A Case Study Replicating Calibration of Large Language Models on Open Models and Datasets

Anonymous ACL submission

Abstract

The popularity of large language models has lead to a growing need to characterize their behavior and understanding beyond the generation grammatical sentences. In this study we attempt to replicate and extend the calibration study of (Kadavath et al., 2022) on open models and datasets, attempting compare model reported response probabilities to their actual calibration, that is the certainty with which a model answers a question with a given completion and the probability that the model is answering the question correctly.

1 Introduction

2 Method

2.1 Introduction

We measure the calibration in keeping with the methodology presented in (Kadavath et al., 2022). The model is queried in either a 0-shot or 5-shot manner. The transformer model is given a question prompt with the each of the multiple choice options. In the 5-shot methodology, in addition to the question under test we proceed the question with 4 additional questions with in the same multiple choice format in order to provide the model with enough context to understand the expected format for answering the question.

For each prompt completion pair we compute the log probability of the completion normalized by its completion length.

The computed probabilities are then grouped in bins in 10 bins from 0 to 1 were we compute the average frequency of answering correctly for each bin.

For ideal calibration, that is when the model completion probability aligns closely with the actual probability of answering the question correctly, these two computed probabilities must be equal, thus a ideal calibration would be represented by a line of slope 1 in the calibration plot. Negative and positive deviations thus represent a model which is under-confident or over-confident in its answers respectively.

2.2 Dataset

The Measuring Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021) is a massive dataset of multiple choice questions which covers 57 tasks in various subjects including elementary mathematics, US history, computer science, law, and more. It serves as a challenging dataset for most modern large language models and can be used to evaluate their calibration.

2.3 Model Calibration By Number of Parameters

We compare the calibration of the Llama (Touvron et al., 2023) model 7b-chat and 13b chat models for answering multiple choice questions in figure ...

We note that that the calibration of the model while not perfect improves with the size of the model with the 13b-chat model being better calibrated and closer to the ideal calibration line than the 7b-chat model. Thus we note that models sizes improves not only the performance of the model but the calibration of the model as well.

2.4 Model Calibration By Fine Tuning

2.5 Model Calibration By Subject Specialization

3 Results

4 Discussion

5 Conclusion

In this study we attempted to study language model calibration under various conditions such as model size, fine tuning and task specialization. We found that we were able to replicate the observed calibration behavior of closed models like GPT-3, and

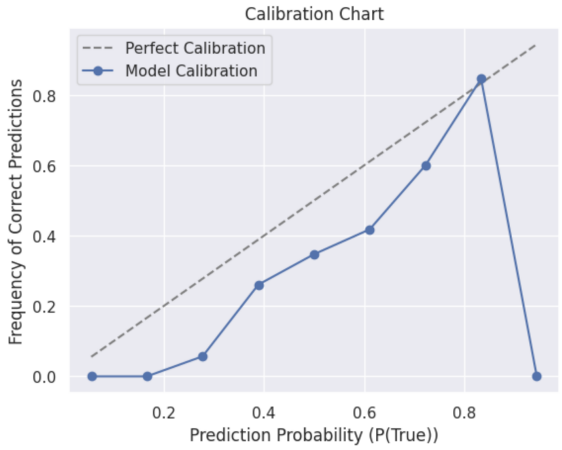

Figure 1: Calibration of the 13b-chat model on the MMLU dataset, we note that it is closer, in calibration than the 7b-chat model.

Claude on the open models like Llama 7b-chat and 13b-chat.

We note that the calibration of the models improved with the size of the model ...

While calibration is one aspect of querying model understanding it is certainly not the only criteria for evaluating model understanding.

Other criteria might include models' ability to reason step by step, and other demonstrate conceptual understanding by generalizing out of distribution.

6 Acknowledgements

7 References

8 Acknowledgements

References

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

A Appendix

This is a section in the appendix.