Language Models as Tools for Research Synthesis and Evaluation

Robin Na [1] **Abdullah Almaatouq** [1]

Abstract

Is academic literature building cumulative knowledge that improves the ability to make predictions under interventions? This question touches not only on the internal validity of individual findings but also on their external validity and whether science is a cumulative enterprise that generates collectively more accurate representations of the world. Such synthesis and evaluation face significant challenges, especially in the social and behavioral sciences, due to the system's complexity and less structured nature of research outputs. Motivated by such challenges, we propose a novel method involving large language models (LLMs) and retrieval-augmented generation (RAG) techniques to measure how various sets of academic papers affect the accuracy of predictive models. We elicit LLMs' predictions on the treatment effect of introducing punishment in public goods games (PGG) under 20 varying dimensions in the game design space that show high heterogeneity. We demonstrate the LLM's ability to retrieve academic papers and alter its distribution of predictions in directions that are expected based on the documents' contents. However, we find little evidence that such updates improve the model's predictive accuracy. The framework introduces a method for evaluating the potential contribution and informativeness of scientific literature in prediction tasks, while also introducing a new human behavior dataset of PGG carefully collected from integrative experiment design that can be used as a benchmark for LLM's performance in making predictions about complex human behavior.

1. Introduction

Artificial intelligence is being increasingly integrated into the scientific processes of many fields, especially in its ability to generate hypotheses and synthetic data (Wang et al., 2023). The social and behavioral sciences have not been an exception to this movement. Recent studies explore data-driven methods to use supervised learning to generate novel hypotheses (Ludwig & Mullainathan, 2024), as well as using large language models (LLMs) as both hypothesis generators (Manning et al., 2024; Zhou et al., 2024b) and as surrogates for human participants (Horton, 2023; Argyle et al., 2023; Mei et al., 2024).

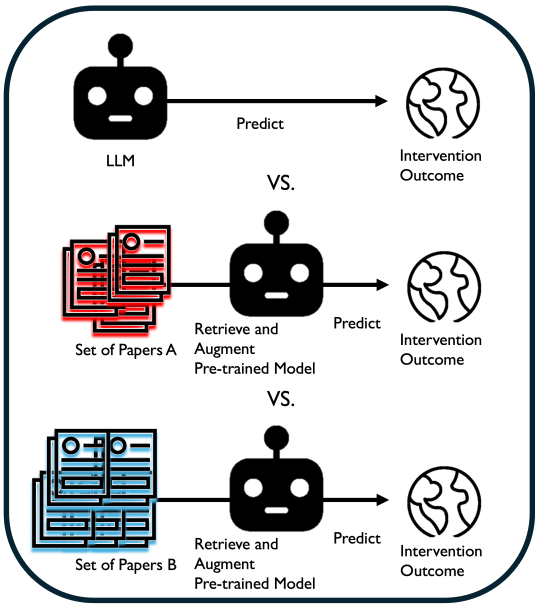

Figure 1. Our experimental design for measuring the treatment effect of different sets of academic papers on LLM's accuracy in making predictions under intervention.

Despite such contributions in scalability and potential novelty in generating new hypotheses in computational social science, evaluating such hypotheses and theories still remains a significant challenge, especially when doing so for the synthetic findings across papers. Such an evaluation of cumulative science is extremely challenging due to both the evaluation part and the synthesis part: we first lack the integrative dataset to serve as a reliable benchmark for evaluating a collection of research findings across different contexts. Also, the less structured nature of academic papers in the field makes it challenging to perform meta-analyses that are scalable both in terms of size and scope.

[1]Massachusetts Institute of Technology.

The challenges in the synthesis part motivate our novel framework of utilizing LLMs and retrieval augmented generation (RAG, Lewis et al. 2020) techniques for measuring how reading a collection of academic papers in certain ways changes the LLM's accuracy in making predictions under heterogeneous interventions. We demonstrate that LLM, at least GPT-4, is a suitable instrument for such operationalization through several levels of manipulation checks upon reading various sources of research outputs. With this validity in hand, we explore which filter of paper selection and which form of paper representation improve the accuracy of retrieval-augmented LLMs.

The unavailability of the integrative benchmark part highlights our contribution in demonstrating concrete applications of the "integrative experiment design" (IED, Almaatouq et al. 2022) as such a benchmark for predicting collective human behavior under highly contingent environments in a well-structured manner. By eliciting LLM's predictions on public goods games (PGG) results collected through IED with 211 unique experiments and 20 dimensions of design space, we systematically evaluate the LLM's ability to make predictions about the effect of punishment on human cooperative behavior by taking these rich features into account and integrating their prior knowledge with academic papers available for retrieval. This not only provides a rich benchmark with well-structured human experimental data for the machine learning community but also demonstrates a concrete application of IED that argues for a wider endeavor for social scientists to generate more of such integrative data in other topics for the sake of contributing to cumulative social science.

The paper is structured as follows: we first outline our motivation by elaborating on the two challenges in evaluating social science on its predictive utility of multiple research outcomes while giving an overview of the replication and generalizability crises that call for a dire need for such an endeavor (Section 2). In Section 3, we describe our PGG dataset from IED and our operationalization using LLM and RAG. We also walk through several manipulation checks to suggest that the changes in LLM's accuracy up on the augmentation is likely due to the content of these academic papers. We show results on how augmenting academic papers contribute to the LLM's predictive accuracy in Section 4, along with further metascientific discussions in Section 5. We discuss the limitation (Section 6) and the impact of our approach that can span across other fields of science as well (Section 7).

2. Challenges in Evaluating Cumulative Knowledge in Science

2.1. The Replication and Generalizability Crises

There are two distinct dimensions that undermine the credibility of science and its practicality: the replication and generalizability crises. While we discuss the symptoms and remedies mostly in behavioral science, the crises are applicable to many other disciplines in the social and natural sciences dealing with high complexity of the system.

The replication crisis refers to the inability to generate data and statistical results that align with the empirical finding from earlier studies despite following the procedures outlined in these studies (Nosek et al., 2022). Given that trying to replicate every paper published is infeasible, many scientific communities have been focusing on predicting whether a paper is likely to be replicated, including supervised-learning methods to scale up the process (Youyou et al., 2023).

Another dimension is the generalizability crisis (Yarkoni, 2022), where the empirical finding does not generalize to other contexts and populations within the paper's (implicit) claim of external validity. Combined with the narrow scope of generalization and most authors' reluctance to explicitly draw the "constraint of generality" (Simons et al., 2017), scientific papers end up being incommensurate with one another which provides less informative signals and even confuses the readers when they were to make predictions based on them.

Our contribution is using machine learning as a tool that zooms into these crises. If a predictive model performs worse than its baseline after being trained on a scientific finding, it suggests that the finding is generating noisy signals either because the study does not replicate (i.e., poor within-distribution performance) or because the study does not generalize to the distribution shift (i.e., poor out-of-distribution performance).

2.2. Benchmark for Evaluating Research Synthesis

Despite the challenges in replication and generalization for each individual paper, one hopeful scenario for science is when the enterprise can filter low-quality papers and integrate different findings from different contexts to build a "map" that informs us which article is relevant in which specific environment (Hu et al.). This calls for benchmarks that can evaluate not one finding from one environment but heterogeneous effects across diverse environments such as high-throughput experimentation (HTE) in natural science.

IED is a novel suggestion of practice in the social and behavioral sciences which shares similarities with HTE. In IED, each experiment lies within a design space that explicitly

specifies as many relevant design choices as possible. This contributes not only to mapping the heterogeneous treatment effect across diverse experimental settings but also to parameterizing each experiment point, enabling adaptive sampling (Bakshy et al., 2018) of subsequent experiments that are most informative to the entire space. This movement produces more reliable benchmarks for integrative science compared to simply collecting experimental data across different papers, as these rarely specify the relevant design choices significant in the IED approach, and each author is not incentivized to sample the experiment that is most informative to the cumulative knowledge of the field.

2.3. Integrating Unstructured Scientific Papers

One motivation for utilizing LLMs relates to the challenges in doing scalable meta-analyses across different sets of papers along with LLMs' promising ability to do so as we show in later sections. While it may be tempting to aggregate the data or statistical analyses across papers, we present several reasons to advocate for making the most out of other language signals in each paper.

First, such statistical aggregations assume that each data generation process in the paper was executed under the same condition and quality of observation. In other words, this assumes that all papers in the set have equal internal validity. Furthermore, it is discovered that the text within each paper can provide highly informative signals for predicting whether a psychology paper will be replicated in the future (Youyou et al., 2023), suggesting a reason not to ignore specific sections within the paper for evaluation.

Furthermore, we need to account for the external and ecological validity. Consider making predictions on the effect of t on y_i under a specific target environment X_i, given a specific paper L and its environment for experiment X_L which specifies to what extent findings in the paper will be externally valid. It is imperative that the predictive model infers how likely X_i will be included in X_L. Since regression tables or data alone do not specify every detail of X_L, we need ways to utilize language models to perform such tasks.

Building on ongoing findings that support LLMs' ability to interact with academic papers critically (Zhou et al., 2024a; Baek et al., 2024), we demonstrate how LLMs can perform as reliable synthesizers in the specific prediction task we set up through the following sections.

3. Material and Methods

3.1. Public Goods Games (PGGs) from Integrative Experiment Design (IED)

We elicit GPT-4's predictions on the effect of punishment in 211 PGG experiments with 20 dimensions of design choices, collected through the IED procedure from a separate project. PGG is a standard multiplayer setting in experimental economics that studies how to foster the players' contribution to the public goods that increase the overall utility of the group, where the default Nash equilibrium is to make zero contribution.

One known prominent feature is enabling costly punishment between players (Fehr & Gächter, 2000; 2002), but some research suggests that punishment contributes to people cooperating even less (Dreber et al., 2008), while many papers suggest that the effect of punishment depends on other factors such as the duration of the game (Gachter et al., 2008), whether players can reward each other (Rand et al., 2009), and whether communication is enabled (Palfrey et al., 2017), just to name a few.

The IED experiment data we use for prediction extracts these features from the literature while adding a few more to create a 20-dimensional design space that can not only place all experiments from prominent PGG papers within but also continually generate new experimental results from game environments yet to be explored from the literature. This serves as a well-structured and reliable experimental dataset for prediction tasks, especially since it is not part of any publication, preventing LLMs' access to the data prior to the prediction.

3.2. Retrieval-Augmented Generation for Prediction

We use GPT-4-turbo and GPT-4o for predicting the effect of introducing punishment mechanisms in 85 different experiment designs. For each analysis, we use the same model and its 'Assistant API' feature for retrieving up to 25 prominent papers in PGG, most of which were cited hundreds to thousands of times and were published in prestigious journals in the general sciences (Nature, Science, PNAS), economics, and psychology.

3.3. Evaluation of Literature Treatment Effect

Given a PGG experiment i, we are ultimately interested in measuring whether an LLM augmented with documents L_j reduces the loss $\mathcal{L}(\cdot,\cdot)$ in making predictions on the efficiency value y_i, after punishment is introduced under a specific PGG design X_i ($dim(X_i) = 20$). Hence, we can measure the effect of L_j by estimating $\mathcal{T}_i^j$ while holding the LLM constant and resetting its memory for every iteration.

$$\mathcal{T}_i^j = \mathcal{L}(y_i^{true}(X_i), \hat{y}_i^{LLM}(X_i)) - \mathcal{L}(y_i^{true}(X_i), \hat{y}_i^{LLM}(X_i; L_j))$$

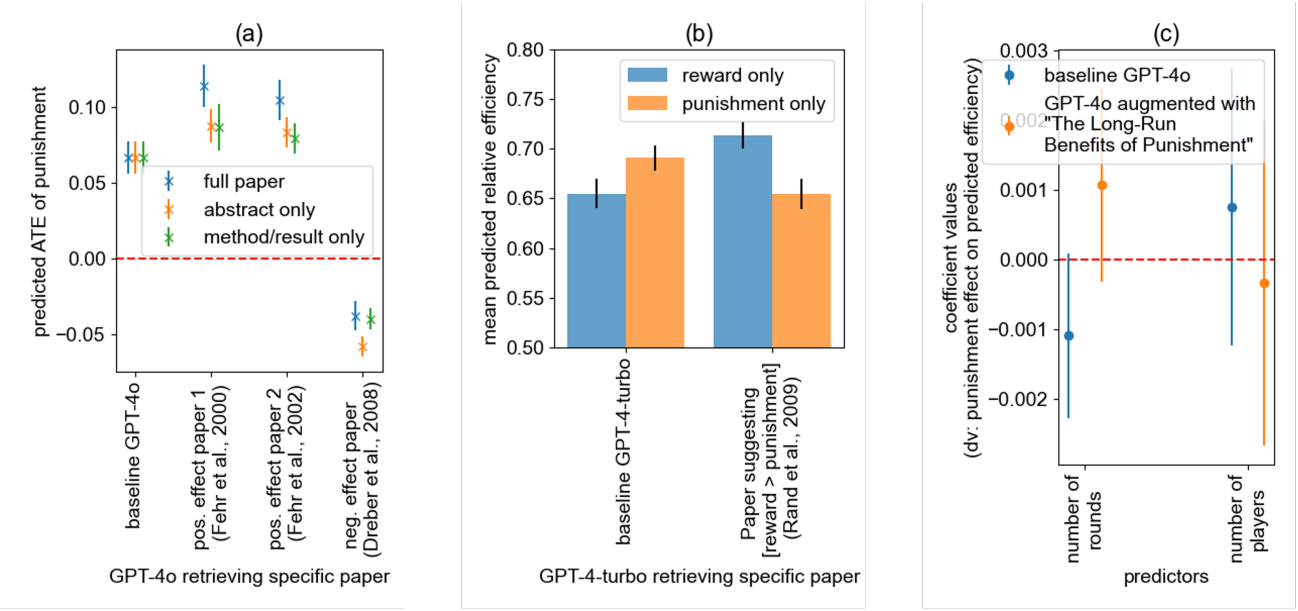

Figure 2. Illustrating the shifting distributions of GPT-4o's predicted outcomes based on its RAG on different papers. Error bars indicate 95% confidence intervals. (a) The model predicts a stronger positive effect of punishment upon reading papers suggesting positive effects of punishment (Fehr & Gächter, 2000; 2002), while switching its prediction to a negative effect upon reading a paper that suggests so (Dreber et al., 2008). (b) The model switches its prior into predicting that the reward will be more effective than punishment upon reading the article suggesting such outcomes (Rand et al., 2009). (c) The model updates its predicted relationship between the number of rounds and efficiency upon reading the paper "the Long-Run Benefits of Punishment (Gachter et al., 2008)." The coefficient for the "number of players" variable remains relatively unchanged since the paper does not highlight its effect.

$\mathcal{T}_i^j > 0$ means that the error of LLM's prediction about experiment i augmented with the literature L_j is less than that of the baseline LLM, indicating that documents L_j improved the prediction. We use squared error loss for every figure that illustrates $\mathcal{T}$.

We also measure $\Delta_i^j = \mathcal{L}(\hat{y}_i^{LLM}(X_i; L_j), \hat{y}_i^{LLM}(X_i))$ to keep track of how much the LLM's response has changed after retrieving from the literature L_j.

3.3.1. Manipulation Check

However, we need to make sure that the LLM updates its responses upon reading L the way they are expected to. We therefore perform multiple levels of manipulation checks. At the lowest level, we confirm that LLMs can correctly retrieve factual information from each paper with near-perfect accuracy across the 25 papers when asked to specify the features in 20 dimensions of the design space.

We then observe whether the distribution of the LLM's response changes in ways that are implied by each paper. For example, does the LLM lean more toward thinking that punishment has positive effects if that is the conclusion drawn by L? We show consistent results from GPT-4-turbo and GPT-4o that it passes the manipulation check via various treatments (Figure 2).

3.3.2. Treatment on Different Parts of Papers

For each paper L_j, we estimate $\mathcal{T}^j$ and Δ^j by performing statistical inference based on 85 PGG experiments (i.e., 85 unique X_i's) from IED. We observe how such changes in distribution and accuracy correlate with various features of each paper.

We also compare the effect of reading the same paper but different versions of it. We compare reading the full paper, reading the title and abstract only, and reading only the design of the experiment and the result that has been summarized by GPT-4o and passed its reading comprehension test.

3.3.3. Treatment on the Collection of Papers

We perform the same analyses on $\mathcal{T}^j$ and Δ^j where L_j is a set of multiple papers instead of one. We classify each treatment by the papers' published venue, published year, and citation counts. While integrating across multiple papers, we prompt GPT-4o to rank each paper it has access to based on the paper's relevance to the given situation, as well as its predicted internal and external validity. We observe

its reasonable ability to do so, especially when it comes to ranking among a PGG paper, a non-PGG collective behavior paper, a behavioral science paper not involving multiple agents, and a physics paper irrelevant to human behavior. GPT-4 consistently ranks the 4 papers in this order, while ignoring the physics paper completely due to irrelevance.

4. Results

4.1. The Negative Impact of Literature Augmentation

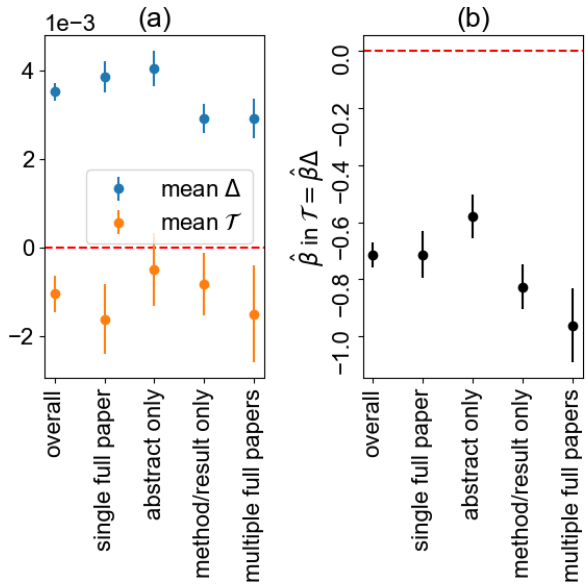

Figure 3. Illustration of Δ, $\mathcal{T}$, and their relationships upon retrieving different versions of papers. Error bars indicate 95% confidence intervals.

We find that augmenting GPT-4o with related academic papers generally does not improve the model's ability to make accurate predictions about the effect of punishment in PGG. While we observe significant shifts in the model's answers (Δ), we also find that the literature treatment effect ($\mathcal{T}$) is significantly negative in almost all cases (Figure 3(a)). It is also noteworthy that the more an LLM's response shifts after augmentation, the less accurate it gets (Figure 3(b)). These results, consistent across different versions of augmented papers, indicate that the more GPT-4o is influenced by highly-reputable PGG literature, the worse it performs in prediction tasks that are supposedly relevant. Hence, the literature is biasing the model instead of debiasing for better accuracy.

We also find no predictive power on $\mathcal{T}$ from features of individual papers that are regarded as critical in the science of science research, such as citation count, published year,

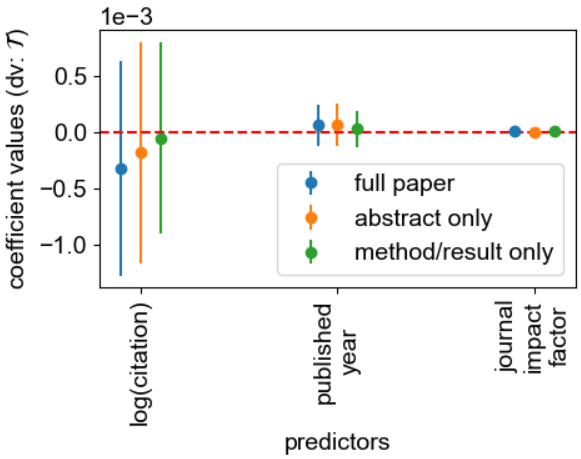

Figure 4. Coefficient plot for the linear regression involving the three metrics as predictors and $\mathcal{T}$ as the dependent variable, upon retrieving different versions of individual papers. Error bars indicate 95% confidence intervals.

and journal impact factor (Figure 4). Out of 25 papers, 4 showed the mean $\mathcal{T}$ significantly below 0 (95% CI) while 0 were significantly above. Traditional proxies for estimating the credibility of the paper's findings do not contribute to explaining our LLM method of evaluation, relating with the fact that such metadata metrics do not predict a paper's likelihood of being replicated (Youyou et al., 2023).

Despite this, one encouraging scenario for science would be when an intelligent agent reads multiple prominent papers in the area and retrieves relevant sources according to their ecological validity. Nonetheless, we do not observe such evidence when augmenting GPT-4o with 25 prominent papers on the topic. In fact, we find that only retrieving from papers with over 1,000 citations significantly worsens the predictive accuracy, suggesting a possibly depressing story of how the spread of "scientific" belief operates (Figure 5).

4.2. Internally Valid, Externally Invalid

Given how the augmented LLMs are faring, we explore whether the challenges are in the internal or external validity of the papers. If a paper L_j is internally valid within its specific design features X_j, then a PGG experiment from IED with feature vector X_i that is identical to X_j should be well-predicted by an LLM augmented with L_j.

To evaluate if this were the case, we select 5 papers where the well-defined experiment designs are similar to one another (Fehr & Gächter, 2000; 2002; Gachter et al., 2008; Herrmann et al., 2008; Nikiforakis, 2010) and measure the augmented GPT-4o's respective performance on X_i that has

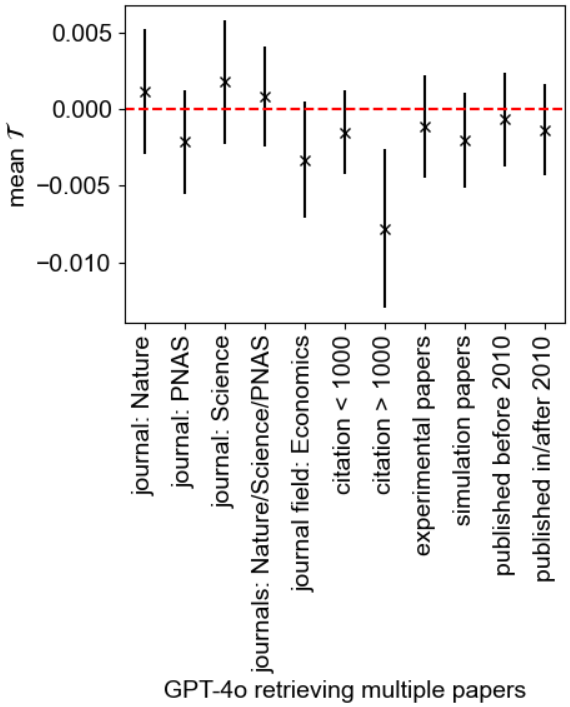

Figure 5. Illustration of literature treatment effect $\mathcal{T}$ upon reading different subsets of multiple papers. Error bars indicate 95% confidence intervals.

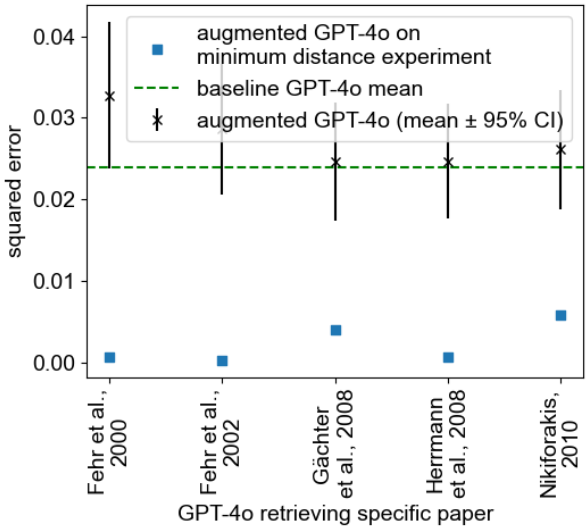

Figure 6. Illustration of the 5 PGG papers' consistent internal validity for the minimum distance experiment and lack of external invalidity on other experiments under different design choices.

the smallest normalized Euclidean distance from each paper (i.e., $X_m = \arg\min_{X_i} ||norm(X_j) - norm(X_i)||_2$). Furthermore, we compare these with each model's performance over all 85 unique X_is in IED. We find consistency over all 5 papers that while the augmented LLM performs no better than the baseline on overall average, it predicts $\hat{y}_m(X_m; L_j)$ with significantly better accuracy (Figure 6). This suggests that while the 5 papers may be internally valid, they are most likely externally invalid as the treatment effect is highly sensitive to the heterogeneity within the PGG design space.

5. Discussion

While the essence of science lies in the systematic testing of theories and hypotheses, the practice is often far from ideal especially if the system of interest is inherently complex. While most scientists are used to testing each hypothesis one at a time through carefully controlled observations, this study suggests a complementary way of evaluating the internal and external validity of a collection of research findings by using a complex instrument of a generative model.

Such an approach is motivated by studies related to inductive biases in machine learning and scientific discovery. For example, physics-informed machine learning has experienced success in improving the predictive power of a neural network by reinforcing the inductive biases that generate predictions that respect known symmetries and conservation laws (Karniadakis et al., 2021). While these symmetries and conservation laws are already shown theoretically and empirically, the fact that algorithms improve under such biases adds more confidence that such laws hold even for extremely complex applications. While the social and behavioral sciences have not discovered such fundamental invariants, an attempt to improve machine learning with behavioral theories as cognitive priors and inductive biases has demonstrated success in outperforming deep neural networks with no such biases, at least when the training data is sparse (Bourgin et al., 2019).

Both cases in distinct disciplines contribute to iteratively evaluating and discovering (Peterson et al., 2021) inductive biases that improve predictions. This approach can be applied much more broadly using the power of generative models that can integrate unstructured data to update models with already rich data and parameters. Whether such updates poison or improve the model is an empirical, meta-scientific question that this paper calls for, and one that can also be applied outside the social and behavioral sciences.

6. Limitations and Future Directions

While our work provides a concrete LLM experimental design that utilizes rich human-behavioral datasets with multiple explicit experiment design features, we expect to have more confidence in the signal of input literature while making the framework more scalable by implementing the following future steps.

First, to add robustness to our findings, we are currently collecting human survey data on the same prediction task given to GPT-4o, both from laypeople and academic experts on PGG-related topics. We expect to present an extensive comparison between baseline LLMs, augmented LLMs, human laypeople, human experts, and supervised machine learning. This will stretch beyond the illustration in Figure 1 and answer the question of whether the literature is improving the humans' and the machines' ability to make predictions under interventions.

Another means of robustness check includes running the same experiment on other available LLM models with comparable capacity to pass the manipulation checks and collecting more PGG data for a few specific X_is from the IED procedure to ensure that we have confidence in labeling the dataset with high precision.

Furthermore, to gain even more confidence that the estimated $\mathcal{T}^j$ is the best signal from L_j we can get, researchers can build on the rapidly improving research topic of LLMs reading and evaluating scientific documents and ensure stricter manipulation checks on this framework. The availability of our benchmark PGG data can attract more scientists from the machine learning community to contribute to LLMs that better predict human behavior by discovering the optimal ways to interact with academic literature through novel knowledge representation and retrieval techniques.

In the case that such advanced implementation produces $\mathcal{T} > 0$, this may imply that scientific literature is producing useful signals but in ways that are less straightforward. If $\mathcal{T}$ stays negative despite all the improvements in upcoming multimodal generative models' capacity and other retrieval techniques, we will be more inclined to diagnose that academic research on this topic is generating more noise than signals.

7. Conclusion

We propose an AI-experimental framework for evaluating the contribution of collective scientific research in making predictions under interventions. Applying the newest LLM's ability to read, evaluate, and synthesize research papers through RAG, we outline the procedure of performing manipulation checks and analyzing treatment of retrieved documents.

Furthermore, we demonstrate a concrete use case of behavioral data collected through IED that not only maps the heterogeneity of human cooperative behavior but also serves as a benchmark for future machine learning researchers developing models that make better predictions of human behavior. We hope such movements encourage more behavioral scientists to collect experimental data with comparable richness and systematic procedures, strengthening the interdisciplinary field of computational social science.

Finally, our work ignites several metascientific discussions about internal and external validity, using generative models as instruments for not only predicting complex systems but also evaluating underlying scientific biases, and open questions on how to quantify the "informativeness" of scientific findings in this integrative and generative framework. At the same time, the paper also contributes to the movement of making social science more solution-oriented (Watts, 2017) by integrating human-curated scientific knowledge into extremely complicated human and machine decision-making processes, ultimately contributing to intelligent decision support systems.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., and Watts, D. J. Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences*, pp. 1–55, December 2022. ISSN 1469-1825. doi: 10.1017/s0140525x22002874. URL http://dx.doi.org/10.1017/S0140525X22002874.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Polit. Anal.*, 31(3): 337–351, July 2023.

Baek, J., Jauhar, S. K., Cucerzan, S., and Hwang, S. J. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*, 2024.

Bakshy, E., Dworkin, L., Karrer, B., Kashin, K., Letham, B., Murthy, A., and Singh, S. Ae: A domain-agnostic platform for adaptive experimentation. In *Conference on neural information processing systems*, pp. 1–8, 2018.

Bourgin, D. D., Peterson, J. C., Reichman, D., Russell, S. J.,

and Griffiths, T. L. Cognitive model priors for predicting human decisions. In *International conference on machine learning*, pp. 5133–5141. PMLR, 2019.

Dreber, A., Rand, D. G., Fudenberg, D., and Nowak, M. A. Winners don't punish. *Nature*, 452(7185):348–351, 2008.

Fehr, E. and Gächter, S. Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994, 2000.

Fehr, E. and Gächter, S. Altruistic punishment in humans. *Nature*, 415(6868):137–140, 2002.

Gachter, S., Renner, E., and Sefton, M. The long-run benefits of punishment. *Science*, 322(5907):1510–1510, 2008.

Herrmann, B., Thoni, C., and Gachter, S. Antisocial punishment across societies. *Science*, 319(5868):1362–1367, 2008.

Horton, J. J. Large language models as simulated economic agents: What can we learn from homo silicus?, 2023. URL https://arxiv.org/abs/2301.07543.

Hu, X. E., Gandhi, L., Whiting, M. E., Watts, D. J., and Almaatouq, A. Tasks beyond taxonomies: A multidimensional design space for team tasks.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, May 2021. ISSN 2522-5820. doi: 10.1038/s42254-021-00314-5. URL http://dx.doi.org/10.1038/s42254-021-00314-5.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Ludwig, J. and Mullainathan, S. Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827, January 2024. ISSN 1531-4650. doi: 10.1093/qje/qjad055. URL http://dx.doi.org/10.1093/qje/qjad055.

Manning, B. S., Zhu, K., and Horton, J. J. Automated social science: Language models as scientist and subjects, 2024. URL https://arxiv.org/abs/2404.11794.

Mei, Q., Xie, Y., Yuan, W., and Jackson, M. O. A turing test of whether AI chatbots are behaviorally similar to humans. *Proc. Natl. Acad. Sci. U. S. A.*, 121(9): e2313925121, February 2024.

Nikiforakis, N. Feedback, punishment and cooperation in public good experiments. *Games and Economic Behavior*, 68(2):689–702, 2010.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., et al. Replicability, robustness, and reproducibility in psychological science. *Annual review of psychology*, 73:719–748, 2022.

Palfrey, T., Rosenthal, H., and Roy, N. How cheap talk enhances efficiency in threshold public goods games. *Games and Economic Behavior*, 101:234–259, 2017.

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

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., and Nowak, M. A. Positive interactions promote public cooperation. *Science*, 325(5945):1272–1275, 2009.

Simons, D. J., Shoda, Y., and Lindsay, D. S. Constraints on generality (cog): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6): 1123–1128, 2017.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., Marks, D., Ramsundar, B., Song, L., Sun, J., Tang, J., Veličković, P., Welling, M., Zhang, L., Coley, C. W., Bengio, Y., and Zitnik, M. Scientific discovery in the age of artificial intelligence. *Nature*, 621 (7978):E33–E33, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06559-7. URL http://dx.doi.org/10.1038/s41586-023-06559-7.

Watts, D. J. Should social science be more solution-oriented? *Nature Human Behaviour*, 1(1):0015, 2017.

Yarkoni, T. The generalizability crisis. *Behavioral and Brain Sciences*, 45:e1, 2022.

Youyou, W., Yang, Y., and Uzzi, B. A discipline-wide investigation of the replicability of psychology papers over the past two decades. *Proceedings of the National Academy of Sciences*, 120(6), January 2023. ISSN 1091-6490. doi: 10.1073/pnas.2208863120. URL http://dx.doi.org/10.1073/pnas.2208863120.

Zhou, R., Chen, L., and Yu, K. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9340–9351, 2024a.

Zhou, Y., Liu, H., Srivastava, T., Mei, H., and Tan, C. Hypothesis generation with large language models, 2024b. URL https://arxiv.org/abs/2404.04326.