be pushed towards the center of the distribution, as compared to the Control. This contrasts with traditional FT in which data points tend to be pushed to the sides of the distribution.

(3) The accuracy of the Control model improved more than the Experimental model within the first epoch; however, in following epochs, both models continued to improve at a similar rate despite skipped weight updates in the Experimental model (Fig. 2).

Discussion and Limitations

With respect to LLM use in healthcare, a model that thinks more critically is advantageous. Many questions are fairly open-ended without an accepted 'truth.' In such cases, decisions should be made more thoughtfully, without undue confidence. We demonstrate an underconfident Experimental Model as a result of Conditional Update FT and believe this can contribute to the desirable, more tentative model personality for healthcare decisions.

Underconfidence in model personality may also have implications for reasoning model applications. Underconfidence may encourage exploration over exploitation in reasoning models that use graph traversal to reach an answer such as Graph of Thought (Besta et al. 2023) and Tree of Thought (Yao et al. 2023) reasoning models as well as LLM knowledge graph exploration (Guo, X et al. 2025).

There are several limitations to this experiment we would like to address. Neither model reached near perfect accuracy despite being trained and validated on the same set of MMLU questions, because a small LoRA rank was used in the experiment. Although the accuracy of both Control and Experimental models improved, the Experimental ended at a lower value; this is expected since the model is being updated progressively less each epoch. Adjusting the threshold parameter to be higher or lower would decrease or increase the number of updates applied to the model, respectively.

Conclusion

This method for affecting confidence calibration while fine tuning LLMs may potentially help in the broader challenge of controlling a model's 'personality' and reliability. Furthermore, a control over the resulting confidence calibration of a model would allow us to either improve calibration or ensure that the model is more under- or over-confident when appropriate. It provides us with more control over how the model truly thinks, ultimately improving AI safety, robustness and explainability.

Acknowledgements

S. S. gratefully acknowledges support from the Students Tackling Advanced Research program at Drexel University for providing a stipend during this research. We would like

to thank the anonymous reviewers for their time, effort and valuable input.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S; Anadkat, S. et al. 2023. Gpt-4 technical report. arXiv:2303.08774.

Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawski, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nyczyk, P.; Hoefler, T. 2023. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. arXiv:2308.09687.

Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. arXiv:1706.04599.

Guo, X.; Li, A.; Wang, Y.; Jegelka, S.; Wang, Y. G1: Teaching LLMs to Reason on Graphs with Reinforcement Learning. arXiv:2505.18499.

Huang, J.; Lu, P.; Zeng, Q. 2025. Calibrated Language Models and How to Find Them with Label Smoothing. arXiv:2311.08298.

Kumar, B.; Lu, C.; Gupta, G.; Palepu, A.; Bellamy, D.; Raskar, R.; Beam A. 2023. Conformal Prediction with Large Language Models for Multi-Choice Question Answering. arXiv:2305.18404.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. Rethinking the inception architecture for computer vision. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp. 2818–2826, 2016.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L. ; Cao, Y.; Narasimhan, K. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:2305.10601.

Zadrozny, B.; Elkan, C. 2001. Learning and making decisions when costs and probabilities are both unknown. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. UC San Diego: Department of Computer Science & Engineering. https://escholarship.org/uc/item/62h3k2mv

Zhu, C.; Xu, B.; Wang, Q; Zhang, Y.; Mao, Z. 2023. On the Calibration of Large Language Models and Alignment. arXiv:2311.13240.