Large Language Models Are Good In-Context Function Learners

Konstantinos Voudouris (konstantinos.voudouris@helmholtz-munich.de)
Helmholtz Munich
Munich, Germany

Elif Akata (elif.akata@helmholtz-munich.de)
Helmholtz Munich
Munich, Germany

Eric Schulz (eric.schulz@helmholtz-munich.de)
Helmholtz Munich
Munich, Germany

Abstract

Human cognitive neuroscience has revealed that humans are capable of flexibly learning complex functions. Large Language Models (LLMs) are increasingly being compared to humans in terms of their intelligence and behavioural sophistication. Here, we use a principled framework to examine whether Large Language Models are able to flexibly learn functions in-context. We find a human-like behavioural motif, in which LLMs are better able to learn smoother, more predictable functions with less noise, and that their in-context learning accuracy approaches the theoretical maximum in the limit.

Keywords: Function learning; Gaussian processes; Large Language Models

Introduction

Learning to associate noisy observations of our world is a fundamental component of our cognitive tool box. We are able to learn the appropriate amount of time to boil the kettle for a perfect cup of tea and determine how long to toast bread for optimal crispiness. Solving these types of problems requires inferring reliable information from inherently noisy sensory and experiential cues. Recent advances in machine learning, particularly with the rise of large language models (LLMs), have shown that this ability to infer underlying functional relationships is central not only to human cognition but also to modern machine learning systems.

Background

Past research on rational models of function learning have demonstrated that humans can generalize from limited, noisy examples by inferring the underlying structure of the environment (Lucas, Griffiths, Williams, & Kalish, 2015; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). In humans, this process of learning functions from noisy data has been effectively modeled using Gaussian Processes (Schulz, Tenenbaum, Reshef, Speekenbrink, & Gershman, 2015), which capture the inherent uncertainty in sensory observations (Williams & Rasmussen, 2006).

Parallel to these human capabilities, recent advances in machine learning have introduced the paradigm of in-context learning for large language models (LLMs). In-context learning describes the ability of LLMs to improve their performance on a given task after being provided with a number of task-relevant demonstrations (Brown et al., 2020). This human-like ability can make LLMs exceed traditional deep learning methods in few-shot tabular data classification (Hegselmann et al., 2023), and even outperform human participants in reinforcement learning scenarios like two-armed bandit tasks (Binz & Schulz, 2023).

Gaussian Processes: A Principled Framework For Function Learning

To produce a principled framework for studying function learning in LLMs, we draw on previous work in function learning in humans (Schulz et al., 2015). Gaussian Process Regression is a method for learning functions from data by placing a prior distribution (a Gaussian Process, GP) over the space of functions, and then inferring the best fitting function from that space using Bayes rule. Formally, a GP defines a probability distribution over function space, $p(f)$ where f is some function, $f(x)$, that takes some input x and returns an output y. By defining a GP, we are able to sample functions with formally defined properties, corresponding, for instance, to the functions' predictability. We can then use these functions as the basis for studying function learning, and associate the properties of those functions with how quickly different systems can learn them.

A GP is parametrised by a mean function $m(x) = \mathbb{E}[f(x)]$ and a covariance function, or kernel, $k(x,x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]$. The mean controls the expected output, y, across the distribution of functions and possible inputs and is usually centred at 0 for convenience. The covariance function controls the variability around this expected output for different inputs across the space of possible functions. A particularly flexible class of covariance functions is the *Matérn kernel*, defined as:

$$k_s(\tau) = \frac{2^{1-\nu}}{\Gamma(\nu)}(\frac{\sqrt{2\nu}\tau}{\gamma})^{\nu}K_{\nu}(\frac{\sqrt{2\nu}\tau}{\gamma}) \quad (1)$$

where τ is the difference (or more generally, for vector inputs, Euclidean distance) between two input values, x and x',

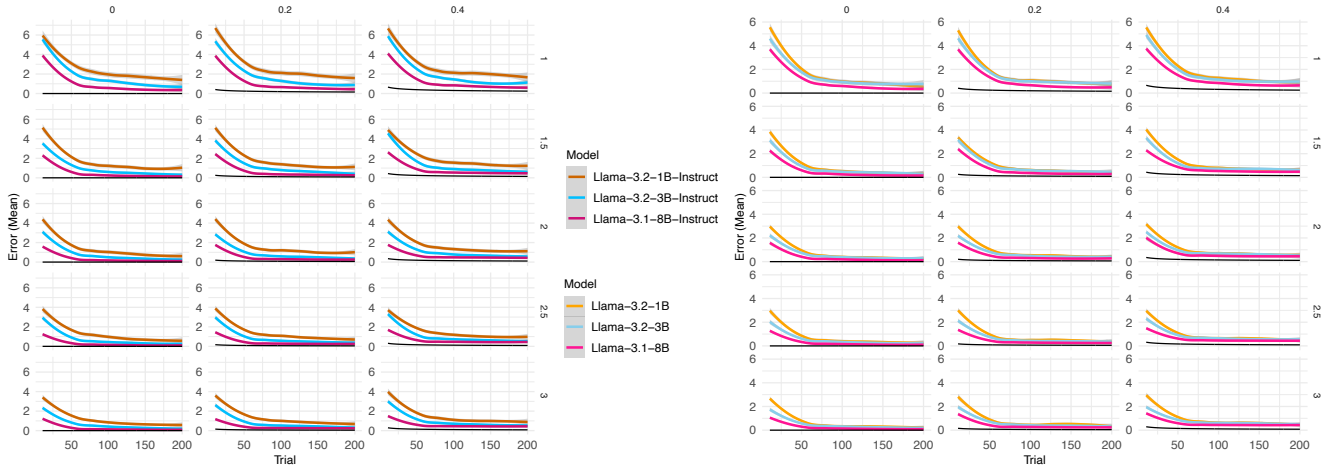

Figure 1: LOESS smoothed learning curves for six LLMs on functions sampled from kernels with different smoothness (rows) and with different amounts of injected Gaussian noise (columns). The theoretical learning curves are shown in black on each plot. Results are shown from Trial 10 onwards as initial errors were very large.

ν is the *smoothness* parameter, γ is the length scale, $\Gamma(\cdot)$ is the gamma function, and $K_\nu(\cdot)$ is the modified Bessel function of order $\nu = s - \frac{1}{2}$. For any two inputs, then, the Matérn kernel essentially encodes how variable the function is across the intervening input space, with γ controlling the scale of that intervening space and higher values of ν effectively increasing the correlation between input values at larger distances from each other. Formally, ν encodes the differentiability of the functions, and indeed, in the infinite limit, produces smooth functions in the technical sense. ν can therefore be thought of as controlling how predictable functions sampled from the GP are, since higher values produce functions that produce more similar outputs for large differences in x.

A benefit of using GPs to sample functions is that we can also describe the theoretical lower bound of the error when learning functions from particular parametrizations of the Matérn kernel as the volume of training data increases

(Opper & Vivarelli, 1998). For the squared error, $\mathcal{L}(y,\hat{y}) = (|y-\hat{y}|)^2$, the expected error after n input points, $\mathcal{E}(n)$, can be expressed as the normalized sum of the non-zero eigenvalues, λ, of the covariance function, scaled by any injected noise, σ:

$$\mathcal{E}(n) \geq \sigma^2 \sum_{i=1}^{N} \frac{\lambda_i}{\sigma^2 + n\lambda_i} \tag{2}$$

Here, we inject noise by adding a scalar sampled from the Gaussian: $y = f(x) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Computing these theoretical learning curves for the Matérn kernel allows us to determine how close empirical learning curves from test subjects (humans, LLMs) are to the theoretical expected minimum for an optimal function learner. Of course, for cases with 0 injected noise, the minimum is 0 from the first input.

Materials & Methods

We studied function learning in the *Base* and *Instruct* versions of three Large Language Models trained by Meta: *Llama-3.2* (1 billion parameters), *Llama-3.2* (3 billion parameters), *Llama-3.1* (8 billion parameters). We prompted these models with text with the following structure:

```
You are a number predictor.
I will give you a number, X, and then you need
to predict a new number, Y. There may be noise
in the true prediction.
Let's begin...
{X: input, Y: output} X: input, Y:
```

Where {`X: input, Y: output`} is repeated 0-199 times with different input-output pairs sampled from a function. We sampled 20 functions each from 5 Matérn kernels with $\nu = [1, 1.5, 2, 2.5, 3]$ respectively and $\gamma = 1000$, and injected Gaussian noise with $\sigma^2 = [0, 0.2, 0.4]$. At each trial, we compute the absolute error between the model's prediction and the true y. We also computed the theoretical learning curves for the absolute error by square-rooting the result of Eq. 2.

Results

Fig. 1 shows the results of six LLMs on our dataset of functions. All models learn to fit the functions over 200 samples with absolute error below 2. We see that the models are ranked by size across all conditions, with the 8 billion parameter model consistently showing lower error during learning, although these differences appear to be attenuated by instruction-tuning. Error is higher during learning for less smooth functions with more noise. All models approach the theoretical learning curves by the end of training. A mixed effects model with noise and smoothness as fixed effects and

function as the random effect found increasing smoothness significantly reduced error but that there was no significant difference for noise, at $\alpha = 0.05$.

Discussion

We found that LLMs tasked with learning scalar functions can do so robustly with 200 examples, and that their performance is affected by how smooth the function is. This aligns with results from human experiments which show that we are better at learning functions that are more predictable (Schulz et al., 2015). Future work will explore whether higher noise can significantly impact performance, why instruction-tuned models are more similar in these tasks, and whether functions learnt in-context are geometrically represented in model activations.

Acknowledgments

This work was supported by Helmholtz Munich, a Jacobs Research Fellowship, and an ERC Starting Grant to E.S.

References

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., & Sontag, D. (2023). Tabllm: Few-shot classification of tabular data with large language models. In *International conference on artificial intelligence and statistics* (pp. 5549–5581).

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic bulletin & review*, *22*(5), 1193–1215.

Opper, M., & Vivarelli, F. (1998). General bounds on bayes errors for regression with gaussian processes. *Advances in Neural Information Processing Systems*, *11*.

Schulz, E., Tenenbaum, J. B., Reshef, D. N., Speekenbrink, M., & Gershman, S. J. (2015). Assessing the perceived predictability of functions. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 37).

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279–1285.

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2) (No. 3). MIT press Cambridge, MA.