CAN LLMS REPLACE ECONOMIC CHOICE PREDICTION LABS? THE CASE OF LANGUAGE-BASED PERSUASION GAMES

Eilam Shapira* **Omer Madmon** **Roi Reichart** **Moshe Tennenholtz**
The Faculty of Data and Decisions Science, Technion - Israel Institute of Technology

ABSTRACT

Human choice prediction in economic contexts is crucial for applications in marketing, finance, public policy, and more. This task, however, is often constrained by the difficulties in acquiring human choice data. With most experimental economics studies focusing on simple choice settings, the AI community has explored whether LLMs can substitute for humans in these predictions and examined more complex experimental economics settings. However, a key question remains: can LLMs generate training data for human choice prediction? We explore this in language-based persuasion games, a complex economic setting involving natural language in strategic interactions. Our experiments show that models trained on LLM-generated data can effectively predict human behavior in these games and even outperform models trained on actual human data.

1 INTRODUCTION

In the digital economy, online platforms have become central to the interaction between consumers and service providers. These platforms, such as e-commerce websites, travel booking sites, and online marketplaces, facilitate a dynamic exchange where service providers present their offerings and consumers make purchasing decisions based on the provided information. A specific example of such an interaction is seen on Booking.com, a popular travel booking platform. On Booking.com, hotel owners (sellers) aim to persuade potential customers to choose their hotels by presenting information, often expressed in natural language, such as textual descriptions and reviews.

These interactions are often repeated, with sellers employing different strategies to attract and retain customers. For instance, some sellers might consistently highlight positive reviews to appeal to potential buyers, even when there are some downsides to their services, a practice can be seen as a `Greedy` persuasion strategy. Other sellers may adopt a more careful approach that aims to build long-term trust and maintain a solid reputation, by presenting both positive and negative aspects of the hotel, conditional on its true quality. This can be seen as adopting an `Honest` persuasion strategy. A platform such as Booking.com is often interested in accurately anticipating consumers' behavior when facing different types of persuasion strategies employed by sellers. By accurately predicting behavior, the platform can asses the expected satisfaction and engagement of its users, and the impact of particular sellers on the overall users' welfare.

In the economic literature, the concept of persuasion has been extensively studied, particularly within the framework of *persuasion games* (Aumann et al., 1995; Farrell & Rabin, 1996; Kamenica & Gentzkow, 2011). These games involve strategic interactions where a *sender*, possessing private information (the actual quality of the hotel, in our example), aims to influence the decision of a *decision-maker* through selective information disclosure (in our case, selection of the review to be presented). Reputation indeed plays a significant role in repeated persuasion games, as demonstrated in previous work (Kim, 1996; Aumann & Hart, 2003; Best & Quigley, 2022; Arieli et al., 2024).

Traditional economic models, however, often abstract these interactions into simplified messages, lacking the nuance and complexity of natural language communication.[1] This abstraction limits the

*Corresponding author: eilamshapira@campus.technion.ac.il.

[1]In economic modeling, the message typically influences the receiver's beliefs solely through the application of Bayes' rule. The content of the message itself is usually abstracted away, meaning that the specific language

applicability of these models to real-world scenarios where language plays a critical role. Consequently, there is a growing need for interdisciplinary research to better understand and predict human decision-making in such contexts.

Indeed, the study of *language-based persuasion games* has recently gained popularity within the natural language processing community. While some previous work focused on optimizing the sender's strategy (Raifer et al., 2022), another important, complementary line of research focused on predicting the behavior of human decision-makers against a given set of reasonable persuasion strategies (Apel et al., 2022; Shapira et al., 2023). Importantly, even without considering a particular business application (e.g., Booking.com), predicting human behavior in strategic interactions has a huge intellectual value for the field of behavioral economics, and its understanding will contribute to the overall understanding of human behavior.

In the language-based persuasion game considered by these works (and first introduced by Apel et al., 2022), a travel agent (expert) is trying to persuade a decision-maker (DM) towards accepting their hotel offer, by presenting the decision-maker with a textual review of the hotel, selected from a given available set of reviews. The true quality of the hotel is the expert's private information, and the DM benefits from accepting the deal only if the hotel is of high quality.[2] The game consists of several rounds played between the same (bot) expert and (human) DM pair, which means that a DM facing a specific expert strategy can potentially learn and adapt over time, based on past experience. Apel et al. (2022) was the first to introduce the study of human choice prediction in the above game, and employed various machine learning (ML) techniques to solve it. Shapira et al. (2023) then studied off-policy evaluation in a similar setup, i.e., predicting human decisions when faced with an expert strategy that was not observed during training time.

Importantly, existing methods for human choice prediction in language-based persuasion games, as well as in other economic contexts (Plonsky et al., 2017; Rosenfeld & Kraus, 2018; Plonsky et al., 2019), rely on training ML models on a human choice dataset. Unfortunately, the collection, storage and usage of human choice data are often fraught with various challenges: first, it requires the development of designated tools and environments (e.g., a mobile application with a user-friendly interface). Moreover, privacy and legal issues must be addressed to permit the collection, storage, and utilization of this data.[3] Human data collection is also a long and tedious process, frequently resulting in issues like participant inattention, which can compromise data quality and must be addressed. These challenges often lead to a process that is extremely inefficient, expensive, and time-consuming.

Meanwhile, Large Language Models (LLMs) have made significant progress in recent years, demonstrating capabilities across a broad spectrum of applications, including text summarization, machine translation, sentiment analysis, and more (Brown et al., 2020; Zhang et al., 2023; Peng et al., 2023; Susnjak, 2023; Wang et al., 2023; OpenAI, 2023). Moreover, recent study demonstrates how LLM-based agents can successfully function as decision-makers in economic and strategic environments, in which agents aim to maximize their gain from – in a complex, possible multiagent interaction (Xi et al., 2023). In the context of human choice prediction, using LLM-based agents to generate synthetic but realistic data represents a groundbreaking proposition. If LLMs can effectively mimic human behavior in these economic settings, they could offer a cost-effective, efficient, and scalable alternative to traditional methods for training human choice prediction models. Importantly, in most real-life scenarios, the generation of a large LLM-based sample is significantly easier than obtaining even a small human choice dataset.

Our contribution In this paper, we introduce a novel in-depth study of the predictive power of LLM-generated data to human choice prediction in language-based persuasion games. We show that a prediction model trained on a dataset generated by LLM-based players can accurately predict human choice behavior. In fact, it can even outperform a model trained on actual human choice data, for a large enough sample size. We also demonstrate how combining LLM-generated training data with actual human-generated data can further improve the accuracy of the prediction task.

---

or framing of the message does not play a role in the analysis. For example, two messages may be called 'good' and 'bad' for ease of exposition, but if they were called m_1 and m_2 nothing would have changed in the analysis, as long as the sender follows the same information revelation strategy in providing them.

[2]As explained in Section 2, the true quality of the hotel is determined by numerical scores associated with its textual reviews.

[3]See Acquisti et al. (2016) for a survey on the economics of privacy.

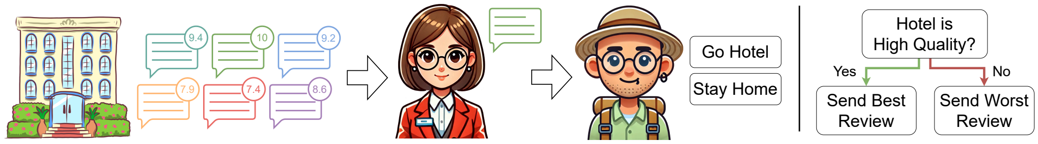


Figure 1: **Left:** Illustration of a single round in the language-based persuasion game. First, the expert observes the interaction history of previous rounds (does not appear in the illustration), as well as the current hotel's review-score pairs. She chooses a single review within this set according to her predefined strategy and sends it to the DM. Then, the DM observes this review (as well as the entire interaction history) and chooses an action. Lastly, both agents get their payoffs based on the DM's action and the hotel's true quality. **Right:** An example of an expert strategy.

2 Task Definition

The language-based persuasion game We begin by describing the language-based persuasion game presented by Apel et al. (2022) and Shapira et al. (2023), which is used to define our human choice prediction task. The game consists of two parties, an expert (she) and a decision-maker (DM; he), interacting for T rounds. At the beginning of each round, the expert is presented with R pairs of textual reviews and numerical scores (in the 1-10 range), describing a hotel the expert aims to promote in the current round. Denote by r_i^t and s_i^t the i'th textual review and score presented at time t, respectively. The expert's hotel in round t is considered of high quality if its average score is at least τ, and otherwise is considered of low quality. We define hotel's quality to be $q_t = \mathbf{I}\left(\frac{1}{R}\sum_{i=1}^{R} s_i^t \geq \tau\right)$, where $\mathbf{I}(\cdot)$ is the indicator function.

Then, the expert selects a single textual review r_i and sends it to the DM. Note that the full set of review-score pairs is the expert's private information and is not available for the DM. Upon observing the message, the DM decides whether to go to the hotel ($a_t = 1$) or not ($a_t = 0$). Lastly, the two players gain utility depending on the quality of the hotel q_t and the DM's action a_t. The expert's utility is simply given by $u(a_t) = a_t$, meaning she always gains if the DM goes to the hotel, regardless of its actual quality. The DM, however, only benefits from opting in when the hotel is of high quality. His utility function is given by $v(a_t, q_t) = \mathbf{I}(a_t = q_t)$. Figure 1 (left) illustrates a single round in the game. The success criteria for both players is gaining as high cumulative utility as possible throughout the rounds of the game.

Expert strategies We restrict our attention to the representative set of strategies considered first by Shapira et al. (2023). These strategies are simple and intuitive, and can be represented as binary decision trees. Figure 1 (right) demonstrates such an example strategy of the expert (this is the `Honest` strategy discussed in the introduction). We highlight that all other strategies also have intuitive behavioral interpretations, and they differ from each other by the extent to which each strategy balances between building trust and exploiting trust. We note that any expert strategy is either *naive* (the rule according to which the presented review is selected is independent of both the hotel's true quality and the interaction history); *stationary* (the selection rule depends solely on the hotel's quality, but not on the history); or *adaptive* (the rule depends on the interaction history). Importantly, our representative set of strategies contains strategies belonging to all three groups. For completeness, in Appendix B we formally introduce and motivate all of the strategies.

The human choice prediction task Given a dataset comprising multiple expert-DM interactions (i.e., multiple games, each consisting of multiple rounds), the task is to predict how other human DMs will engage in the game against the same experts that appear in the train set. Given a training dataset of expert-DM interactions in language-based persuasion games, the goal is to predict the behavior of human DMs against the same set of expert strategies that generated the training data (this stands in contrast to Shapira et al., 2023, in which there is a mismatch between training-time and test-time experts). However, unlike previous work on prediction in persuasion games, we aim to evaluate the prediction quality when the training data does not consist of any actual human DMs data, and instead consists of data generated by LLM-based players. Similarly to Shapira et al. (2023), we evaluate a prediction model with respect to the per-DM per-expert average accuracy.

3 Data Collection

Human dataset We use the human-bot interaction dataset of Shapira et al. (2023), which was collected via a mobile application (published on both Apple's App Store and Google Play), in which human DMs interact with 6 experts in a multi-stage game. In each stage, the human DM plays multiple games against the same bot expert, denoted e_i (recall that each game has T rounds, where they set $T = 10$). The i'th stage ends whenever the human player reaches a pre-defined cumulative utility $\bar{v}_i$ in a $T-$round single game.[4] Human data was collected from May 2022 to November 2022. The human dataset contains 71,579 decisions made by 210 distinct human DMs that completed the game, i.e. completed all six stages.[5] Hotel reviews were taken from Booking.com. The hotels and the game parameters were chosen such that each hotel has $R = 7$ reviews, and about half of the hotels are defined to be of high quality (i.e., have an average score of at least $\tau = 8$). The hotel dataset used by Shapira et al. (2023) contained 1068 distinct hotels. Appendix C.2 provides an example of a single hotel review and its corresponding score, taken from Shapira et al. (2023).

LLM datasets To solve the human choice prediction task without using human-generated data in the train set, we created an LLM-generated dataset by replicating the human dataset data collection process, and replacing human DMs with LLMs. To do so, we implemented an identical pipeline with the exact same set of experts, hotels, and game parameters (T, R and τ). We utilized 5 different state-of-the-art LLMs to generate the LLM datasets: Google's Chat-Bison (Google, 2023) and Gemini-1.5 (Gemini Team, 2024), Alibaba's Qwen-2 72B (Yang et al., 2024), and Meta's Llama-3 70B and 8B (Bhatt et al., 2024), all in their chat versions. Similarly to the human choice dataset, each LLM-based player played two games against each expert (unless the LLM player completed the stage in the first game). We repeat this process over and over to create many LLM players. The prompt given to the LLMs is similar to the instructions and messages presented to the human players in the mobile application developed by Shapira et al. (2023). Appendix C contains this prompt, as well as a conversation example.

4 Models and Baseline

The primary goal of this work is to demonstrate the effectiveness of training a human choice prediction model using LLM-generated data. To do so, we compare the performance of prediction models trained on actual human-choice data with the same model trained on LLM-generated data, across various prediction model architectures. In addition, we compare to a baseline method in which data is collected using agents that rely solely on the sentiment analysis abilities of an LLM on the review text rather than a combination of both linguistic and behavioral understanding. All prediction models are evaluated on the choice data corresponding to 100 human players in the dataset, randomly selected for 50 different test sets. Prediction models that used human training data were trained with $K \in [32, 64, 110]$ human players which do not belong to the evaluated test set.

We consider four types of prediction models: LSTM (Hochreiter & Schmidhuber, 1997) and Mamba (Gu & Dao, 2024) as sequential models, transformer (Vaswani et al., 2023) as an attention-based model, and XGBoost (Chen & Guestrin, 2016) as a history independent model.[6] All predictors use the same representation of the choice data and are being trained on data generated by the three different paradigms (human, LLM, and baseline). That is, throughout the experiments, we fix the prediction model architecture and only modify the data it is trained on. This enables the evaluation of the generated data quality in terms of enhancing the human choice prediction.

[4] For our prediction task, we consider only the first two games each human player played against each expert (or a single game if the player completed the stage in a single game). We restrict our analysis to the initial games due to a noticeable decline in participants' response times in subsequent games, suggesting reduced attention and non-strategic behavior. This approach is widely accepted in social sciences and experimental economics, see e.g. Rubinstein (2013).

[5] Players were incentivized to complete all six stages in various ways, including participation in lotteries and receiving academic credit for completing the game.

[6] We evaluate our methods with LSTM, Mamba and a 4-headed transformer, all with a learning rate of $4 \cdot 10^{-4}$, 64 hidden dimensions, and two layers. For XGBoost, we used 300 estimators with a max depth of 3.

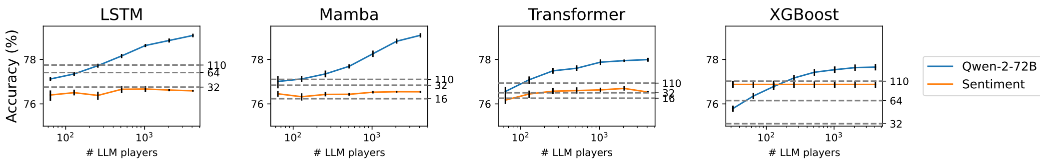


Figure 2: Accuracy obtained by prediction models trained on different data sources. Grey lines represent the accuracy obtained by a model trained on human data with a different number of players. Results are shown for LSTM, transformer, Mamba and XGBoost. Notably, for all prediction models training on LLM-generated data outperforms training on actual human choice data when the number of LLM players is large enough. In addition, the LLM-based training paradigm outperforms the sentiment analysis baseline, implying that allowing simulated players to determine behavior (and not just to interpret the textual signal) yields a better predictor.

Sentiment baseline The baseline method, instead of considering both the interaction history and the text, considers only the sentiment of the text. To implement this baseline, we first asked an LLM to predict the scores τ of all reviews in the dataset, and for each such review, we extracted the score distribution induced by the LLM using its logits (see Appendix D for more details). Then, we use this review-score joint distribution to simulate a choice dataset for the human choice prediction task. For every given review, we sample a score and apply the following decision rule: going to the hotel if and only if the sampled score is above the threshold τ that defines the hotel's quality.

Importantly, unlike the LLM-based player, the baseline DM is forced to act solely based on the current linguistic signal (i.e., the agent's decision rule is history-independent). For instance, the baseline player cannot use punishment strategies in the form of ”trust the expert until she turns out to have sent a great recommendation for a bad hotel”. In contrast, both human and LLM players may condition their current actions not only on the current message but also on the interaction history, and potentially learn such complex strategies that involve patterns of cooperation and punishment.

For the baseline method, we use the same LLM that generated the data which achieved the best performance when looked both at the text and at the history. This baseline allows us to examine the advantage of a model trained on LLM data created with both linguistic and behavioral knowledge, compared to a model trained on an equal amount of data generated solely based on linguistic knowledge, thereby measuring the improvement in prediction that comes as a result of the LLM's economic understanding.

5 Results

Figure 2 presents the accuracy obtained by the Qwen-2-72B LLM, which achieved the best performance for the prediction task.[7] It shows the results for varying training sizes, along with the sentiment analysis baseline. The x-axis represents the number of different players used to create the expert-LLM players interaction dataset (logarithmic scale), while the y-axis displays the accuracy averaged over 50 runs. The results are presented with bootstrap confidence intervals, at a confidence level of 95%, and are distinguished by the players in the train and test sets and the initial weights of the neural networks. The grey horizontal lines indicate the accuracy achieved by a model trained with human players. The number of players used during training is displayed to the right of each line.

Notably, for all types of prediction models, LLM-generated data outperforms human data in the human choice task for a large enough sample size. Importantly, in most real-life applications, obtaining human data of sufficient sample size is significantly more complex and expensive compared to LLM data generation. Results for all evaluated LLMs are provided in Appendix E. Aside from Qwen-2-72B, most LLMs (except Llama-3-8B) also generated data that allowed the training of a predictive model with quality surpassing that of a model trained with data from 16 humans.

Next, we compare the performance of training with LLM-generated data to training with the sentiment baseline method. We recall that data generated by the baseline method is history-independent, in the

[7]Note the atomic unit of evaluation is a *decision* nor a *player*. We report the number of players since each player makes a different number of decisions.

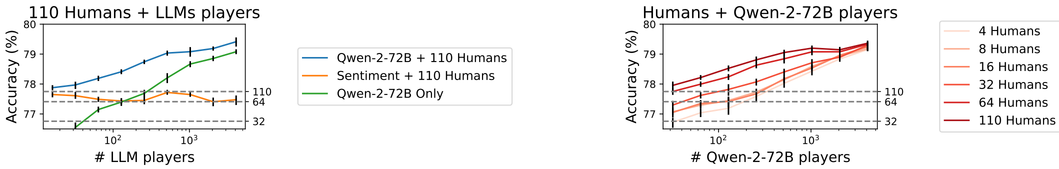


Figure 3: **Left:** Accuracy of models trained with data containing 110 humans and a varying number of LLM players (x-axis). **Right:** Accuracy of a model trained with data containing a varying number of humans (color) and a varying number of players generated by Qwen-2-72B (x-axis). Gray lines: Models trained with data comprised of only human players. Combining human and LLM players outperforms training solely on human players.

sense that decisions of the current round are made based only on the current review score prediction. We note that using LLMs to simulate an end-to-end interaction (and, in particular, allowing such history-depending behavior) leads to better predictions of human choice behavior compared to this baseline.[8] This implies that simulation interaction based only on linguistic aspects is insufficient for human choice prediction. In contrast, we hypothesize that our LLM-generated data encodes not only a linguistic understanding of the textual reviews but also an economic understanding of the interaction: LLM players may condition their current choice not only on the linguistic signal but also on the outcome of previous interactions, leading to a variety of behavioral patterns. Evidently, these patterns increase the predictive power compared to the baseline method. Since LSTM outperforms the other prediction models for each training data configuration, and Qwen-2-72B is the best data generator for our task, the results of all following experiments are shown only for the LSTM prediction model trained with the dataset created using the Qwen-2-72B LLM.

Combination of LLM + human data We now investigate how mixing LLM and human data impacts performance. Figure 3 (left) shows the performance of different prediction models whose training set consisted of 110 human players, supplemented by a varying number of LLM players. It can be seen that enriching the human data with synthetic data significantly improved performance. Figure 3 (right) shows the performance of a model trained with a varying number of human players supplemented by players generated by Qwen-2-72B, the best-performing LLM. It can be observed that a model trained on a mixture of LLM-based and human players outperforms a model trained on human data only. Hence, even if human data has already been collected, enriching it with LLM-based players will yield greater benefits than collecting additional data from humans.

6 Conclusions

This paper illustrates the potential of using LLM-generated data for human choice prediction in language-based persuasion games. We built upon the language-based persuasion game and showed that training a choice prediction model on a dataset containing no human choice data at all can even outperform the same model trained on an actual human-generated dataset, given enough generated data points. This observation has major implications for understanding synthetic data potential in the context of enhancing human choice prediction. We compared the results to a naive sentiment-analysis baseline, in which decisions of synthetic players are made solely based on the current hotel review, without considering the entire history of interaction between the DM and the expert. The fact that our method significantly outperforms the baseline highlights the importance of generating synthetic decision data in a way that captures strategic behavior. We then demonstrated the effectiveness of combining LLM-generated data with existing human data to further enhance predictive power.

While the findings of this paper are specific to our experimental context, they offer a novel approach to studying and predicting human behavior. Future research may focus on exploring the predictive power of LLM-generated data beyond the context of language-based persuasion games, as well as characterizing the boundaries and limitations of this approach, e.g. by utilizing explainability methods to better understand the difference between models trained on different data sources.

[8] The fact that this naive language-only baseline obtains decent results in terms of prediction accuracy is consistent with previous literature on sentiment analysis in economics environments, e.g. (Pagolu et al., 2016; Venkit et al., 2023).

Ethical Statement

This work may also have several ethical implications. In behavioral economics studies, ethical issues concern keeping participants' privacy, and this work suggests ways to make that process easier. From this perspective, our suggested approach offers a solution for such ethical considerations involved with experimental economics. On the other hand, this very same approach, which serves effective human choice prediction, has the potential to be utilized maliciously. The potential of using LLMs to enhance human choice prediction, demonstrated in this work, calls for clear ethical guidelines to safeguard against its potential for harm, emphasizing the importance of responsible use.

Acknowledgements

R. Reichart and E. Shapira have been partially supported by a VATAT grant on data science. The work of O. Madmon, E. Shapira and M. Tennenholtz is funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement n° 740435). We would like to thank Itamar Reinman for providing valuable comments on an earlier version of this paper.

References

Alessandro Acquisti, Curtis Taylor, and Liad Wagman. The economics of privacy. *Journal of economic Literature*, 54(2):442–492, 2016.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pp. 337–371. PMLR, 2023.

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *arXiv preprint arXiv:2305.16867*, 2023.

Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. Do llms exhibit human-like reasoning? evaluating theory of mind in llms for open-ended responses, 2024. URL https://arxiv.org/abs/2406.05659.

Reut Apel, Ido Erev, Roi Reichart, and Moshe Tennenholtz. Predicting decisions in language based persuasion games. *Journal of Artificial Intelligence Research*, 73:1025–1091, 2022.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Itai Arieli, Omer Madmon, and Moshe Tennenholtz. Reputation-based persuasion platforms. *Games and Economic Behavior*, 147:128–147, 2024. ISSN 0899-8256. doi: https://doi.org/10.1016/j.geb.2024.07.002. URL https://www.sciencedirect.com/science/article/pii/S0899825624000964.

Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas, and Suzan Verberne. Generating synthetic documents for cross-encoder re-rankers: A comparative study of chatgpt and human experts. *arXiv preprint arXiv:2305.02320*, 2023.

Robert J Aumann and Sergiu Hart. Long cheap talk. *Econometrica*, 71(6):1619–1660, 2003.

Robert J Aumann, Michael Maschler, and Richard E Stearns. *Repeated games with incomplete information*. MIT press, 1995.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks, 2024. URL https://arxiv.org/abs/2406.18403.

James Best and Daniel Quigley. Persuasion for the long run. *Available at SSRN 2908115*, 2022.

Manish Bhatt, Sahana Chennabasappa, Yue Li, Cyrus Nikolaidis, Daniel Song, Shengye Wan, Faizan Ahmad, Cornelius Aschermann, Yaohui Chen, Dhaval Kapil, David Molnar, Spencer Whitman, and Joshua Saxe. Cyberseceval 2: A wide-ranging cybersecurity evaluation suite for large language models, 2024. URL `https://arxiv.org/abs/2404.13161`.

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. Deep blue. *Artificial intelligence*, 134 (1-2):57–83, 2002.

Gary Charness and Matthew Rabin. Understanding social preferences with simple tests. *The quarterly journal of economics*, 117(3):817–869, 2002.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016. URL `http://arxiv.org/abs/1603.02754`.

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, pp. 354–372. PMLR, 2021.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. Using large language models in psychology. *Nature Reviews Psychology*, pp. 1–14, 2023.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 2023.

Joseph Farrell and Matthew Rabin. Cheap talk. *Journal of Economic perspectives*, 10(3):103–118, 1996.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*, 2023.

Google Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, June 2024. URL `http://arxiv.org/abs/2403.05530`. arXiv:2403.05530 [cs].

Alon Goldstein, Miriam Havin, Roi Reichart, and Ariel Goldstein. Decoding stumpers: Large language models vs. human problem-solvers. *arXiv preprint arXiv:2310.16411*, 2023.

Google. Palm 2 technical report, 2023.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL `https://arxiv.org/abs/2312.00752`.

Shangmin Guo, Haoran Bu, Haochuan Wang, Yi Ren, Dianbo Sui, Yuming Shang, and Siting Lu. Economics arena for large language models. *arXiv preprint arXiv:2401.01735*, 2024.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Zak Hussain, Marcel Binz, Rui Mata, and Dirk U Wulff. A tutorial on open-source large language models for behavioral science. 2023.

Daniel Kahneman, Jack L Knetsch, and Richard Thaler. Fairness as a constraint on profit seeking: Entitlements in the market. *The American economic review*, pp. 728–741, 1986.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6): 2590–2615, 2011.

Jeong-Yoo Kim. Cheap talk and reputation in repeated pretrial negotiation. *The RAND Journal of Economics*, pp. 787–802, 1996.

Aman Kishore, Tae Eun Choe, Junghyun Kwon, Minwoo Park, Pengfei Hao, and Akshita Mittel. Synthetic data generation using imitation training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3078–3086, 2021.

Tuan Anh Le, Atilim Giineş Baydin, Robert Zinkov, and Frank Wood. Using synthetic data to train neural networks is model-based reasoning. In *2017 international joint conference on neural networks (IJCNN)*, pp. 3514–3521. IEEE, 2017.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477, 2022.

Samarth Mishra, Rameswar Panda, Cheng Perng Phoo, Chun-Fu Richard Chen, Leonid Karlinsky, Kate Saenko, Venkatesh Saligrama, and Rogerio S Feris. Task2sim: Towards effective pre-training and transfer from synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9194–9204, 2022.

OpenAI. Gpt-4 technical report, 2023.

Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. Sentiment analysis of twitter data for predicting stock market movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pp. 1345–1350, 2016. doi: 10.1109/SCOPES.2016.7955659.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Towards making the most of chatgpt for machine translation. arxiv. *Preprint posted online March*, 24, 2023.

Ori Plonsky, Ido Erev, Tamir Hazan, and Moshe Tennenholtz. Psychological forest: Predicting human behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Ori Plonsky, Reut Apel, Eyal Ert, Moshe Tennenholtz, David Bourgin, Joshua C Peterson, Daniel Reichman, Thomas L Griffiths, Stuart J Russell, Evan C Carter, et al. Predicting human decisions with behavioral theories and machine learning. *arXiv preprint arXiv:1904.06866*, 2019.

Raul Puri, Ryan Spring, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Training question answering models from synthetic data. *arXiv preprint arXiv:2002.09599*, 2020.

Maya Raifer, Guy Rotman, Reut Apel, Moshe Tennenholtz, and Roi Reichart. Designing an automatic agent for repeated language–based persuasion games. *Transactions of the Association for Computational Linguistics*, 10:307–324, 2022. doi: 10.1162/tacl_a_00462. URL `https://aclanthology.org/2022.tacl-1.18`.

Ariel Rosenfeld and Sarit Kraus. Predicting human decision-making. In *Predicting Human Decision-Making: From Prediction to Action*, pp. 21–59. Springer, 2018.

Ariel Rubinstein. Response time and decision making: An experimental study. *Judgment and Decision Making*, 8(5):540–551, 2013. doi: 10.1017/S1930297500003648.

William Samuelson and Richard Zeckhauser. Status quo bias in decision making. *Journal of risk and uncertainty*, 1:7–59, 1988.

Eilam Shapira, Reut Apel, Moshe Tennenholtz, and Roi Reichart. Human choice prediction in non-cooperative games: Simulation-based off-policy evaluation. *arXiv preprint arXiv:2305.10361*, 2023.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Karthik Sreedhar and Lydia Chilton. Simulating human strategic behavior: Comparing single and multi-agent llms, 2024. URL `https://arxiv.org/abs/2402.08189`.

C Stevenson, I Smal, M Baas, R Grasman, and H van der Maas. Putting gpt-3's creativity to the (alternative uses) test. arxiv, 2022.

Teo Susnjak. Applying bert and chatgpt for sentiment analysis of lyme disease in scientific literature. *arXiv preprint arXiv:2302.06474*, 2023.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. Systematic biases in llm simulations of debates, 2024. URL `https://arxiv.org/abs/2402.04049`.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences. *arXiv preprint arXiv:2309.10621*, 2023.

Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 969–977, 2018.

Dennis Ulmer, Elman Mansimov, Kaixiang Lin, Justin Sun, Xibin Gao, and Yi Zhang. Bootstrapping llm-based task-oriented dialogue agents via self-talk. *arXiv preprint arXiv:2401.05033*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL `https://arxiv.org/abs/1706.03762`.

Pranav Narayanan Venkit, Mukund Srinath, Sanjana Gautam, Saranya Venkatraman, Vipul Gupta, Rebecca J Passonneau, and Shomir Wilson. The sentiment problem: A critical survey towards deconstructing sentiment analysis. *arXiv preprint arXiv:2310.12318*, 2023.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. Is chatgpt a good sentiment analyzer. *A Preliminary Study*, 2023.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey, 2023.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 Technical Report, July 2024. URL `http://arxiv.org/abs/2407.10671`. arXiv:2407.10671 [cs].

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. Zerogen: Efficient zero-shot learning via dataset generation. *arXiv preprint arXiv:2202.07922*, 2022.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. Extractive summarization via chatgpt for faithful summary generation. *arXiv preprint arXiv:2304.04193*, 2023.

A Related Work

LLMs and human behavior Recent studies examine the extent to which LLMs can mimic human behavior. Previous works demonstrate the abilities of LLMs to solve stumpers (Goldstein et al., 2023), take creativity tests (Stevenson et al., 2022), and simulate human samples from sub-populations in social science research (Argyle et al., 2023). Another line of research focused on exploring whether and when LLMs can replace human participants in psychological science (Aher et al., 2023; Hussain et al., 2023; Dillion et al., 2023; Demszky et al., 2023; Taubenfeld et al., 2024). Amirizaniani et al. (2024) evaluate the Theory of Mind capabilities of LLMs through open-ended responses. Closer to our work, Horton (2023) evaluated LLMs in experiments motivated by classical behavioral economics experiments of Kahneman et al. (1986), Samuelson & Zeckhauser (1988), and Charness & Rabin (2002). Sreedhar & Chilton (2024) used LLMs to simulate human strategic behavior in the classical ultimatum game. These growing lines of research inspired us to ask whether the ability of LLMs to behave like humans implies they can function as training data generators for human choice prediction. Importantly, we demonstrate this novel approach in a well-motivated economic setup, where, in contrast to previous work, natural language plays a crucial role in the interaction.

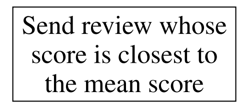
LLMs as data generators There is a vast literature on training ML models using synthetic data (Le et al., 2017; Tremblay et al., 2018; Puri et al., 2020; Kishore et al., 2021; Mishra et al., 2022). The recent rise of LLMs offers promising advancements in this area by providing scalable and high-quality methods for generating synthetic data. LLMs have been successfully used as data generators for tabular data (Borisov et al., 2022), medical dialogue summarization (Chintagunta et al., 2021), text classification (Meng et al., 2022; Ye et al., 2022), and more. LLMs have also been used to annotate data (Chiang & Lee, 2023; Thomas et al., 2023), improve document ranking (Askari et al., 2023), and replace human judges in NLP tasks (Bavaresco et al., 2024). To the best of our knowledge, we are the first to demonstrate the effectiveness of LLMs as training data generators for human choice prediction.

LLMs as rational agents LLMs have also emerged as potential rational agents in economic setups. This new paradigm marks a significant shift from past approaches where algorithms devoid of language capabilities were utilized for solving complex games such as Chess and Go (Silver et al., 2017; Campbell et al., 2002). LLMs offer a novel perspective by acting as rational agents, as demonstrated in previous work: Guo et al. (2024) show that LLMs may converge to Nash-equilibrium strategies, and Akata et al. (2023) demonstrate how LLMs tend to cooperate in repeated games. There is a vast recent literature on concrete applications of LLMs as rational agents, including negotiation (Fu et al., 2023) and task-oriented dialogue handling (Ulmer et al., 2024). Here we leverage these capabilities of LLMs to simulate strategic behavior in a fundamental economic setup in which language is coupled with strategic behavior. We then use this simulated data to predict human behavior.

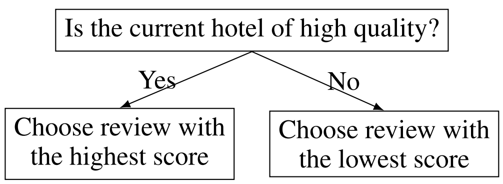
B Expert Strategies

Our paper is concerned with solving the task of predicting human decisions in a language-based persuasion game, as studied by Apel et al. (2022). Shapira et al. (2023) collected and published a dataset of human-bot interactions, in which human players interact with six expert bots. We use these interactions to define the prediction task, hence we use the same expert strategies to collect our LLM dataset. We devote this section to discussing the six strategies introduced by Shapira et al. (2023) and claiming that this set is somewhat representative of the entire strategy space of the expert in the persuasion game. Figure 4 provides the binary tree representation of all six expert strategies.

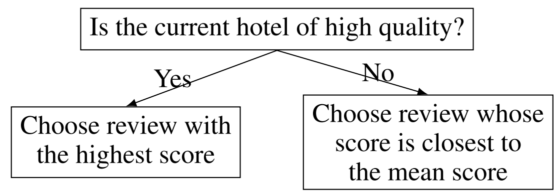
(A) `Greedy` strategy



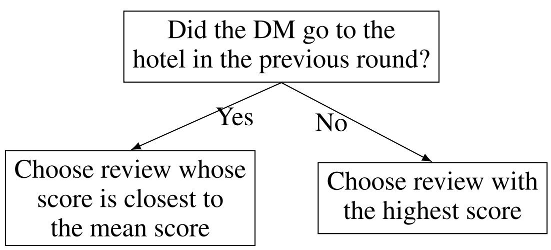
(B) `Average` strategy



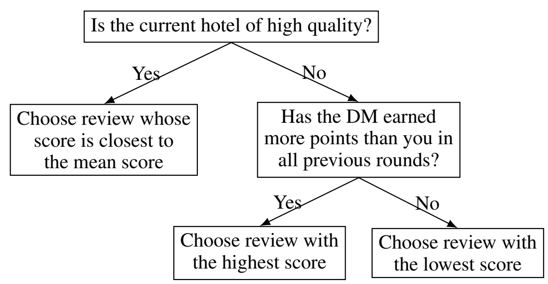
(C) `Honest` strategy



(D) `Ambiguous` strategy



(E) `Choice-based Adaptive` strategy



(F) `Points-based Adaptive` strategy

Figure 4: Expert strategies, represented as binary trees.

Strategies interpretation Importantly, we argue that each of the six strategies has a clear and intuitive behavioral interpretation. The `Greedy`, briefly discussed in the introduction, is the strategy according to which the expert always reveals the most positive review. The `Average` strategy is a more careful strategy that always displays the (closest to the) average review, aiming to build trust by revealing more representative information to the DM.

Another strategy that aims to build trust with the DM is the `Honest` strategy, also discussed briefly in the introduction. In this strategy, the expert reveals the most positive review when the hotel is of high quality, but ”warns” the DM when the hotel is of low quality by revealing the worst possible review. A similar, yet more sophisticated and manipulative strategy is the `Ambiguous` strategy. In this strategy, the expert also reveals the most positive review for good hotels, but when the hotel is bad she ambiguously selects a not-too-negative review. This principle of revealing the state when it is good and ”bluffing” when it is bad is fundamental in the economic literature, and turns out to be an optimal sender strategy in economic setups that can be modeled as persuasion games, such as product adoption games (e.g., Arieli et al., 2024).

The last two strategies, `Choice-based Adaptive` and `Points-based Adaptive`, are slightly more sophisticated as they aim to adaptively control the behavior of the DM. The `Choice-based Adaptive` completely ignores the true quality of the hotel and only conditions the review selection rule on whether the DM opted-in in the previous round or not. This can be seen as trying to push the hotel more aggressively after failing to do so in the previous round. `Points-based Adaptive` follows a similar high-level principle, with two major differences: it first takes into account the actual quality of the hotel, and it selects the review to present based on the number of points gathered by the DM, and not by the selected action.

Strategies classification We argue that these can be classified into three different strategy classes:

- **Naive strategies.** These are strategies in which the expert's action is independent of both the actual quality of the hotel and the interaction history.
- **Stationary strategies.** These are strategies in which the expert's action depends solely on the actual quality of the hotel (but it is independent of the interaction history).
- **Adaptive strategies.** These are strategies in which the expert's action depends on the interaction history.

It is clear that `Greedy` and `Average` are naive, `Honest` and `Ambiguous` are stationary, and `Choice-based Adaptive` and `Points-based Adaptive` are adaptive. Trivially, every possible strategy is either naive, stationary, or adaptive. While the entire strategy space is infinite, we argue that the consideration of two strategies of each class within the experimental setup is somewhat representative.

C Additional Dataset Information

C.1 Dataset Statistics and Persona Diversification

To diversify the LLM-generated dataset, we have associated each LLM player with a *persona*. Each persona type specifies a typical behavior the LLM is instructed to follow. This is implemented by simply concatenating a sentence to the beginning of the initial prompt of the LLM player, that specifies the required behavior. For example, for the `optimistic` persona type, the initial prompt contains the game instructions (as they were written for the human players), followed by the following instruction: *"behave like an optimistic person"*.

Table 1 provides all descriptions and prompts for all 8 persona types we used. The first two persona types are the `optimistic` and `pessimistic` personas. The other six persona types were selected based on textual features extracted from Booking.com reviews, as described by Apel et al. (2022). These reviews were represented using binary features that indicated whether various hotel aspects were discussed positively or negatively. We used these features to create different personas, each focusing on a specific aspect of the hotel.

Persona Type	**Prompt Prefix**
Optimistic	"Behave like an optimistic person."
Pessimistic	"Behave like a pessimistic person."
Price	"Behave like a person to whom the hotel's price is important."
Facilities	"Behave like a person who values the facilities offered by the hotel."
Room	"Behave like a person who cares about the quality of the room in the hotel."
Location	"Behave like a person for whom the location of the hotel is important."
Staff	"Behave like a person who cares about the treatment they will receive from the hotel staff."
Sanitary	"Behave like a person to whom the sanitary conditions of the hotel are important."

Table 1: Persona types and the initial prompts of the LLM players.

For each language model, we generated decision data using all different persona types (personas, hereafter). For each specific persona, we generated a decisions dataset using a varying number of LLM-based players, as specified in Table 2.

Persona	**Qwen-2-72B**	**Llama-3-8B**	**Llama-3-70B**	**gemini-1.5-flash**	**Chat-Bison**
Optimistic	519	512	208	129	514
Pessimistic	528	512	208	144	512
Price	538	512	192	128	514
Facilities	525	517	208	128	514
Room	513	512	145	138	514
Location	512	516	193	132	514
Staff	512	528	208	128	514
Sanitary	530	512	208	129	514
All	4177	4121	1570	1056	4110

Table 2: Number of players created by each LLMs.

Persona diversification for sample size reduction While using similar techniques has already been shown to encourage LLM agents to take a variety of social behaviors (Park et al., 2023), in our context, we observed that persona diversification contributes to reducing the sample size required to obtain accurate predictions. Figure 5 shows the number of players required to achieve certain levels of accuracy in two different settings (here the LLM used to simulate DMs is Chat-Bison, and the prediction model is LSTM):

1. **Without persona diversification.** LLM players were not assigned any persona type, i.e., no persona-associated prefix was added to their prompts.

2. **With persona diversification.** Any LLM player was randomly assigned one of the 8 persona types defined above, and the corresponding prefix was added to its prompt.

Notably, the persona diversification technique indeed improves sample complexity. Moreover, the gap between the two settings increases as the desired level of accuracy increases. After observing the phenomenon in Chat-Bison, which was the first LLM we considered, we decided to continue the data collection process using persona diversification, in order to reduce time and costs.

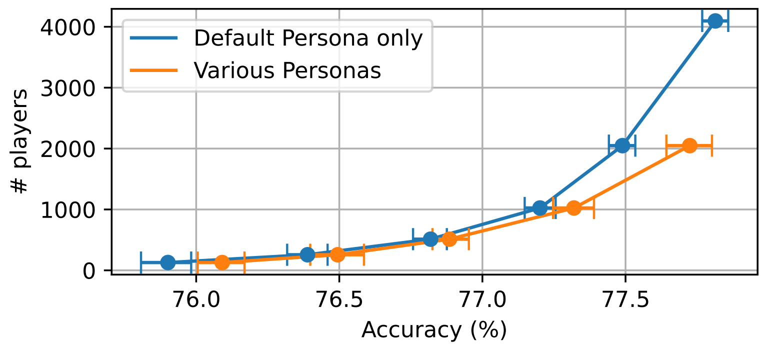


Figure 5: Number of LLM-generated players (using Chat-Bison) required for achieving different levels of accuracy (with the LSTM predictor), with persona diversification (various personas) and without persona diversification ('default' persona only). Interestingly, note that the higher the desired accuracy, the larger the gap between the required sample sizes of the two methods.

C.2 Booking.com Review Example (taken from Shapira et al., 2023)

Figure 6 provides an example review from the hotel reviews dataset.

Positive: Great location, walking distance from the town centre and Menin Gate. Nice roof terrace and garden. Quiet at night. Comfortable bed. Excellent breakfast

Negative: The room was very small, but everything we needed was provided.

Score: 9.6 / 10

Figure 6: A sample review from the hotel reviews dataset.

C.3 Prompts and Example Conversation

This appendix explains how we collected the data through interaction with an LLM that simulated a DM. In addition, we introduce the beginning of an interaction and its first two rounds. The colors represent the message sender: messages from nature appear in purple, messages from the Sender appear in blue, and messages from the LLM appear in orange. Every round, the LLM-Chat agent receives a message containing all the purple and blue parts that have been sent since its last message.

> Behave like a person to whom the hotel's price is important.
> Let's play a game!
> ###
> Introduction:
> Are you the vacation planner at your house? Think you always know how to choose the best hotel? Start to plan your 10-days trip with our travel agents. Just remember

Figure 7: The data collection pipeline of a single LLM-based player. The player is initialized with a prompt that contains the persona prefix followed by the introduction message. Then, nature (purple) manages the conversation between the agent (blue) and the LLM-based player (orange). Practically, the LLM-Chat agent receives a concatenation of all purple and blue parts sent since its last message as a single message.

- they don't always want the best for you, and might have their own strategy to make you book the hotel they try to promote! Travel or Trouble is a strategy game in which you will try to outsmart our traveling agents and plan the perfect vacation for you. Each game consists of 10 rounds, in each round, one of our traveling agents will introduce you to a review for a new hotel they think might suit you, and you will have to choose: either book the hotel or stay home.
###
The game:
You will play as the traveler and encounter several travel agents in the game. In each round, the agent will provide you with a message about the hotel, and you will decide whether to go to the hotel or stay at home. Your goal is to go to the good hotels and avoid the bad ones.
Meet your new travel agent: David!
You'll be playing the next game with David as your travel agent.
You need to earn 10 points to win the game.
###
Round: 1 | You have 0 points
David's review about the hotel:
~

Positive: Location
Negative: Bad hotel. Bad service! First problems with the doors keycard, had to go up 2 times before receptionist believed us. Then no light in the bathroom. Unable to fix..so to toilet in the dark. Compensation was free breakfast. Next morning, no water on top floor, clearly a pressure problem, new receptionist had no cloe, "new problem, never had complaints". Then he wanted us to pay for breakfast :(I'm not a complainer but this was bad! Do not get fooled by rating..hotel without a soul, and nobody who takes responability for it.
~

Choose your action: [Go | Don't Go]
Go
Round results:
This hotel is bad, You should have skipped it.
This round, you earn no points.
###
Round: 2 | You have 0 points
David's review about the hotel:
~

Positive: Big and spacious. This apartment was EXCEPTIONALLY CLEAN and a great value in Brussels. If given the chance to stay here again I would not hesitate to do so. The staff was also very nice.
Negative: I have nothing negative to say about our experience.

~

Choose your action: [Go | Don't Go]
Go

D Review-Score Distribution Extraction for the Linguistic Baseline Method

This appendix explains how we extract the score distribution induced by the LLM for a given review. We use this distribution in the baseline method, as explained in section 4. We asked the language model to transform a review into a numerical score, and extracted the underlying score distribution from the distribution the model assigns to the different numerical tokens.

For instance, if upon observing review r the model assigns a probability of 0.4 to the first token of the output to be 8, then $P(8 \leq s < 9|r) = 0.4$.

The following text describes the format prompt we used for extracting the probabilities. The red parts represent the review itself.

> Rank the value of the hotel as presented by the review, from 1 to 100, with 80 being the minimum score for a hotel you would like to stay in.
> Positive: Big and spacious. This apartment was EXCEPTIONALLY CLEAN and a great value in Brussels. If given the chance to stay here again I would not hesitate to do so. The staff was also very nice.
> Negative: I have nothing negative to say about our experience.
> Answer only with your value!

E FULL RESULTS: ACCURACY OF PREDICTION MODELS TRAINED ON DIFFERENT DATA SOURCES

Table 3 shows the accuracy obtained by prediction models trained on different data sources.

Training Size	**64**	**128**	**256**	**512**	**1024**	**2048**	**4096**
Qwen-2-72B	77.12	77.35	77.72	**78.16**	**78.63**	**78.85**	**79.08**
gemini-1.5-flash	76.13	76.30	76.67	77.16	77.34	-	-
Chat-Bison	73.65	74.12	74.72	75.34	75.75	75.93	76.04
Llama-3-70B	75.84	75.71	75.53	75.42	75.74	-	-
Llama-3-8B	72.04	71.45	71.72	71.38	71.54	71.70	71.93
Sentiment	76.40	76.51	76.38	76.66	76.67	76.62	76.59

Table 3: Accuracy obtained by LLM datasets for varying training sizes, along with the Sentiment baseline, for different numbers of players in the training set. Configurations that provide better accuracy than a model trained with data of 16 humans are underlined. Configurations that provide better accuracy than a model trained with data of 110 humans are **both underlined and bolded**. Some of the values here are missing due to the high cost of generating LLM players.

F Predicting Against a Specific Strategy

Our experiments in the main paper show that our LLM-based data generation paradigm obtains high prediction accuracy, and even outperforms models trained on human data whenever sufficient synthetic data points are available. This appendix provides a more careful analysis of the effectiveness of this approach, by evaluating accuracy with respect to each *individual* expert strategy separately.

Strategy evaluation from the expert's perspective We begin with evaluating the different strategies from the expert's perspective: Figure 8 (Left) shows the expert's winning rate for each expert strategy, against human players and the different LLM players. Figure 8 (Right) shows the same metric *conditional on the actual hotel's quality being low.* That is, the latter shows the persuasion power of each expert in cases where the DM should not opt-in.

A first observation is that most LLM players behave similarly to human DMs. The only exception is Chat-Bison players, which seem to be significantly easier to manipulate. Additionally, as one could expect, for all players, the opt-in frequency is significantly higher when we do not condition the hotel as being of low quality. In terms of evaluating the strategies from the expert's perspective, it is notable that `Greedy` is indeed the best strategy for the expert, which is also consistent with Raifer et al. (2022).

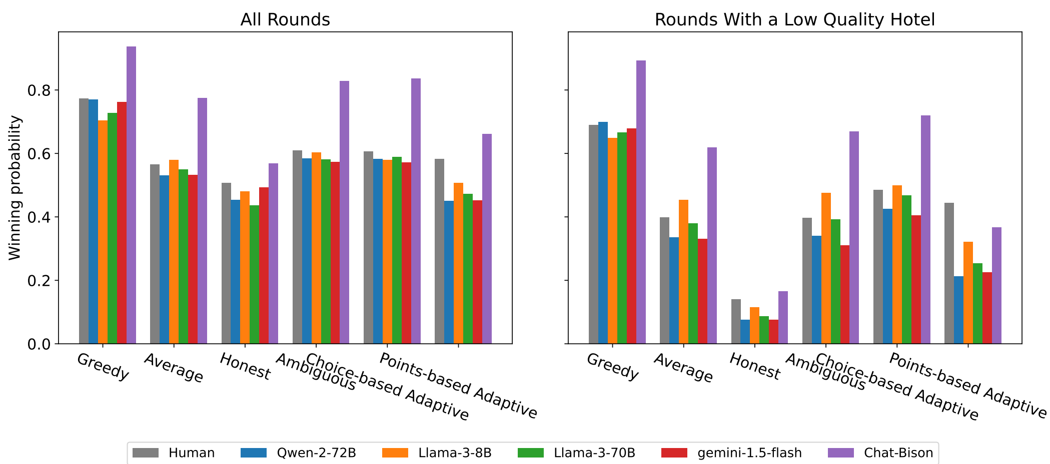


Figure 8: Winning rates of the expert against both LLM-based DMs and human DMs for each expert strategy.

Comparison to human-generated data Figure 9 shows the accuracy with respect to each expert separately for all training paradigms, as in the previous experiment. Each subplot corresponds to an individual expert strategy.[9] The LLM-based approach with Qwen-2-72B generated data always outperforms the models trained on 32 human players. Moreover, in some cases it even outperforms models trained on 110 human players. These results indicate that the LLM-based approach is quite robust: not only that averaging over strategies yields a high accuracy in the prediction task, but also human actions against each strategy separately can be accurately predicted.

Interestingly, one of those cases in which our result is significantly effective, is the case of the `Greedy` strategy, which simply means the expert always sends the review corresponding to the highest score, regardless of the actual quality of the hotel, or the history of interactions with the DM player. This particular expert strategy is important for two reasons: first, it represents a very common and typical behavior of a naive expert who aims to greedily persuade non-sophisticated users towards opting in. Second, it turns out that the `Greedy` expert strategy is very effective in terms of expert utility maximization. Raifer et al. (2022) empirically studied a similar language-based persuasion

[9] One can ask whether training a prediction model only on the individual expert interaction data (instead of using data that includes interactions of players with other experts) yields a higher accuracy. We next show that the answer is negative.

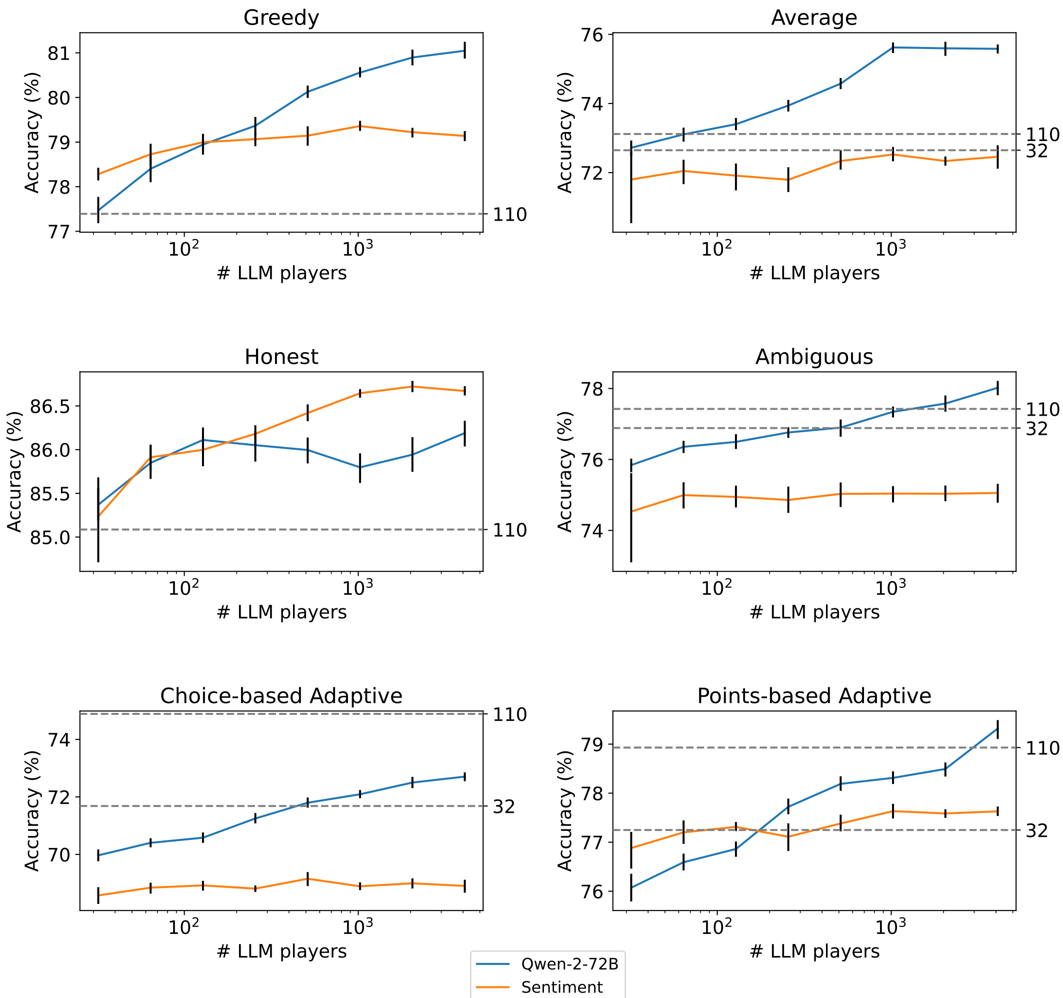


Figure 9: Accuracy on individual expert interactions, obtained by prediction models for the three training configurations: human data, LLM-generated data, and the sentiment baseline. Prediction models trained on LLM-generated data outperform those trained on 32 human players and almost always surpass those trained on 110 human players.

game, and demonstrated the effectiveness of the `Greedy` strategy when used against human DMs. This aligns with our experimental setting, in which the `Greedy` expert strategy is shown to be the best expert strategy among all six strategies considered, against both human and LLM-based DMs (see Figure 8). These arguments suggest that `Greedy` experts are expected to exist in many realistic cases.

Comparison to the sentiment baseline The LLM-based approach indeed outperforms the sentiment baseline in the majority of the cases. The only exception is the `Honest` strategy, on which the naive baseline outperforms both the LLM-based approach and the standard approach of training the predictor on human data. This is exactly the strategy discussed in the introduction and described in Figure 1 (Right), according to which the expert reveals the most positive review when the hotel is of high quality, and the most negative review when it is of low quality. In particular, this is a *stationary* strategy (i.e., independent of the interaction history). Hence, from the DM perspective, once trust in the expert is established, the task of making a decision boils down to a simple sentiment analysis task, which is exactly what the sentiment baseline does. It is therefore reasonable that the baseline method performs well against this particular strategy, while against all other (non-truthful) strategies it is outperformed by LLM-based training.

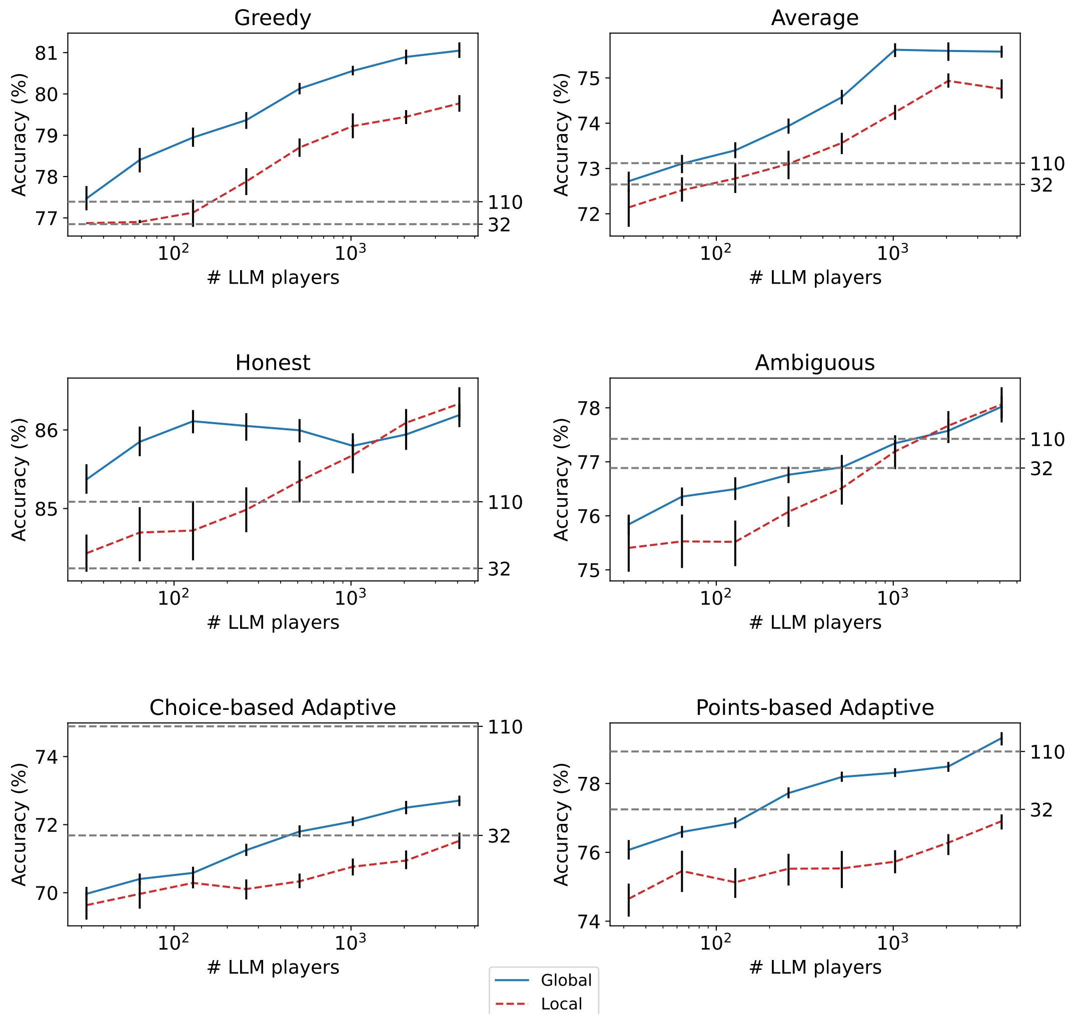


Figure 10: Accuracy on individual expert interactions only, obtained by prediction models trained on LLM-generated data (both local and global). For any individual expert, the global model outperforms the local model.

Global vs. local models We now answer the following natural question: whenever there is only a single expert for which human choice prediction is required, is it better to train a *global model* (i.e., use all experts' interaction data to train the model) or a *local model* (i.e., train only using interaction data corresponding to the individual expert)? This is a question of data quality vs. quantity trade-off: while a global model relies on more observations (many of which are somewhat irrelevant), the local model uses fewer observations, but all of them are collected with respect to the individual expert. Note that throughout this appendix we always trained a global model (namely, included interaction with experts other than the individual expert within the training set), and the results of this experiment will justify this approach. For each individual expert, Figure 10 shows both the results of a local and a global model, for models trained on Qwen-2-72B players. Evidently, the global model outperforms the local model for any possible individual expert. This implies that although the interaction data across expert strategies may be different, it is still beneficial in terms of enhancing the capabilities of an expert-specific prediction model.

G LIMITATIONS

This work suffers from several limitations. First, it focuses on a specific class of games, namely language-based persuasion games. While these games have major importance both in the economics literature and in the NLP community, the fact that the approach of utilizing LLMs for training data generation is demonstrated solely in this setup is indeed restrictive. We view this work as a first attempt to apply the approach in a complex and realistic economic setup, which is also grounded in a particular real world application. Focusing on this particular setup also enables a richer analysis of the results and discussion of the assumptions. For instance, discussing the specific persuasion strategies considered and the evaluation with respect to each strategy separately, is a kind of contribution that is particularly relevant in the context of persuasion games. We hope that this contribution will encourage the community to consider different applications of the proposed approach, as well as further suggest extensions and improvements.

Second, even in the context of persuasion games, the restriction to a specific set of expert strategies is limiting by nature. We highlight that this limitation comes from a practical argument of balancing budget constraints and the need for collecting a large enough sample for each strategy, as well as the need for avoiding the collection of another human choice dataset in addition to the dataset of Shapira et al. (2023) (as doing so may impose non trivial challenges in equalizing the conditions among human participants). These constraints indeed call for taking a critical view of the particular strategies considered, and we believe that enriching the discussion on the extent to which these strategies are relevant and representative can be viewed as a major contribution. Here we suggest two alternative justifications for the particular set of strategies considered: (1) an intuitive behavioral interpretation for each strategy; and (2) a classification of all strategies to types covering the entire strategy space (*naive*, *stationary* and *adaptive*), where each type is indeed represented in our strategies set (see Appendix B). We acknowledge the need for continuously discussing, questioning and criticizing the limitations in the selection of the strategies that define the human choice prediction task.

Lastly, a key question that remains unanswered is *when* and *why* the LLM-based approach is less effective. Namely, can we characterize those specific cases (e.g., specific expert strategies, or even specific decisions) in which models trained on LLM-generated data fall short? One can view our per-expert analysis (Appendix F) as a first step towards this end, yet the complete characterization is left as an interesting future question.

H Compute Information

We utilized a hardware configuration consisting of 8 NVIDIA A100-SXM4-40GB GPUs and 128 CPUs to collect data from Qwen-2-72B and Llama-3-70B. Using this setup, generating a single Qwen-2-72B player took an average of 7.5 minutes, as same as Llama-3-70B player. In total, we generated 4177 Qwen-2-72B players for our experiments, which took approximately 21.75 days to complete. To generate one LLM player using Llama-3-8B, we used a single GPU for 2 minutes. We have used API calls to the Google Cloud platform to generate players with Bison-chat and Gemini-1.5. Each LLM player of this setup costs around 0.5$. We utilized one NVIDIA GeForce GTX 1080 GPU with 8GB of memory to train the prediction models. Training the LSTM model with 4096 LLM players took approximately 16 minutes.