Adaptación de modelos grandes de lenguaje con *few-shots learning* y calibración post-hoc

Juan Ignacio Tollo[1]

[1]FCEyN, Universidad de Buenos Aires | jtollo@dc.uba.ar

|ORCID iD: 0009-0003-6069-3355

Abstract. En el contexto de adaptación de modelos de lenguaje a una tarea específica, utilizar *prompt engineering* suele mostrar una ganancia en la performance sin necesitar acceder a los pesos del modelo. Otra forma de adaptación, mucho menos estudiada en la literatura, es a través de técnicas de calibración post-hoc, en donde se accede únicamente a la salida del modelo y se hace una transformación lineal sobre ella que redunda en una mejor performance en la tarea. Este enfoque de "caja gris", en donde sólo se accede a los valores de la capa de salida del modelo, ofrece una alternativa de adaptación computacionalmente más barata que las técnicas de fine-tuning y no ha sido estudiado en profundidad en la literatura. Este trabajo muestra algunos resultados preliminares en donde la combinación de *prompt engineering* y calibración post-hoc logran una mejora en tareas de preguntas de opción múltiple sobre comportamiento social en dos modelos de lenguaje grandes, Phi-1.5 y Phi-2. La complementariedad entre estas técnicas abre un camino de investigación más amplio para potenciar la performance del modelo a partir de combinar sistemáticamente *prompt engineering* y calibración post-hoc.

Keywords: aprendizaje con pocos ejemplos, calibración, adaptación de modelos grandes de lenguaje

Large language model adaptation through few-shots learning and post-hoc calibration

Abstract. In the context of adapting language models to a specific task, using prompt engineering often yields performance gains without requiring access to the internal parameters of the model. Another form of adaptation, much less studied in the literature, is achieved through post-hoc calibration techniques, where only the model's output scores are accessed and modified via a function to enhance task performance. This "gray-box" approach, which only utilizes the values from the model's output layer, offers a computationally inexpensive alternative to supervised learning techniques and has not been thoroughly investigated in the literature. This work presents some preliminary results in which the combination of prompt engineering and post-hoc calibration shows improvements in multiple-choice social behavior question tasks on two large-scale language models (Phi-1.5 and Phi-2). The results obtained so far suggest that these two techniques are complementary, paving the way for the development of techniques that systematically combine prompt engineering with post-hoc calibration to improve model performance.

Keywords: few-shot learning, calibration, large language model adaptation

1 Introduction and related work

Large Language Models (LLMs) are useful for many Natural Language Processing tasks, such as question answering and summarization. However, it is also known that LLMs can benefit from adaptation to a downstream task when samples from that task are available. This adaptation often relies on fine-tuning, which is computationally expensive and requires full access to the model parameters. In contrast, prompt engineering and few-shot learning in particular offer a lightweight and flexible alternative by embedding task examples directly into the input prompt (Brown et al., 2020). Nonetheless, for multiple-choice question answering tasks, model performance can be highly sensitive to the prompt structure. This includes its format, the selected few-shot examples, and the order in which those examples are presented (Zhao et al., 2021).

For Pezeshkpour and Hruschka (2024) LLM prompt sensitivity primarily stems from two factors: inherent model uncertainty when distinguishing between top candidate answers and a positional bias that predisposes the models to favor options based on their placement. Moreover, Guo et al. (2017) show that models often fail to distribute probability mass among valid options in a calibrated sense. Even in simple scenarios with well-defined likelihoods, their predictions exhibit systematic biases rather than reflecting true uncertainty. These challenges highlight the need for methods that address miscalibration directly at the output level. In this sense, calibration is an appealing "gray-box" adaptation lever, as it requires access only to the model's inputs and outputs.

In response to these challenges, Pezeshkpour and Hruschka (2024) propose two mitigation strategies: a) majority vote approach over 10 random reorderings and b) a Multiple Evidence Calibration (MEC), where for a query with two possible answers, then a separate agent selects the best answer. However, the authors note that majority voting is resource-intensive, while MEC fails to improve performance. As an alternative, Zhao et al. (2021) work on contextual calibration using context-free inputs and then correct biases in model outputs. They showed that combining few-shot learning with contextual calibration can improve the overall performance, i.e. calibration plus discrimination.

While these approaches operate at the level of final label predictions — and, in some cases, on the associated probabilities after applying the softmax function — less attention has been paid to combining few-shot learning with calibration methods applied directly at the logit level. Working directly on the logits can be advantageous because the softmax function can compress their richer uncertainty into a single, sharply peaked probability vector, often inflating confidence even when the prediction is incorrect (Guo et al., 2017). To address this issue, post-hoc calibration adjusts a model's output scores to minimize a proper scoring rule (usually the cross-entropy) on task-specific data (Kull & Flach, 2015). Following this approach, this work explores two main questions:

- Can few-shot learning alone achieve posterior quality comparable to that obtained through post-hoc calibration?
- How sensitive—in terms of both mean performance and variance—are few-shot and post-hoc calibration methods to minor changes in prompt format?

2 Experiments and preliminary results

In this section we present preliminary results on the use of few-shot learning in combination with post-hoc calibration. We explored three minimal prompt formats: (A) the original prompt as extracted from the dataset, (B) replacing the ")" after the label options with a ".", and (C) changing "answer:" to "the most correct answer is". Once the prompt was fixed, we extracted the logits for the tokens corresponding to each possible answer choice. This procedure was combined with adding 0, 2, 4, 6, or 8 few-shot examples to the start of the prompt. We calibrate the raw scores, s_{raw}, using an affine transformation $s_{cal} = \alpha \cdot s_{raw} + \beta$, where α and β are learned parameters (Brümmer & Van Leeuwen, 2006).[1]

A pool of 1000 instances constituted our *training set*, used exclusively to estimate the calibration parameters. In addition, we defined a separate fixed *shot-context set* of 50 instances from where we randomly select shots to add in the prompt. Thus the context examples the model sees during calibration and during evaluation are identical, and they never overlap with the training set. We repeated this process for 3 seeds.

We report results on HellaSwag and SocialIQA, two social behavior question-answering dataset tasks (Lourie et al., 2021). As models, we used Microsoft's Phi-1.5 and Phi-2 models, which are instruction fine-tuned models that originally report results on the above mentioned datasets (Javaheripi & Bubeck, 2024; Li et al., 2023). We report results on a held-out split of the data corresponding to a subset of the original validation set. We use Normalized Cross Entropy (NCE) as evaluation metric which assesses the quality of the scores by directly comparing them with the true likelihood of the data. Additionally, this metric allows the decomposition between calibration (the difference between the uncalibrated model and the calibrated one), and discrimination (the value of the calibrated model). For more details on this see Ferrer and Ramos (2025).

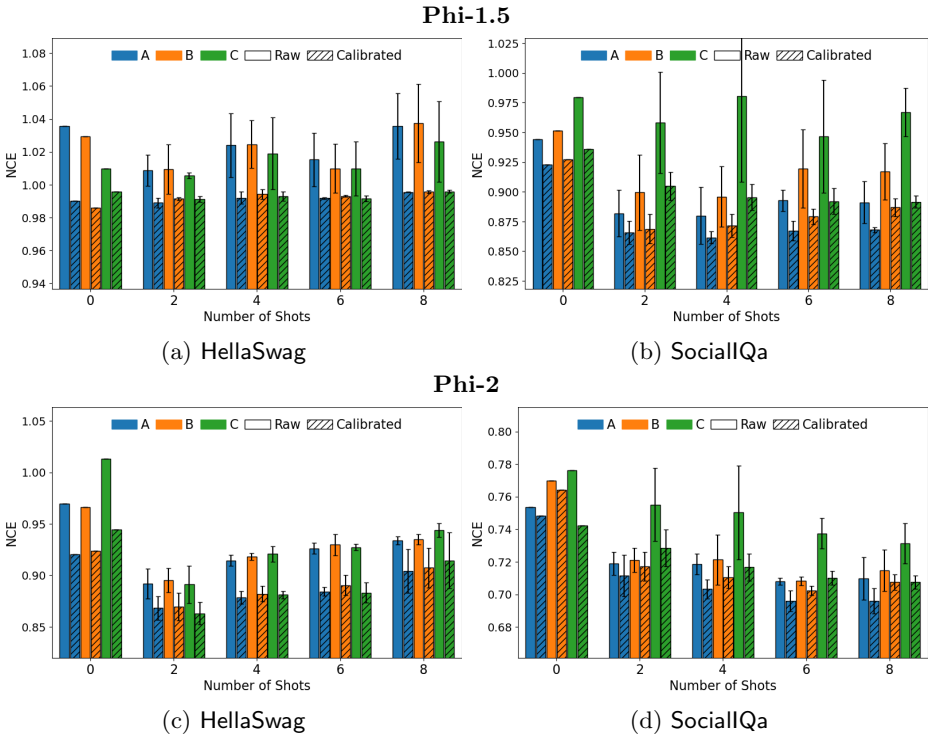
Figure 1 shows performance improvement in NCE in the calibrated posteriors. We can see that minor changes in prompts—such as variation (C)—can significantly deteriorate performance, but after calibration, the posteriors perform comparably to other formats.

Regarding variance, both prompt format and random sample selection affect the metrics. First, prompt format variance is reflected in the height of the bars. As shown in Figure 1, the calibrated posteriors show smaller differences between prompt formats compared to raw posteriors. Second, in few-shot settings, adding samples to the prompt introduces variance due to random sample selection, which is reflected in the error bars of raw posteriors. Here, post-hoc calibration also helps reduce this source of variance.

3 Discussion and Conclusion

Our preliminary experiments confirm that prompt engineering and calibration are complementary techniques, jointly improving performance and reducing vari-

[1] For implementation we follow https://github.com/luferrer/CalibrationTutorial



(a) HellaSwag (b) SocialIQa

(c) HellaSwag (d) SocialIQa

Fig. 1: For each model–dataset pair, this figure shows the NCE achieved across all combinations of prompt format (A,B and C) and number of shots. Solid bars show performance before calibration, while striped bars show performance after.

ance (measured by NCE) on HellaSwag and SocialIQa datasets with Phi-1.5 and Phi-2 models. From Figure 1, key findings and open questions are:

- Most of the prompting techniques tested so far result in non-calibrated systems. However, there is no reason to think that there are some specific format of the prompt or a number of shots that produce a model that is systematically calibrated. We open the question of how to achieve that case.
- The number of shots used is not following a proportional tendency with respect to the calibrated performance in the tested tasks, raising the question if this is the same for other tasks and other number of shots.
- For all prompt format and number of shots considered, performance seems to be more stable in terms of variance after calibration. However, additional tasks and models need to be considered to confirm this tendency.
- There are some cases in which worst-performing prompt format can become the best after calibration (e.g., Phi-2, SocialIQa zero-shot). This raises the following interesting question: are there any prompting techniques that only correct calibration instead of discrimination and calibration at once? If this is the case, this question could lead to produce more controllable and interpretable prompting techniques, and even to provide models that produce high-quality scores and interpretable outputs.

Future actions involve extending analyses to different models and additional datasets, exploring whether chain-of-thought posteriors are calibrated, and systematically integrating prompt engineering with calibration.

4 Acknowledgments

Research in this publication was supported by an Amazon Research Award Fall 2023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not reflect the views of Amazon.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 1877–1901, Vol. 33). Curran Associates, Inc.

Brümmer, N., & Van Leeuwen, D. (2006). On calibration of language recognition scores. *Proceedings of IEEE Odyssey 2006:The Speaker and Language Recognition Workshop, 2006*, *1*, 1–8.

Ferrer, L., & Ramos, D. (2025). Evaluating posterior probabilities: Decision theory, proper scoring rules, and calibration. *Transactions on Machine Learning Research*.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (pp. 1321–1330, Vol. 70). PMLR.

Javaheripi, M., & Bubeck, S. (2024). *Phi-2: The surprising power of small language models*. Microsoft Research Blog.

Kull, M., & Flach, P. (2015). Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. *Machine Learning and Knowledge Discovery in Databases*, 68–85.

Li, Y., Bubeck, S., Eldan, R., Giorno, A. D., Gunasekar, S., & Lee, Y. T. (2023). Textbooks are all you need ii: Phi-1.5 technical report.

Lourie, N., Bras, R. L., Bhagavatula, C., & Choi, Y. (2021). Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI-21)*, 13480–13488.

Pezeshkpour, P., & Hruschka, E. (2024). Large language models sensitivity to the order of options in multiple-choice questions. *Findings of the Association for Computational Linguistics: NAACL 2024*, 2006–2017.

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. *Proceedings of the 38th International Conference on Machine Learning*, *139*, 12697–12706.