Steering at the Frontier: Extending the Power of Prompting

Published December 12, 2023

By Eric Horvitz, Chief Scientific Officer; Harsha Nori, Director, Research Engineering; Yin Tat Lee, Principal Researcher

Share this page

We're seeing exciting capabilities of frontier foundation models, including intriguing powers of abstraction, generalization, and composition across numerous areas of knowledge and expertise. Even seasoned AI researchers have been impressed with the ability to steer the models with straightforward, zero-shot prompts. Beyond basic, out-of-the-box prompting, we've been exploring new prompting strategies, showcased in our Medprompt work, to evoke the powers of specialists.

Today, we're sharing information on Medprompt and other approaches to steering frontier models in *promptbase*, a collection of resources on GitHub. Our goal is to provide information and tools to engineers and customers to evoke the best performance from foundation models. We'll start by including scripts that enable replication of our results using the prompting strategies that we present here. We'll be adding more sophisticated general-purpose tools and information over the coming weeks.

As an illustration of the capabilities of the frontier models and on opportunities to harness and extend the recent efforts with reaching state-of-the-art (SoTA) results via steering GPT-4, we'll review SoTA results on benchmarks that Google chose for evaluating Gemini Ultra. Our end-to-end exploration, prompt design, and computing of performance took just a couple of days.

Let's focus on the well-known MMLU (Measuring Massive Multitask Language Understanding) challenge that was established as a test of general knowledge and reasoning powers of large language models. The complete MMLU benchmark contains tens of thousands of challenge problems of different forms across 57 areas from basic mathematics to United States history, law, computer science, engineering, medicine, and more.

In our Medprompt study, we focused on medical challenge problems, but found that the prompt strategy could have more general-purpose application and examined its performance on several out-of-domain benchmarks—despite the roots of the work on medical challenges. Today, we report that steering GPT-4 with a modified version of Medprompt achieves *the highest score ever achieved on the complete MMLU*.

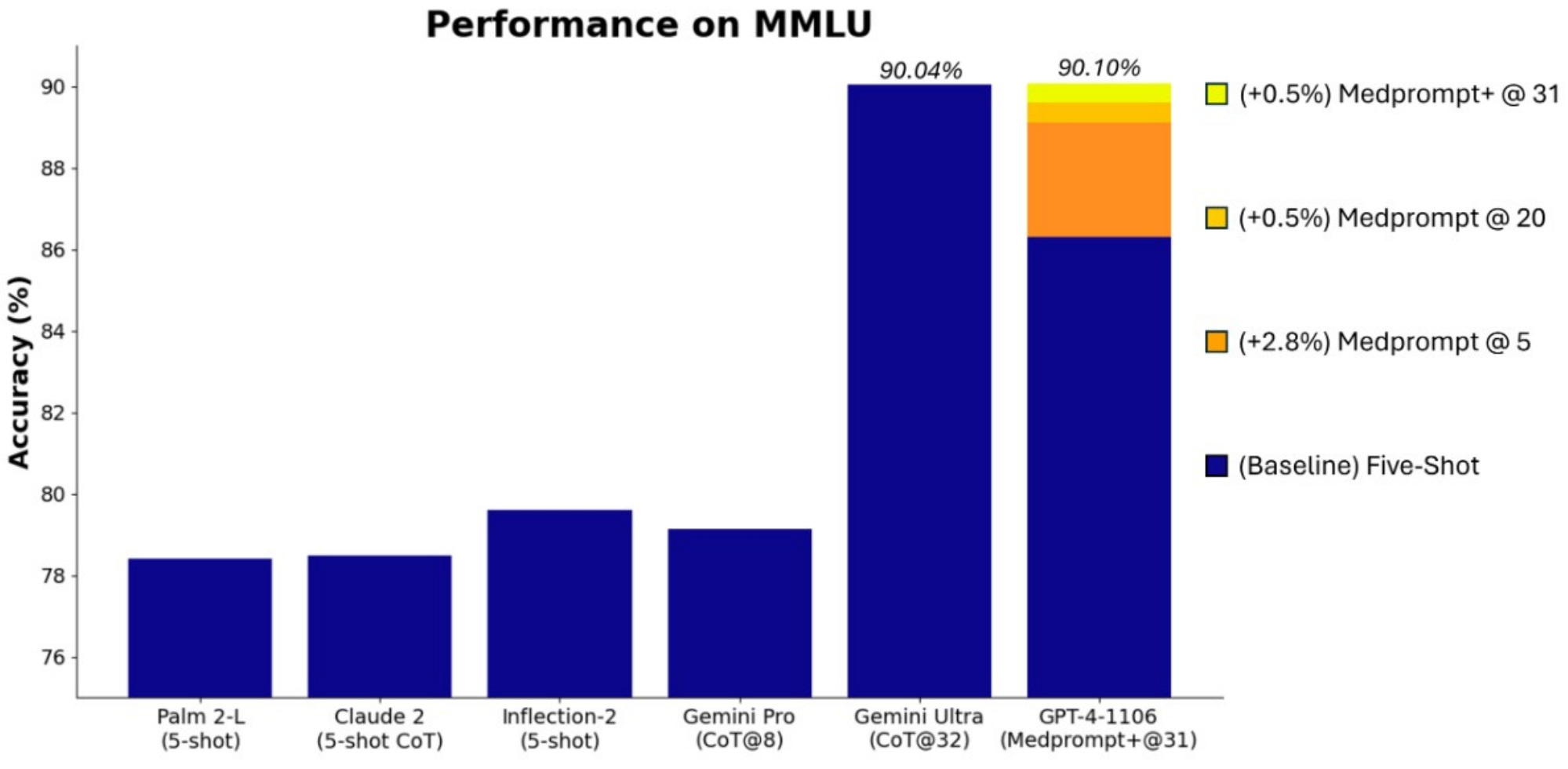

Figure1. Reported performance of multiple models and methods on the MMLU benchmark.

In our explorations, we initially found that applying the original Medprompt to GPT-4 on the comprehensive MMLU achieved a score of 89.1%. By increasing the number of ensembled calls in Medprompt from five to 20, performance by GPT-4 on the MMLU further increased to 89.56%. To achieve a new SoTA on MMLU, we extended Medprompt to Medprompt+ by adding a simpler prompting method and formulating a policy for deriving a final answer by integrating outputs from both the base Medprompt strategy and the simple prompts. The synthesis of a final answer is guided by a control strategy governed by GPT-4 and inferred confidences of candidate answers. More details on Medprompt+ are provided in the promptbase repo. A related method for coupling complex and simple queries was harnessed by the Google Gemini team. GPT-4 steered with the modified Medprompt+ reaches a record score of 90.10%. We note that Medprompt+ relies on accessing confidence scores (logprobs) from GPT-4. These are not publicly available via the current API but will be enabled for all in the near future.

While systematic prompt engineering can yield maximal performance, we continue to explore the out-of-the-box performance of frontier models with simple prompts. It's important to keep an eye on the native power of GPT-4 and how we can steer the model with zero- or few-shot prompting strategies. As demonstrated in Table 1, starting with simple prompting is useful to establish baseline performance before layering in more sophisticated and expensive methods.

Benchmark	GPT-4 Prompt	GPT-4 Results	Gemini Ultra Results
MMLU	Medprompt+	**90.10%**	90.04%
GSM8K	Zero-shot	**95.27%**	94.4%
MATH	Zero-shot	**68.42%**	53.2%
HumanEval	Zero-shot	**87.8**%	74.4%
BIG-Bench-Hard	Few-shot + CoT*	**89.0%**	83.6%
DROP	Zero-shot + CoT	**83.7%**	82.4%
HellaSwag	10-shot**	**95.3%****	87.8%

* followed the norm of evaluations and used standard few-shot examples from dataset creators

** source: Google

Table 1: Model, strategies, and results

We encourage you to check out the promptbase repo on GitHub for more details about prompting techniques and tools. This area of work is evolving with much to learn and share. We're excited about the directions and possibilities ahead.