Evaluating Predictive Performance and Learning Efficiency of Large Language Models with Think Aloud in Risky Decision Making

Hanbo Xie (hanboxie1997@arizona.edu)
Department of Psychology, 1503 E. University Blvd
Tucson, AZ 85719 United States

Huadong Xiong (hdx@arizona.edu)
Department of Psychology, 1503 E. University Blvd
Tucson, AZ 85719 United States

Robert Wilson (bob@arizona.edu)
Department of Psychology, 1503 E. University Blvd
Tucson, AZ 85719 United States

Abstract

Predicting human behaviors and explaining the underlying mental processes have long been objectives in psychology. Traditionally, researchers have used computational models to hypothesize cognitive processes and test these models against behavioral data. However, this approach is often constrained by the researchers' theoretical insights and may not directly reflect the actual underlying cognitive functions. The Think Aloud protocol offers a more direct approach by capturing participants' thought processes on a trial-by-trial basis during cognitive tasks. However, analyzing verbal data from Think Aloud protocols is labor-intensive and subjective. Advancements in Natural Language Processing and Large Language Models (LLM) significantly mitigate these challenges, enabling comprehensive and scalable analysis of verbal data. This study evaluates the effectiveness of LLMs combined with Think Aloud data as cognitive models in risky decision-making tasks. We compare the predictive performance and learning efficiency to previously well-established symbolic and small neural network models. Our results indicate that LLMs, enhanced with Think Aloud data, accurately predict human decisions and show superior training efficiency and generalizability with minimal data. This approach advances our computational understanding of human decision-making processes, shedding light on the mechanisms of human cognitive computation.

Keywords: Large Language Models; Computational Models; Neural Networks; Risky decision-making; Think Aloud Protocol

Introduction

One of the goals of computational modeling is to predict human behaviors and explain the underlying mental processes. However, most models infer the underlying cognitive processes by hypothesizing computations, which can be indirect and limited by the researcher's understanding of the domain.

To bridge this gap, researchers can utilize human languages, which act as a conduit between internal thought representations and external symbols. Thinking, a complex cognitive process involving world modeling, inference, and decision-making, can be directly probed through the 'Think Aloud' method Simon and Ericsson (1984). By asking participants to verbalize their thought processes while performing tasks, we gain direct insights into the cognitive steps they take. Although this approach provides a more immediate connection to human cognition, the analysis of verbal data collected from Think Aloud is labor-intensive and subjective. This complexity presents significant challenges in scaling the method and integrating it with both behavioral and computational analyses in contemporary research settings.

The advent of Natural Language Processing and Large Language Models (LLM) mitigates the limitations of the Think Aloud protocol. It can analyze and quantify text data at scale. In our study, we investigate how effectively LLMs, enhanced with Think Aloud verbal data, can serve as cognitive models to predict human decisions in risky decision-making tasks. We evaluate the efficiency of various LLMs, comparing their predictive performance and learning efficiency with well-established symbolic models, such as Prospect Theory (Kahneman and Tversky (1979)), and neural network models, like the context-dependent model (Peterson, Bourgin, Agrawal, Reichman, and Griffiths (2021)). In general, our study highlights the potential of integrating Think Aloud into LLM to advance new insights into human cognition.

Results

We ran two online studies to collect human choices and Think-Aloud in risky decision-making. In Experiment 1, we inherit the exact same setting as Kahneman and Tversky (1979) with 19 trials for Prospect Theory. 72 undergraduate students were recruited. In each trial, we present two options with different outcomes and probabilities. Participants are asked to choose what they prefer and speak out their thoughts (Figure 1). The verbal data of Think-Aloud is recorded as audio and later transcribed by Whisper (Radford et al., 2023), and further checked by research assistants. We also ran a larger Experiment 2, where we recruited 468 participants, each of them completing 60 to 100 trials of choice-making and Think Aloud. Trials for each participant are randomly sampled from the 'choice13k' dataset (Peterson et al., 2021).

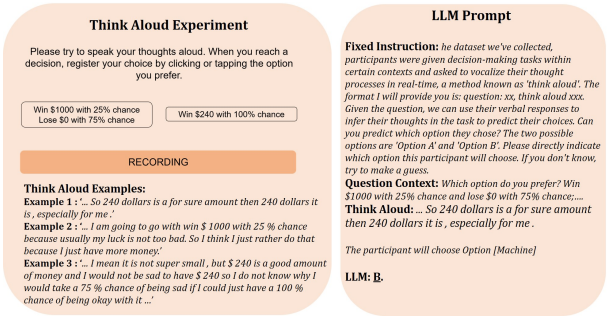

Figure 1: Experiment setting and LLM Prompt. An example trial setting of collecting both choice behavior and Think Aloud data. Participants were instructed to speak out their thoughts before pressing a button. Prompts to LLMs consist of question contexts and corresponding Think Aloud.

LLMs Outperform Symbolic Models and Neural Network Models in Predictive Performance and Learning Efficiency

To assess the predictive accuracy of LLMs using Think Aloud data on behaviors, we input the settings, questions, and Think Aloud responses from each trial into the models. We selected LLaMA2 in various model sizes (7B, 13B, and 70B) to explore the impact of model capacity (Touvron et al., 2023). We also include a recent version of GPT-4 (GPT4-Turbo-0125-preview, Achiam et al. (2023)). We use LLM to do **Zero-**

Shot learning to acquire the results. Compared to LLMs, we picked the two most representative symbolic models, Expected Utility Theory (Von Neumann & Morgenstern, 2007) and Prospect Theory (Kahneman & Tversky, 1979). We also include best-performed neural network models (Value-Based Model and Context-Depedent Model) reported from Peterson et al. (2021). To ensure a fair comparison, we employed binary cross-entropy loss to individually optimize the parameters of each symbolic model. We adopted the neural network model structure from the original study, adding an individual embedding layer with 16 neurons to better capture individual differences in the value function.

In the smaller dataset from Experiment 1, we used Leave-One-Trial-Out cross-validation that allows each participant to leave out the same trial for testing. As Figure 2 illustrates, nearly all LLMs outperform the symbolic and neural network models in predicting human choices based on test likelihoods. Among the LLMs, performance improves with increasing model size, suggesting that greater language model capacity enhances their effectiveness as cognitive models with the Think-Aloud content. We also controlled the LLM input by providing only the question context without Think Aloud data, which significantly reduced performance (e.g., GPT-4-0125-Preview: $83.0 \pm 5.8\%$ vs. $51.1 \pm 21.6\%$). This result suggests that Think Aloud data contributes considerably to the superior predictive performance in LLM.

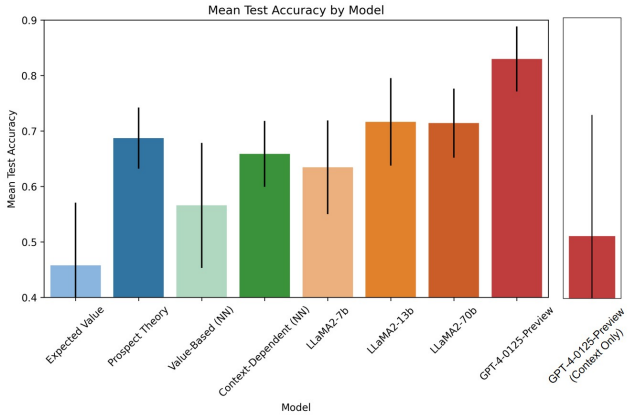

Figure 2: Test Accuracy in Smaller Dataset. LLMs outperform traditional SOTA cognitive symbolic models and neural network models in **Zero-Shot** training. The performance in LLM with only context information drops drastically, suggesting the importance of Think Aloud data in understanding the behaviors.

To further evaluate performance and learning efficiency, we conducted a similar analysis using a large dataset from Experiment 2. In this dataset, we reserved 10% of the data as a test set and incrementally added data from the remaining training set for each training step. We increased the training data in 20 incremental steps, sampling different subsets 10 times at each step. The result shows that both symbolic models and neural network models can become better at predicting behaviors in the test set, but their performance is still far from the best LLM performance even with the full training data. Most importantly, these LLM results were obtained on a **Zero-Shot** basis, highlighting their superior learning efficiency and generalizability as cognitive models. Finally, we fine-tuned LLaMA2-70b only with their final layer's embeddings to fit the dataset. With the same progressive setting, LLaMA2-70b shows super training efficiency that with only 5% of training data, it could surpass the zero-shot performance of three LLMs, and reach an accuracy of $(69.4 \pm 0.4\%)$, equaling around 85% of training data in the Prospect Theory Model. This result demonstrates that LLMs equipped with Think Aloud data can be further optimized with minimal data, enhancing both their predictive accuracy and training efficiency.

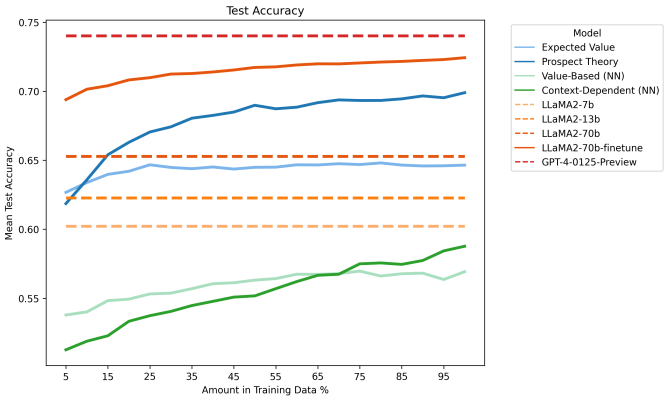

Figure 3: Progressive Test Accuracy in Larger Dataset. LLMs with Think Aloud show robustly higher performance than other types of models in the larger dataset. Fine-tuning on LLaMA2-70b shows super training efficiency to be a strong cognitive model.

Discussion

LLMs with Think Aloud as A New Form of Cognitive Models LLMs enhanced with the Think Aloud data can serve as a new form of cognitive models. In our study, LLMs with Think Aloud data exhibit strong predictive performance in **Zero-Shot** training and high learning efficiency during fine-tuning, surpassing both SOTA symbolic cognitive models and neural network models from previous research. This suggests that Think Aloud data does contain a deeper insight into human cognitive processes, compared to raw behavioral data, which LLMs can well decode.

Limited Interpretability However, despite the boosted performance, the interpretation of Think Aloud data remains poorly understood. It's challenging to separate semantic representations and cognitive representations of Think Aloud from LLM. A recent work has proposed an approach mapping text embeddings of Think Aloud to decision-related embeddings with a neural network model (Xie, Xiong, & Wilson, 2023), which could be the potential to make Think Aloud identifiable in cognitive space.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... others (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kahneman, D., & Tversky, A. (1979). 0prospect theory: An analysis of decision under risk, 1 econometrica. *March*, *47*(2), 2635291.

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492–28518).

Simon, H. A., & Ericsson, K. A. (1984). Protocol analysis: Verbal reports as data.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... others (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Von Neumann, J., & Morgenstern, O. (2007). Theory of games and economic behavior: 60th anniversary commemorative edition. In *Theory of games and economic behavior.* Princeton university press.

Xie, H., Xiong, H., & Wilson, R. C. (2023). Text2decision: Decoding latent variables in risky decision making from think aloud text. In *Neurips 2023 ai for science workshop*.