Literature Review: Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback (Tian et al., 2023)

Ali Bakly | 383057 | ali.bakly@epfl.ch
ab-eh-me

1 Summary

This paper examines methods for obtaining well-calibrated confidence estimates from large language models (LLMs) that have been fine-tuned using reinforcement learning from human feedback (RLHF). A well-calibrated confidence score implies that the model's confidence in a prediction accurately reflects the likelihood that the prediction is correct. The authors find that RLHF generally worsens the calibration of an LLM's conditional probabilities when compared to those from unsupervised pre-training. However, surprisingly, RLHF-LMs can directly verbalize confidence estimates that are relatively well-calibrated. Further improvements in the calibration of these verbalized confidences are achieved by prompting the model to consider multiple possible answers before responding. These findings are demonstrated across multiple state-of-the-art RLHF-LMs on various factual question-answering datasets, with the methods employed reducing the expected calibration error (ECE) by 50% relative to the model's own conditional probabilities.

The authors propose two primary methods to extract the internal conditional probabilities of closed-source models. The simplest method, labeled **Label prob**. utilizes the conditional probability distribution $P(y|x)$ of the model given a question x, estimated from $n = 10$ samples. For verbalized confidence scores, they introduce five different methods, largely inspired by research in human psychology that suggests overconfidence can be mitigated by considering alternative answers before responding (Lord et al., 1984). The method **Verb. 1S top-** k prompts the model to generate k guesses and assigns a probability to each, selecting the guess with the highest probability as the model's output and confidence indicator. **Ling. 1S-human** employs linguistic likelihood expressions to articulate uncertainty, where the model assigns confidences to its predictions by selecting from a predefined set of expressions, which are quantitatively mapped to probabilities based on results from a human survey. The remaining three methods are variations on these themes, each adapted to refine the expression of calibrated confidence

2 Strengths

For a prediction system to be reliable in practical applications, it should output confidence estimates that accurately reflect its probability of being correct. Specifically, when the system expresses low confidence in a prediction, that prediction should be more likely to be incorrect. The proposed novel methods to improve the calibration of RLHF-LMs are the main strengths of this paper. For example a well calibrated NLP system could defer to a human expert in cases of uncertainty, avoiding the pitfalls of overconfident predictions that prove erroneous.

Another key strength of this work is the extensive empirical evaluation conducted by the authors. They measure calibration on five different RLHF-LMs: GPT-3.5-turbo, GPT-4, Claude-1, Claude-2, and Llama-2-70b-chat across the TriviaQA, SciQ, and TruthfulQA benchmarks. Furthermore, they report four different metrics for calibration: ECE (expected calibration error), ECE-t (ECE with temperature scaling), BS-t (Brier Score on temperature-scaled confidences), and AUC (area under the curve of selective accuracy and coverage). Most importantly, they evaluate numerous prompting techniques for the models to verbalize confidence scores, but they also assess various methods to extract internal conditional probabilities, as most of these models are closed source. Due to this extensive evaluation, it is straightforward to assert that their results are reliable.

The results of the paper consistently show that closed-source large RLHF-LMs (such as Claude and ChatGPT) often verbalize better-calibrated confidences than the models' internal conditional probabilities. Moreover, the study demonstrates that allowing the model to produce several possible answers improves calibration—an interesting finding that suggests generating and evaluating multiple hypotheses enhances calibration, aligning with human psychological processes. Notably, RLHF-LMs can express their uncertainty with numbers very effectively, which is particularly significant as it challenges the longstanding problem in NLP of difficulty representing numbers(Thawani et al., 2021). This can be seen as a strength of the paper, as many of the results have intriguing applications beyond calibration. For example, implementing methods grounded in human psychology to improve certain aspects of LLMs might extend further than improvement of calibration.

3 Weaknesses

A particular weakness of this paper is that improvements in calibration are predominantly observed in the closed-source models, ChatGPT and Claude, while they are much less consistent in the open-source model, Llama-2-70b-chat. For the closed-source models, the conditional probabilities are typically estimated using methods such as **Label prob**., but the paper does not explicitly detail how the conditional probabilities are extracted for Llama-2-70b-chat. Given that Llama-2-70b-chat is open source, it would be reasonable to assume that these probabilities are accessed directly, without the need for estimation. This distinction is critical because if Llama-2-70b-chat directly accesses probabilities, it is troubling that this model does not show a clear improvement in calibration. Since Llama-2-70b-chat does not rely on estimation, its calibration metrics could arguably provide the most accurate reflection of RLHF-LMs intrinsic capability of verbalizing calibration. The inconsistent results for Llama-2-70b-chat could imply that the estimation of the conditional probabilities of the closed source models are unreliable or that Llama-2-70b-chat has inherent limitations within its architecture or training regimen that affect its ability to achieve better calibration. The authors do not adequately discuss the implications of Llama-2-70b-chat's inconsistent calibration improvements or explore the possible reasons behind this discrepancy.

Another limitation of this study is that the experiments focus exclusively on factual question-answering tasks that predominantly assess knowledge retrieval capabilities. While the authors' techniques demonstrate promising results within this specific context, it remains uncertain how well these methods would generalize to tasks requiring more complex reasoning. Examples of such tasks include mathematical problem solving, coding challenges, or multi-step scientific questions, where the model must generate its own solution paths rather than merely retrieving factual information. The robust calibration of verbalized confidences observed may not seamlessly translate to these more dynamic and open-ended settings. Furthermore, incorporating tasks that involve greater elements of creativity or subjective judgment, such as generating artistic content or speculative writing, could also provide insights into the versatility of these calibration techniques. Broadening the scope of evaluation to include a wider range of established language benchmarks would help ascertain the full spectrum of the potential impact of these methods.

References

Charles G Lord, Mark R Lepper, and Elizabeth Preston. 1984. Considering the opposite: a corrective strategy for social judgment. *Journal of personality and social psychology*, 47(6):1231.

Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. 2021. Representing numbers in NLP: a survey and a vision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–656, Online. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.