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.

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.

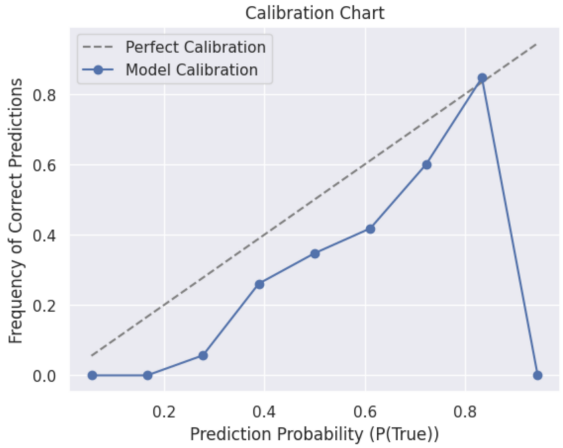

Figure 1: Calibration of the 13b-chat model on the MMLU dataset, we note that it is closer, in calibration than the 7b-chat model.

2.3 Model Calibration

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.

3 Results

The first line of the file must be

```
\documentclass[11pt]{article}
```

To load the style file in the review version:

```
\usepackage[review]{ACL2023}
```

For the final version, omit the `review` option:

```
\usepackage{ACL2023}
```

Output	natbib command	Old ACL-style command
(Cooley and Tukey, 1965)	`\citep`	`\cite`
Cooley and Tukey, 1965	`\citealp`	no equivalent
Cooley and Tukey (1965)	`\citet`	`\newcite`
(1965)	`\citeyearpar`	`\shortcite`
Cooley and Tukey's (1965)	`\citeposs`	no equivalent
(FFT; Cooley and Tukey, 1965)	`\citep[FFT;][]`	no equivalent

Table 1: Citation commands supported by the style file. The style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.

To use Times Roman, put the following in the preamble:

```
\usepackage{times}
```

(Alternatives like txfonts or newtx are also acceptable.) Please see the LaTeX source of this document for comments on other packages that may be useful. Set the title and author using `\title` and `\author`. Within the author list, format multiple authors using `\and` and `\And` and `\AND`; please see the LaTeX source for examples. By default, the box containing the title and author names is set to the minimum of 5 cm. If you need more space, include the following in the preamble:

```
\setlength\titlebox{<dim>}
```

where `<dim>` is replaced with a length. Do not set this length smaller than 5 cm.

4 Discussion

5 Conclusion

6 Acknowledgements

7 References

8 Acknowledgements

References

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

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.