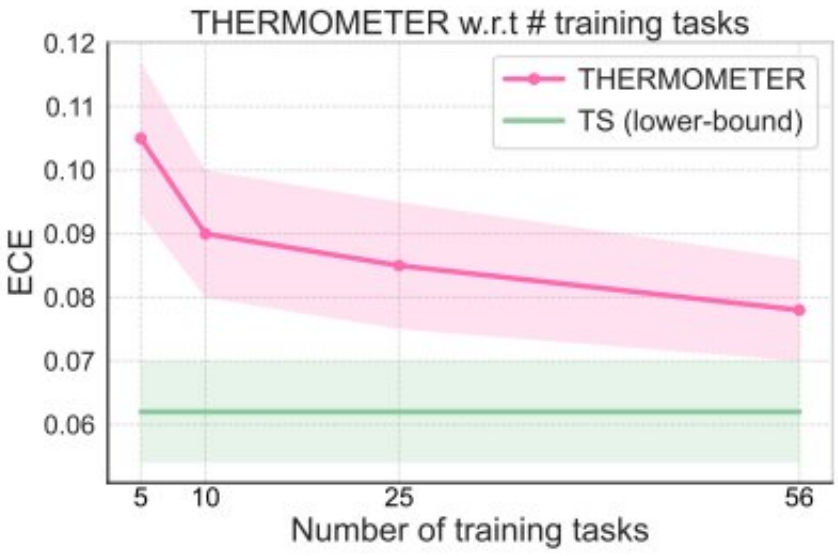
'Thermometer' technique prevents an AI model from being overconfident about wrong answers

July 31 2024, by Adam Zewe



Thermometer performance vs. number of training tasks. Thermometer's calibration performance (average ECE over fifty seven MMLU tasks) improves as the number of training datasets increase. The shaded region represents two standard error. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2403.08819

People use large language models for a huge array of tasks, from translating an article to identifying financial fraud. However, despite the incredible capabilities and versatility of these models, they sometimes generate inaccurate responses.

On top of that problem, the models can be overconfident about wrong answers or underconfident about correct ones, making it tough for a user to know when a model can be trusted.

Researchers typically calibrate a machine-learning model to ensure its level of confidence lines up with its accuracy. A well-calibrated model should have less confidence about an incorrect prediction, and vice-versa. But because large language models (LLMs) can be applied to a seemingly endless collection of diverse tasks, traditional calibration methods are ineffective.

Now, researchers from MIT and the MIT-IBM Watson AI Lab have introduced a calibration method tailored to large language models. Their method, called Thermometer, involves building a smaller, auxiliary model that runs on top of a large language model to calibrate it.

Thermometer is more efficient than other approaches—requiring less power-hungry computation—while preserving the accuracy of the model and enabling it to produce better-calibrated responses on tasks it has not seen before.

By enabling efficient calibration of an LLM for a variety of tasks, Thermometer could help users pinpoint situations where a model is overconfident about false predictions, ultimately preventing them from deploying that model in a situation where it may fail.

"With Thermometer, we want to provide the user with a clear signal to tell them whether a model's response is accurate or inaccurate, in a way that reflects the model's uncertainty, so they know if that model is reliable," says Maohao Shen, an electrical engineering and computer science (EECS) graduate student and lead author of a paper on Thermometer.

Shen is joined on the paper by Gregory Wornell, the Sumitomo Professor of Engineering who leads the Signals, Information, and Algorithms Laboratory in the Research Laboratory for Electronics, and is a member of the MIT-IBM Watson AI Lab; senior author Soumya Ghosh, a research staff member in the MIT-IBM Watson AI Lab; as well as others at MIT and the MIT-IBM Watson AI Lab.

The research was recently presented at the International Conference on Machine Learning (ICML 2024) held in Vienna, Austria, from July 21 to July 27. It is available on the *arXiv* preprint server.

Universal calibration

Since traditional machine-learning models are typically designed to perform a single task, calibrating them usually involves one task-specific method. On the other hand, since LLMs have the flexibility to perform many tasks, using a traditional method to calibrate that model for one task might hurt its performance on another task.

Calibrating an LLM often involves sampling from the model multiple times to obtain different predictions and then aggregating these predictions to obtain better-calibrated confidence. However, because these models have billions of parameters, the computational costs of such approaches rapidly add up.

"In a sense, large language models are universal because they can handle various tasks. So, we need a universal calibration method that can also handle many different tasks," says Shen.

With Thermometer, the researchers developed a versatile technique that leverages a classical calibration method called temperature scaling to efficiently calibrate an LLM for a new task.

In this context, a "temperature" is a scaling parameter used to adjust a model's confidence to be aligned with its prediction accuracy. Traditionally, one determines the right temperature using a labeled validation dataset of task-specific examples.

Since LLMs are often applied to new tasks, labeled datasets can be nearly impossible to acquire. For instance, a user who wants to deploy an LLM to answer customer questions about a new product likely does not have a dataset containing such questions and answers.

Instead of using a labeled dataset, the researchers train an auxiliary model that runs on top of an LLM to automatically predict the temperature needed to calibrate it for this new task.

They use labeled datasets of a few representative tasks to train the Thermometer model, but then once it has been trained, it can generalize to new tasks in a similar category without the need for additional labeled data.

A Thermometer model trained on a collection of multiple-choice question datasets, perhaps including one with algebra questions and one with medical questions, could be used to calibrate an LLM that will answer questions about geometry or biology, for instance.

"The aspirational goal is for it to work on any task, but we are not quite there yet," Ghosh says.

The Thermometer model only needs to access a small part of the LLM's inner workings to predict the right temperature that will calibrate its prediction for data points of a specific task.

An efficient approach

Importantly, the technique does not require multiple training runs and only slightly slows the LLM. Plus, since temperature scaling does not alter a model's predictions, Thermometer preserves its accuracy.

When they compared Thermometer to several baselines on multiple tasks, it consistently produced better-calibrated uncertainty measures while requiring much less computation.

"As long as we train a Thermometer model on a sufficiently large number of tasks, it should be able to generalize well across any new task, just like a large language model, it is also a universal model," Shen adds.

The researchers also found that if they train a Thermometer model for a smaller LLM, it can be directly applied to calibrate a larger LLM within the same family.

In the future, they want to adapt Thermometer for more complex text-generation tasks and apply the technique to even larger LLMs. The researchers also hope to quantify the diversity and number of labeled datasets one would need to train a Thermometer model so it can generalize to a new task.

More information: Maohao Shen et al, Thermometer: Towards Universal Calibration for Large Language Models, *arXiv* (2024). DOI: 10.48550/arxiv.2403.08819