UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Predicting Human Choice Between Textually Described Lotteries

Permalink

https://escholarship.org/uc/item/7838t9zr

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors

Marantz, Eyal

Plonsky, Ori

Publication Date

2025

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Predicting Human Choice Between Textually Described Lotteries

Eyal Marantz (Eyalmarantz@campus.technion.ac.il)
Faculty of Data and Decision Sciences
Technion - Israel Institute of Technology, Haifa, 3200003, Israel

Ori Plonsky (Plonsky@technion.ac.il)
Faculty of Data and Decision Sciences
Technion - Israel Institute of Technology, Haifa, 3200003, Israel

Abstract

Predicting human decision-making under risk and uncertainty is a long-standing challenge in cognitive science, economics, and AI. While prior research has focused on numerically described lotteries, real-world decisions often rely on textual descriptions. This study conducts the first large-scale exploration of human decision-making in such tasks using a large dataset of one-shot binary choices between textually described lotteries. We evaluate multiple computational approaches, including fine-tuning Large Language Models (LLMs), leveraging embeddings, and integrating behavioral theories of choice under risk. Our results show that fine-tuned LLMs, specifically GPT-4o, outperform hybrid models that incorporate behavioral theory, challenging established methods in numerical settings. These findings highlight fundamental differences in how textual and numerical information influence decision-making and underscore the need for new modeling strategies to bridge this gap.

Keywords:

Decision making; Artificial Intelligence; Machine Learning; Natural Language Processing; Computational modeling

Introduction

Predicting and understanding human choice under uncertainty is a fundamental challenge in economics, psychology, and the cognitive sciences, with clear implications for many real-world scenarios, including financial investments, health-related choices, and risk management. Most of the systematic study in this domain has focused on investigating how people choose between lotteries or gambles, with these lotteries explicitly and accurately described using numerical format. This line of research, which goes back more than eight decades, assumes that the response to these numerical descriptions captures the basic properties of human decision making under risk and uncertainty. Therefore, the insights gained in these studies should generalize to more natural settings. Importantly, many of the most important insights such research reveals concern the ways by which people seem to deviate from clear theoretical benchmarks like maximization of Expected Value or of Expected Utility. It is convenient that the numerical format of presentation thus allows computing the predictions of these benchmarks.

Yet, in the real world, people rarely face precise numerical descriptions of choice options. Instead, potential options may often be described using natural language. For example, people may face signs that warn against choosing certain options or ads that promote the choice of other options. That is, in many real-world situations, rather than relying on precise numerical information, individuals must rely on qualitative descriptions and make subjective interpretations of textual information before reaching a decision. In this paper, we investigate—and try to predict—people's decisions between textually described choice options that do not contain precise numerical information.

Under a textual description format, almost any behavior may be considered "rational" (i.e., adhering to the prescriptions of expected value or utility maximization). For example, Figure 1 presents a binary choice task presented in two formats. Under a numerical format, the task has a clear theoretical prediction: Option B that provides "5 with probability .23; 2 otherwise" dominates—and should be chosen over—Option A that provides "1 for sure". Yet, when described textually, this no longer holds. While the textual descriptions are accurate (in the sense that they faithfully describe the underlying payoff distributions), the choice of Option A (*This option may seem appealing for its consistency, but it cannot offer any surprisingly high rewards*) over B (*This alternative holds an advantage for the risk-takers who seek the excitement of a larger possible gain*) is quite reasonable and depends on both subjective interpretations of the texts and on idiosyncratic preferences. Under the textual format, it is also quite hard to elicit clear predictions of extant computational models of choice that lack the ability to process the textual inputs.

Lacking clear benchmarks, we chose to start the investigation of this domain with a prediction-based study. Using a recently collected dataset of 1000 one-shot binary choice tasks, **TextualChoices-1K** (Erev, Plonsky, Marantz, & Roth, in preperation) we conduct the first large-scale exploration of human decision-making in tasks framed through textual descriptions, rather than numeric lotteries. We systematically test various computational approaches, all of which use Large Language Models (LLMs) that can accept the textual descriptions as input. Our study contrasts and compares different ways to use LLMs to predict behavior in this task, including both purely data-driven methods and approaches that aim to enhance the predictive ability of the LLMs with extant behavioral theories of choice under risk and uncertainty. In so doing, we also aim to bridge the gap between extant numeric-focused models and modern language-based decision frameworks, advancing our understanding of hu-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

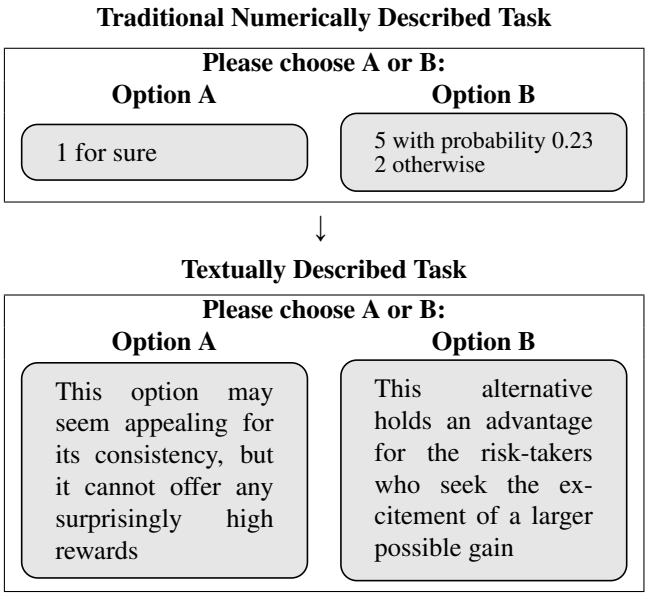


Figure 1: Comparison of Numerical and Textual Task Descriptions

man decision-making under uncertainty while highlighting the strengths and limitations of LLMs in this context.

Recent works on predicting numerically described tasks revealed that hybrid methods that complement data-driven computational methods with behavioral theories lead to the most accurate models of choice prediction (Plonsky et al., 2024). In contrast, our findings revealed that behavioral-theory-free machine learning models outperform theory-driven models in predicting decisions based on textual descriptions. This has led us to test similar data-driven methods, namely fine-tuning of LLMs, in numerically described tasks. Our results suggest hybrids of behavioral theory and machine learning still outperform the pure LLM approach in these settings.

This divergence may hint of a fundamental difference in the choice processes involved in numerically vs. textually described options. While modern computational models excel at interpreting and predicting decisions based on natural language cues, they face challenges when precision and numeric reasoning are required. These findings cast doubt on the assumptions underlying much of the classical behavioral research on choice under risk and uncertainty and underscore the need for task-specific strategies in computational modeling, tailoring predictive approaches to the structure of the decision problem.

Related Work Our work is related mainly to two lines of research. First, it relates to studies that aim to predict human decision-making using behavioral models, ML, or a combination of both. Historically, models such as Expected Utility Theory assumed that individuals make decisions by maximizing utility. However, decades of empirical research have shown systematic deviations from this rational framework. This led to the development of behavioral models like Prospect Theory (Kahneman & Tversky, 1979) and many others, including Best Estimate and Sampling Tools (BEAST) that has shown high accuracy in predicting choice under risk uncertainty (Erev, Ert, Plonsky, Cohen, & Cohen, 2017).

More recently, machine learning (ML) techniques have been combined with behavioral theories to create hybrid models that improve predictive accuracy (Peterson, Bourgin, Agrawal, Reichman, & Griffiths, 2021; Plonsky, Erev, Hazan, & Tennenholtz, 2017). For example, BEAST-GB (Plonsky et al., 2024), integrates behavioral insights based on the model BEAST with ML tools to achieve state-of-the-art predictive performance for choice between numerically described choice options. While these approaches have advanced decision-making research in settings that involve explicit numeric outcomes and probabilities, they primarily focus on such structured scenarios, leaving a gap in understanding decision-making in less structured, real-world tasks.

Second, our work is related to recent studies that use LLMs to mimic, augment, or predict human behavior. Advances in LLMs have demonstrated their capacity to process and interpret qualitative information effectively. For example, CENTaUR (Binz & Schulz, 2023; Binz et al., 2024) and Arithmetic-GPT (Zhu, Yan, & Griffiths, 2024) have shown that LLMs can accurately predict human decisions in numeric and arithmetic contexts. However, challenges remain, as LLMs often default to overly rational behavior and struggle with inconsistencies in reasoning (R. Liu, Geng, Peterson, Sucholutsky, & Griffiths, 2024; Macmillan-Scott & Musolesi, 2024). Our work examines the usefulness of LLMs for prediction of human choice when clear benchmarks of behavior are lacking.

Dataset

Human decision-making under risk and uncertainty is often studied through tasks involving choices between m lotteries (or gambles), $\{L_1, L_2, \ldots, L_m\}$, where for each $i \in [m]$, L_i is defined by N possible payoffs $\{x_i^m\}_{i=1}^N$ and their respective probabilities $\{p_i^m\}_{i=1}^N$. Whereas traditionally, the options' payoff distributions are explicitly and numerically described, we study choices where these lotteries are described using free text. The dataset we use, **TextualChoices-1K** (Erev et al., in preperation), includes 1,000 tasks of choice between $m = 2$ lotteries labeled *Option A* and *Option B*. To create this dataset, (numeric) payoff distributions for the choice tasks were first randomly sampled from a large space. Then, an LLM converted these distributions to natural language, avoiding direct references to specific payoffs or probabilities. Multiple descriptions were generated for each option, with one randomly selected for inclusion in the dataset. Further details on the creation of the dataset is given in (Erev et al., in preperation) and in Part I of the online Supplementary Material (SM).

Each textually-described choice task was completed by, on average, 31 participants, recruited using Prolific. Each par-

ticipant completed 5 tasks, making a single decision without feedback. Participants were incentivized to maximize earnings: Their bonus payment depended on the realized payoffs from the options they selected. Our study aims to predict the proportion of participants who chose *Option A* based solely on the textual descriptions of each option.

Method

We explored four complementary approaches to using large language models (LLMs) to predict human decisions under risk: (1) fine-tuning LLMs (Binz et al., 2024; Jeong, 2024) on **TextualChoices-1K**; (2) leveraging pre-trained embeddings with regression models (Binz & Schulz, 2023); (3) "LLMs as subjects" - prompting LLMs to act as decision makers (R. Liu et al., 2024; Shapira, Madmon, Reichart, & Tennenholtz, 2024); and (4) extracting interpretable behavioral features that past research has suggested are central to choice under risk and uncertainty from the textural descriptions.

Experimental Setup

We used **TextualChoices-1K** as the main dataset throughout our experiments. Across all approaches, we allocated 90% of the dataset ($N = 900$) for training and validation, reserving 10% ($N = 100$) as a fixed held-out test set used consistently for model evaluation. Model selection (e.g., hyperparameter tuning, early stopping) was conducted using the validation subset of the training set. For all predictive models, we report: (a) mean-squared error (MSE) between predicted and observed proportions of Option A choices, and (b) directional accuracy—i.e., whether the prediction and observed label fall on the same side of the 0.5 threshold.

Some of our approaches involved training regression models to predict human choices based on either data representations (e.g., embeddings) or structured outputs from LLMs. We evaluated a range of regression techniques, including Linear Regression, Ridge, Lasso, Support Vector Regression (SVR), XGBoost, K-Nearest Neighbors (KNN), and Multi-Layer Perceptrons (MLPs). While all were explored during development, only the most relevant and best-performing models are reported in the results section.

Incorporating Psychological Theory

Recent research highlights the benefits of hybrid models that combine psychological theory with machine learning techniques to improve both predictive accuracy and interpretability. In this work, we investigate several such integrations, focusing on the BEAST model (Best Estimate and Sampling Tools) (Erev et al., 2017)—a highly successful behavioral model developed to explain and predict decisions under risk and uncertainty. BEAST models choice as the outcome of a partially biased mental sampling process, modulated by sensitivity to expected values. It has been shown to capture 14 well-documented choice anomalies and has won two international choice prediction competitions (Plonsky et al., 2024; Erev et al., 2017).

Furthermore, BEAST has served as the foundation for previous hybrid approaches that combined psychological features with machine learning algorithms to predict human choice in numerically described lotteries (Plonsky et al., 2017; Bourgin, Peterson, Reichman, Griffiths, & Russell, 2019; Plonsky et al., 2024). Following this line of work—which has achieved state-of-the-art results on the largest available datasets—we adopt BEAST as the core theoretical framework for guiding and augmenting our prediction models. In what follows, we describe how BEAST is incorporated into multiple components of our workflow. We incorporated BEAST in three ways:

1. **BEAST-labeled pretraining:** In addition to fine-tuning solely on **TextualChoices-1K** (Approach 1), we also experimented with pretraining models on a synthetic dataset (N=20,000) of numerically described choices labeled by BEAST.

2. **Personality-driven prompting:** For LLMs-as-subjects (Approach 3), we designed prompts to encode BEAST-inspired "personalities" aligned with cognitive dimensions.

3. **Feature extraction:** In Approach 4, we proposed an alternative approach that used LLMs to extract BEAST-inspired features from tasks, following (Plonsky et al., 2017), and trained a regression model on these representations.

Approach 1: Fine-Tuned LLMs

We fine-tuned multiple pre-trained LLMs based on the training data. We utilized BERT-based models, including BERT (Devlin, Chang, Lee, & Toutanova, 2019), RoBERTa (Y. Liu, 2019), and DeBERTa (He, Liu, Gao, & Chen, 2021), due to their ability to generate rich, context-aware text representations that are well-suited for regression and predictive modeling. Additionally, we trained OpenAI's GPT-4o and GPT-4o-mini (OpenAI, 2023), leveraging their advanced capacity to interpret complex textual patterns and perform qualitative reasoning, making them highly adaptable across diverse predictive scenarios.

Some fine-tuned models, such as GPT-4o and GPT-4o-mini, produce stochastic outputs during inference. To mitigate this variability, we generated 10 predictions for each sample and averaged them. Formally, for a sample i, the prediction in these cases is: $\hat{p}_i = \frac{1}{10}\sum_{j=1}^{10} \hat{p}_i^j$.

Additional Training Data Because the size of the **TextualChoices-1K** dataset is limited, we explored two strategies for incorporating additional data into the training (i.e. fine-tuning) phase. Notably, to our knowledge, no other dataset of choice between textually described options exists. We thus chose to supplement the pre-training phase with data on choice between numerically described lotteries. First, we pre-trained our model with real data on choices between lotteries, using the large numerical dataset, **Choices13k** (Peterson et al., 2021). Here, we used the 1039 choice tasks that do not include feedback and ambiguity, to

align with our experimental setting. Of these, we used 935 for training and validation and 104 as the held-out test set. Second, we also tried pretraining using a large synthetic dataset that we generated specifically for this study (N = 20,000). The labels for this dataset were derived from the BEAST model (Erev et al., 2017), a strong behavioral model rooted in psychological theory.

Approach 2: Text Embeddings

We transformed the textual data into numerical embeddings using OpenAI's *text-embedding-ada-002* and *text-embedding-3-large* models. These embeddings capture semantic relationships in continuous vector spaces, enabling downstream regression tasks. *text-embedding-ada-002* emphasizes efficiency and cost-effectiveness, while *text-embedding-3-large* offers richer semantic representation with higher dimensionality. For each task, we transformed the description of each option into an embedding vector, denoted as $\mathbf{v}_A$ for *Option A* and $\mathbf{v}_B$ for *Option B*. To capture the relationship between the options, we computed the task representation as the difference between the two embedding vectors: $\mathbf{d} = \mathbf{v}_A - \mathbf{v}_B$ where $\mathbf{d}$ represents the embedding difference vector for the task. Using this task representation,[1] we applied various regression techniques, as described above, to predict the outcome.

We also investigated the effect of using PCA to reduce the dimensionality of the embedded vectors on regression performance. Dimensionality reduction helps mitigate computational costs and overfitting, especially with high-dimensional data. PCA transforms data into a set of orthogonal components ranked by their contribution to variance. Using PCA, we retained 5%, 10%, 25%, and 33% of the original dimensions and evaluated the trade-off between model complexity and predictive accuracy across these dimensions.

Approach 3: LLMs as Simulated Subjects

We designed an experimental framework where LLM agents acted as *"experimental subjects"*. Each agent faced and provided its choices for 50 of the choice tasks (See Figures A.3, A.4, A.5, in the SM. The responses from all agents were aggregated for each task to generate the final LLM's prediction. Then, a regression model was trained to learn the relationship between the LLM's predictions and human choices, providing an optimized mapping between the two.

Prompting Conditions The LLM agents' responses were elicited under three distinct prompting conditions. In the *Binary* condition, the LLM made a direct choice between the two options. In the *Percentage* condition, the LLM provided a continuous preference score between 0 and 100. Finally, the *Confidence* condition required the LLM to make a binary choice and then assign that choice a confidence level (0–100), which was used for predictions.

[1] Other representations were tested, but we focused on vector difference as it performed best.

Personalities To investigate the influence of psychological theory on the model's performance, we developed ten distinct *Personalities*, each reflecting an assumption (or a combination of assumptions) embedded in the BEAST model (Erev et al., 2017). For instance, one of BEAST's assumptions is that people are sometimes more sensitive to the sign of the reward (gain or loss) than the actual values. Accordingly, one of the personalities is *The Guardian*, which was defined to behave as someone who is "Sensitive to gains vs. losses, impacting risk tolerance". The interpretations and details of these personalities are presented in Table A.12 in the SM.

For comparison, we included a baseline model where all agents operated without any assigned personality. To improve predictions, we aggregated the outputs from each predefined personality profile and trained a weighted regression model, where each personality contributes to the final prediction according to its optimized weight. This approach captures the collective predictive power of the different personalities while accounting for their unique contributions to overall prediction performance.

Approach 4: Theory-Guided Feature Extraction

We used the LLM to extract from the textual descriptions numeric values for behavioral features, transforming the task into a numerical prediction task with a well-established solution. Building on the work of Plonsky et al. (2024), which demonstrated that human choices can be effectively predicted using ML and features derived from the behavioral model BEAST, we aimed to extract a set of features that capture various elements of BEAST. For instance, one of BEAST's assumptions is that people tend to exhibit *pessimism*, expecting the worst possible outcome. To reflect this, we extracted a "worst-case" feature, which identifies the option with the better payoff under the worst-case scenario. All the extracted features appear in Table A.14 in the SM.

The primary objective was not to assess the accuracy of the LLM's feature extraction but to ensure that its process mirrored human-like reasoning. For instance, when a description emphasized disadvantages, it was reasoned that human subjects might "extract" a set of perceived values different from the actual numerical values (which were unknown to them) and base their decisions on these perceptions.

To implement this, we designed specific prompts for each feature and instructed the LLM to classify which option was preferred under the assumption of that feature. To account for ambiguity, we allowed the LLM to provide a neutral response when no clear preference could be inferred. The results were aggregated and converted into numeric scores, which were then trained using a regression model (as mentioned above) for final predictions.

Evaluation on Numerical Tasks

We also evaluated how some of the best models perform with tasks involving numeric descriptions. To do so, we used the **Choices13k** (Peterson et al., 2021) dataset, the largest dataset of risky choice publicly available. Of this dataset, we used

the subset of tasks that excluded feedback and ambiguity to match our experimental conditions. 90% (N = 935) of this set was used for training and validation while the rest of the data (N=104) was used as a held-out set. We fine-tuned GPT-4o on this dataset and, as a benchmark, also trained BEAST-GB (Plonsky et al., 2024), which is currently considered state-of-the-art in this numerical description setting.

Results

We present the results of all different models we tried in the SM. Here, we focus on the best set of models within each approach. Table 1 presents these results.[2] Fine-tuning of LLMs outperformed alternative methods, highlighting its effectiveness in adapting pre-trained linguistic representations to the task. Fine-tuning a GPT-4o model achieved strong results, with an MSE of `0.0121` and an accuracy of `0.87`. Incorporating numerical data into the training process further improved performance, reducing the MSE to `0.0110`, the best result achieved in terms of MSE. In contrast, adding synthetic BEAST data slightly degraded performance, increasing the MSE to `0.0123`.

Using embeddings extracted from textual descriptions and training traditional machine learning models such as MLP and Ridge regression represented the best non-fine-tuning approach, achieving MSEs of `0.0138` and `0.0159`, respectively. The "LLM as Subjects" approach, where out-of-the-box LLMs or BEAST-personalities were prompted directly, resulted in higher MSEs (`0.0170` and `0.0220`). Feature extraction based on BEAST-derived representations performed worst, with a relatively high MSE of `0.0395`.

Other fine-tuned models are also presented in Table 1. GPT-4o-mini, the smaller variant of GPT-4o, performed slightly worse than the larger model, achieving an MSE of `0.0130` with a similar accuracy of `0.87`, yet still outperforming all other models. Traditional transformer models such as RoBERTa and DeBERTa demonstrated decent performance, with RoBERTa reaching an MSE of `0.0169` and DeBERTa achieving `0.0162`. BERT lagged behind, with a substantially higher MSE of `0.0260`. Interestingly, the accuracy of DeBERTa was found to be highest of all models we tried.

We further analyzed the errors of three models: GPT-4o, our best overall model; the best embeddings model, which was the second-best approach; and DeBERTa, chosen for its high directional accuracy. Table 2 shows that for tasks with non-extreme target values (i.e., between 0.2 and 0.8), GPT-4o and the embeddings model performed similarly but in tasks with extreme target values, GPT-4o was clearly better. DeBERTa also performed better on extreme tasks than on non-extreme ones. These results suggest that GPT-4o's overall advantage may stem from its ability to handle extreme decisions more effectively.

Finally, to evaluate the robustness of fine-tuned LLMs when applied to numerically described gambles, we tested the best models on the numeric dataset Choices13k (Table 3). Here, GPT-4o maintained relatively strong performance with an MSE of `0.0104` and an accuracy of `0.89`. However, it still underperformed compared to BEAST-GB (Plonsky et al., 2024), a hybrid model combining behavioral theories with ML, which achieved a lower MSE of `0.0092`.

Discussion

Human choice under risk has been extensively studied for decades, but this research has predominantly focused studying tasks with accurate numeric descriptions. This approach, while valuable, does not fully capture the richness and complexity of real-world decisions, which often involve potentially ambiguous textual information. We take an important step by examining choice behavior in textually described contexts, offering a closer approximation of how people navigate decisions in naturalistic settings. Our findings reveal important differences between these two domains, highlighting their distinct challenges and opportunities for behavioral theories and for ML models.

Our findings reveal a significant gap between textual and numeric decision-making tasks. While theory-free ML approaches excelled in the textual domain, a hybrid of behavioral theories and ML, specifically BEAST-GB, demonstrated its continued advantage in the numeric setting. This discrepancy highlights potentially fundamental differences in how textual and numeric data are processed. Textual descriptions often include interpretive ambiguity, allowing language models to leverage fine-tuning for task-specific optimization. Numeric data, by contrast, benefits from the structured assumptions provided by behavioral theories, which align well with predefined, explicit representations of choices.

We find that fine-tuning GPT-4o achieved the best performance on **TextualChoices-1K**. These results became even stronger when we incorporated additional pretraining on numerical data, which may imply that despite the aforementioned potential differences between the underlying processes involved in choices between textually and numerically described lotteries, they also share some similarities, and people's choices in one type of tasks are associated with their choices in the other. Future research should focus on the key similarities and differences between the two types of tasks.

Furthermore, GPT-4o demonstrated remarkable robustness across both textual and numerical tasks. Although it did not achieve the top score in the numerical domain, where BEAST-GB outperformed it, it still performed competitively. This resilience, particularly when leveraging numerical and BEAST synthetic data, highlights GPT-4o's ability to handle diverse and noisy data sources. We attribute this strength to its broader pretraining and superior generalization capacity. Overall, these findings underscore GPT-4o's versatility and position it as a strong candidate for general-purpose decision-making tasks across a wide range of domains.

Despite the success of BEAST-GB in numeric tasks, attempts to integrate psychological theory into textual decision-

[2] A previous version of this paper mistakenly reported different results due to a data processing error.

Table 1: Models Performance on **TextualChoices-1K** (Textually Described Choices)

Approach	Model	Training Data	Test MSE	Test Accuracy
Fine-Tuning	BERT	Textual only	0.0260	0.84
	RoBERTa	Textual only	0.0169	0.87
	DeBERTa	Textual only	0.0162	0.92
	GPT-4o-mini*	Textual only	0.0130	0.87
	GPT-4o*	Textual only	0.0121	0.87
		Textual + Numerical	**0.0110**	0.88
		Textual + Synthetic BEAST	0.0123	0.85
Embeddings	MLP	Textual only	0.0138	0.89
	Ridge	Textual only	0.0159	0.88
LLM as Subjects	Out-of-box LLM	–	0.0170	0.87
	BEAST-personalities LLM	–	0.0220	0.81
Feature Extraction	XGBRegressor	BEAST-derived model	0.0395	0.73

Note: * Results are based on stochastic models averaged over 10 inference runs.

Table 2: MSE by target extremity.

Model	Extreme (n=29)	Non-Extreme (n=71)
GPT-4o	0.0066	0.0128
Embedding	0.0136	0.0127
DeBERTa	0.0140	0.0172

Note: Extreme denotes tasks with target values below 0.2 or above 0.8.

Table 3: Comparison of model's performance on Numeric Dataset **Choices13k**

Model	Test MSE	Test Accuracy
BEAST-GB	0.0094	0.89
GPT-4o	0.0104	0.89

making were less effective. BEAST-derived models and synthetic data did not enhance performance compared to theory-free versions of the same models. Feature extraction, which is a fully theory-driven method approach performed particularly poorly. This is surprising given that psychological theory has historically improved predictive accuracy in numeric settings. One possible explanation is that the richness and complexity of textual data dilute the utility of predefined behavioral constructs, which are inherently designed for structured numeric inputs. It is important to note that all our approaches to incorporate psychological theory were based on the model BEAST. Hence, our results do not necessarily imply that integrating behavioral theory based on different models or theories would also be ineffective. However, BEAST has a strong track record in numerical settings and, even in our analysis, BEAST-based models outperformed all other models, highlighting its strengths in structured, quantitative tasks. Furthermore, when using BEAST as a foundation for LLM personalities or feature extraction, our approach may not have effectively captured key elements of the model, as some aspects are non-trivial to process. Adapting such frameworks to qualitative, language-based representations remains a significant challenge.

These results underscore the need to develop hybrid models better suited for textual tasks, combining insights from behavioral theories with the capabilities of modern LLMs. One promising avenue is to refine feature engineering to align behavioral constructs with the nuances of textual data. Additionally, exploring how LLMs process qualitative, ambiguous information could yield valuable insights into computational decision-making models. Future research should also investigate how task-specific fine-tuning can be further optimized to bridge the gap between textual and numeric settings.

While this study provides valuable insights, some limitations should be noted. The relatively small size of **TextualChoices-1K** may limit the generalizability of the findings, particularly for complex models like GPT-4o. Additionally, the inherent differences between controlled numeric tasks and naturalistic textual descriptions may pose challenges for direct comparisons. Finally, it is important to acknowledge that we have not tested all possible LLMs, and as this field evolves rapidly, more advanced models may already exist or emerge in the near future. This highlights the need for ongoing research to evaluate and compare the latest advancements in ML for decision-making tasks.

Conclusion

Our work highlights the effectiveness of task-specific fine-tuning for textual decision-making tasks, with GPT-4o achieving state-of-the-art performance. However, the gap between textual and numeric settings, along with the challenges of incorporating psychological theory, points to the need for further research. By bridging these gaps, future studies can advance our understanding of human decision-making and improve the predictive capabilities of computational models.

References

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., ... others (2024). Centaur: a foundation model of human cognition. *arXiv preprint arXiv:2410.20268*.

Binz, M., & Schulz, E. (2023). *Turning large language models into cognitive models.* Retrieved from https://arxiv.org/abs/2306.03917

Bourgin, D. D., Peterson, J. C., Reichman, D., Griffiths, T. L., & Russell, S. J. (2019). *Cognitive model priors for predicting human decisions.* Retrieved from https://arxiv.org/abs/1905.09397

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*, *124*(4), 369.

Erev, I., Plonsky, O., Marantz, E., & Roth, Y. (in preperation). choice between verbally described lotteries.

He, P., Liu, X., Gao, J., & Chen, W. (2021). Deberta: Decoding-enhanced bert with disentangled attention. In *International conference on learning representations (iclr).*

Jeong, C. (2024). Fine-tuning and utilization methods of domain-specific llms. *arXiv preprint arXiv:2401.02981*.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263–291.

Liu, R., Geng, J., Peterson, J. C., Sucholutsky, I., & Griffiths, T. L. (2024). *Large language models assume people are more rational than we really are.* Retrieved from https://arxiv.org/abs/2406.17055

Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, *364*.

Macmillan-Scott, O., & Musolesi, M. (2024). (ir) rationality and cognitive biases in large language models. *Royal Society Open Science*, *11*(6), 240255.

OpenAI. (2023). *Gpt-4 technical report.* Retrieved from https://openai.com/research/gpt-4

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.

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., ... others (2024). Predicting human decisions with behavioral theories and machine learning. *arXiv preprint arXiv:1904.06866*.

Plonsky, O., Erev, I., Hazan, T., & Tennenholtz, M. (2017). Psychological forest: Predicting human behavior. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 31).

Shapira, E., Madmon, O., Reichart, R., & Tennenholtz, M. (2024). Can large language models replace economic choice prediction labs? *arXiv preprint arXiv:2401.17435*.

Zhu, J.-Q., Yan, H., & Griffiths, T. L. (2024). *Language models trained to do arithmetic predict human risky and intertemporal choice.* Retrieved from https://arxiv.org/abs/2405.19313