Predicting Field Experiments with Large Language Models

Yaoyu Chen
ychen563@uic.com
University of Illinois at Chicago
Chicago, Illinois, USA

Yuheng Hu[*]
hu.3331@osu.edu
The Ohio State University
Columbus, Ohio, USA

Yingda Lu[**]
yingdalu@uic.com
University of Illinois at Chicago
Chicago, Illinois, USA

Abstract

Large language models (LLMs) have demonstrated unprecedented emergent capabilities, including content generation, translation, and simulation of human behavior. Field experiments, on the other hand, are widely employed in social studies to examine real-world human behavior through carefully designed manipulations and treatments. However, field experiments are known to be expensive and time consuming. Therefore, an interesting question is whether and how LLMs can be utilized for field experiments. In this paper, we propose and evaluate an automated LLM-based framework to predict the outcomes of a field experiment. Applying this framework to 276 experiments about a wide range of human behaviors drawn from renowned economics literature yields a prediction accuracy of 78%. Moreover, we find that the distributions of the results are either bimodal or highly skewed. By investigating this abnormality further, we identify that field experiments related to complex social issues such as ethnicity, social norms, and ethical dilemmas can pose significant challenges to the prediction performance.

CCS Concepts

• Information systems → Collaborative and social computing systems and tools.

Keywords

Large Language Models, Field Experiments

1 Introduction

Field experiments allow researchers to manipulate variables of interest in a real-world setting, such as human behaviors like ad clicking or donation giving, establishing causal relationships between interventions and outcomes. They typically begin by designing an intervention aligned with a specific research question, randomly assigning participants to treatment or control groups, and then measuring outcomes under natural conditions. The resulting data are collected and analyzed to evaluate the causal impact of the intervention [27]. It is adopted by a wide range of disciplines across academia and industry, such as finance [5, 14], marketing [17, 28], and organizational studies [23, 31]. In recent years, online field experiments, also known as A/B testing, have revolutionized numerous online platform designs, advertising strategies, etc. While effective, field experiments are also known to be expensive and time consuming. For example, some experiments may take months or years to conduct [35]. Moreover, recruiting high-quality participants for the experiments is challenging and costly [3], potentially affecting the experiment's outcomes.

In recent years, Large language models (LLMs) have demonstrated unprecedented emergent capabilities, including content generation, translation, and simulation of human behavior. For example, existing studies have explored the alignments between simulated data generated by LLMs and real data collected from human participants across various aspects, including human responses, traits [16], moral standards [13], preferences [36], and emotions [18]. Therefore, it is becoming increasingly popular to leverage LLMs for simulating human behaviors [19, 48]. For example, there are a handful of studies in which scholars have successfully instructed LLMs to replicate existing lab experiments across several disciplines, including psychology [2, 11], sociology [22, 26, 34], and economics [2, 24]. Their aim is to replicate existing lab experiments by treating LLMs as participants in lab experiments.

However, directly applying these works to field experiment simulations in our context has several challenges. First, they mainly focused on lab experiment settings. But field experiments are inherently more challenging to conduct than laboratory experiments, due to diverse participant backgrounds, complex workflows, and multifaceted treatment designs. Second, most of the previous studies mainly relied on manual processes, and because of that, tested a handful of experiments with a limited scope of topics. While some recent studies tested on relatively large data, the selected experiments were limited to Likert-based psychological or social surveys [11, 22]. In order to explore LLMs' capabilities, robustness, and generalizability in field experiment simulation, we need to test them on a much larger scale and broader range of experiments on different topics.

In this paper, we fill the literature gap by proposing an automated LLM framework to predict the outcome of a wide range of field experiments. Our framework has several major components, such as an information extraction module, a variant generation module, and a prediction module. Specifically, the information extraction module extracts key experiment settings, while the variant generation module generates false variants as distractors to confuse LLMs. Finally, the prediction modules leverage two prompt templates with a Chain-of-Thought design, prompting the LLM to predict the outcomes of a field experiment.

We test this framework on 276 field experiments reported in premier academic journals from 2000 to 2024. Those experiments contain a total of 1261 conclusions with a wide range of topics

[*] Both authors contributed equally to this research.

SLW, Toronto

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

such as labor-market discrimination, educational incentives, household finance behavior, and the impact of healthcare enrollment. Without any alignment techniques or fine-tuning, our framework achieves an average prediction accuracy of 78%. We also test for data memorization effects by examining prediction results on recent experiments that appeared in 2024, which would be less likely to be included in the training data of LLMs. More interestingly, we also find that the prediction results are either bimodal or highly skewed. For example, our framework achieves nearly 100% prediction accuracy on 71% of conclusions while it completely fails to predict 18% of the conclusions with close to 0% accuracy. Further analyses reveal that our LLM-based framework has limitations in predicting experiments related to complex social issues such as ethnicity, social norms, and ethical dilemmas.

Our research makes several contributions:

1 We extend the current literature on LLMs' emergent capabilities by demonstrating that LLMs can simulate field experiments by predicting conclusions. While using LLMs to simulate human behavior is well studied in recent years, to the best of our knowledge, this is the first work in the literature to replicate large-scale field experiments that require a more complex environment setting and workflow design.
2 Our proposed framework enables the prediction of field experiment conclusions in a fully automated, large-scale fashion. As a result, our framework is robust and can be generalized to a wider range of downstream applications in field experiments.
3 This paper examines the prediction performance and reveals the limitations of LLMs in field experiment simulations.

2 Literature Review

Our study is related to the LLM literature on simulating human behavior, with a particular focus on experimental simulations by LLMs. Here, we highlight our contributions by comparing and contrasting our work with existing studies.

While past literature mainly focused on agent-based social simulation [9, 45, 49], there is an increasing trend to adopt LLMs as simulation tools. Existing studies have found that LLMs' ability to simulate human behavior stems from their possession of human-like reasoning skills and their adaptivity to personas of diverse characters [40, 43]. Upon those features of LLMs, Aher et al. [2] proposed the concept of "Turing Experiment", in which LLMs are profiled as synthetic participants of experiments with integrated prompt of experimental settings and demographic information. Similarly, Horton [24] demonstrated LLMs' ability to simulate lab experiments and promoted it as a method of experimental pilot testing. The usage of LLMs in Horton [24]'s work is similar to Aher et al. [2], consisting of two stages: prompting LLMs as synthetic experimental participants and collecting responses from conversations, reporting a few successful simulation cases of lab experiments. Leng & Yuan [26] harnessed a three-phase procedure to complete the lab experimental simulation, which includes the initialization phase, interaction phase, and decision analysis phase. In the initialization phase, separate conversations of GPT-4 are prompted as vanilla experimental participants without specifying demographic information. Then, in the interaction stage, synthetic participants are prompted with the actions of other participants and asked to take actions and rationales according to the experiment design. Last, agents' actions and rationales are collected for analysis to conclude. Leveraging this procedure, the study simulates five existing lab experiments. Manning et al. [34] proposed an automated framework for various lab experimental simulations. Although the entire workflow is divided into seven steps, it is essentially similar to the manual procedures of the three papers above. However, the sole input is the backstory of the experiment. Based on the input information and powered by an LLM, the framework continues with subsequent steps: identifying dependent and independent variables, generating treatment values, profiling agents, organizing the interaction of agents, collecting data, and establishing causal relationships and conclusions. The paper reports the results of four social scenarios. Besides the aforementioned studies, which were tested only on a small scale of cases, there are extant papers that implemented large-scale testing on LLM-based experimental replication. Cui et al. [11] replicated 154 psychological experiments. Specifically, they profiled LLMs as either students or adults, retrieving Likert-like responses from LLMs, which is oversimplified compared to most lab and field experiments. By comparison, Ashokkumar et al. [22] extended such a massive replication to 70 more complex survey experiments, though their selected experiments were limited exclusively to US social surveys and did not include any field experiments in the primary tests.

In summary, while existing studies discussed above pioneered the application of Large Language Models (LLMs) to laboratory experiment simulations, several research gaps remain unaddressed. For instance, existing methods are not tested on field experiments that are inherently more complex than laboratory experiments, encompassing diverse participant backgrounds, more intricate workflows, and multifaceted treatment designs. Besides, despite extensive discussions on LLM bias [8, 15, 46] and the fact that recognized biases in LLMs—such as gender [25] and social norm bias [39]—can compromise performance on downstream tasks [21], only one of these studies mentioned the impact of certain topics on the replication successful rate of psychological surveys [11]. It is still unclear how the joint effect of topics and sentiments would undermine the fidelity of field experiment prediction. Another limitation is that these studies tested their simulation strategies on only a small number of experiments, restricting generalizability. To address these gaps, we evaluated our proposed framework at scale on field experiments from published papers, demonstrating not only that LLMs can accurately predict the outcomes of established experiments, but also clarifying boundary conditions under which they cannot provide reliable experimental predictions. Furthermore, current automated simulation framework [11, 22, 34] are incompatible with scenarios involving complex treatments involving human-object interactions and limit only to lab experiments with Likert response or survey. By contrast, our framework offers a broadly adaptable approach that supports experiments across a wide range of contexts.

3 Data Collection and Filtering

To explore LLMs' ability to predict field experiments in a large-scale fashion, we first need to collect existing field experiments. Inspired by prior studies that focus on using LLMs to simulate human behaviors in lab experiments from existing psychology literature, we also consider field experiments studied in premier journals in economics, such as The Review of Economic Studies, American Economic Review, Journal of Political Economy, and The Quarterly Journal of Economics. It is worth noting that we focus on field experiments in economics because they are generally a larger scale in terms of participant size and robust in terms of careful design.

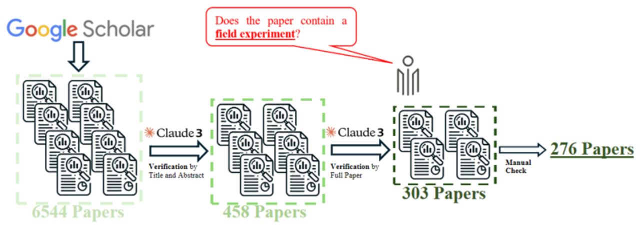

Google Scholar
Does the paper contain a field experiment?
Claude 3
Verification by Title and Abstract
Claude 3
Verification by Full Paper
Manual Check
6544 Papers
458 Papers
303 Papers
276 Papers

Figure 1: The Data Collection Workflow.

Figure 1 shows the data collection and filtering workflow. Initially, 6544 papers containing keywords related to field experiments published between 2000 and 2024 of these top journals were selected. Then, we applied a two-layer verification process powered by Claude (Claude-3-opus-20240229). First, we prompted the title and abstract of each paper to Claude and asked it to judge if the paper designs and implements a field experiment. Upon this, the second verification prompted the entire paper to Claude, asking the same question. This strategy balances the accuracy of verification and the cost of calling Claude, as prompting the entire paper exponentially increases the cost. A final rule-based manual check to ensure the fidelity of the automated selection and filter out 276 papers for testing. More details about the manual check are available in Section 4.4. It is also worth mentioning that using Claude as the verification tool, instead of GPT, prevents potential data leakage as GPTs are leveraged as the prediction tool in our framework. The distribution of the selected papers is shown in Figure 7 (In Appendix A).

4 Framework

We present the details of our automated framework for predicting field experiments in this section. Figure 2 shows the workflow. Overall, our framework is divided into three stages: Extraction from Papers, Variant Generation, and Prediction. Notably, Claude (Claude-3-opus-20240229) powers all preprocess tasks in the first two stages, whereas GPT completes the prediction at the last stage. We use two different LLMs in different stages to prevent potential data leakage from one another.

4.1 Extraction

Specifically, the framework uses Claude to extract information related to a field experiment from the selected paper. To realize this, the framework leverages a manually crafted prefixed prompt, which has proven to be efficient for various downstream tasks [7]. As shown in Figure 8 (In Appendix B.1), the prompt template contains a placeholder "*Paper*", an information form consisting of bullet points from "A" to "G", and clear instructions that ask the LLM to extract information according to the form from the paper. As underscored in the prompt, the first six bullet points "A" to "F" are key experimental settings that shape the experiment context, whereas the last point "G" is about conclusions that are true outcomes in the prediction task.

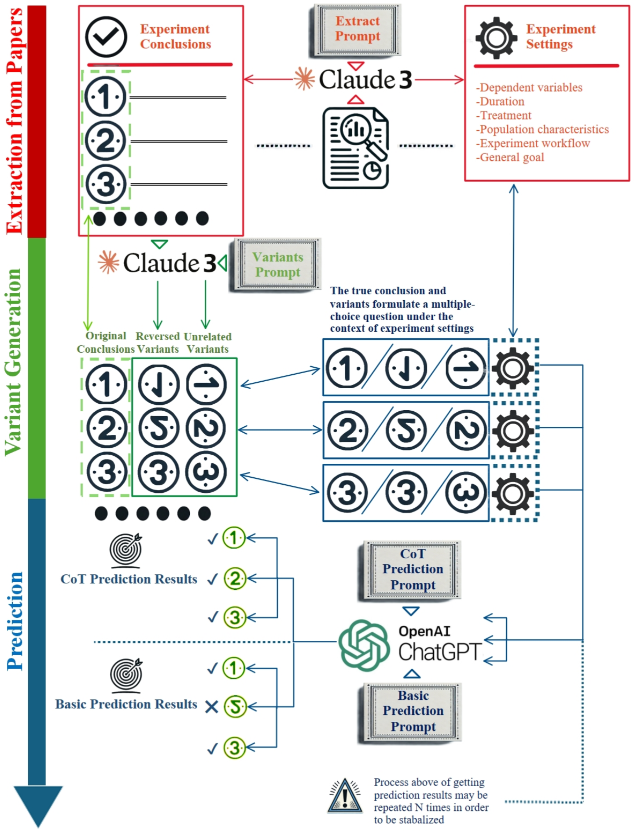


Figure 2: Prediction Framework.

Based on the response from Claude, the framework formulates "Experiment Settings" directly from bullet points "A" to "F", while it polishes point "G" to generate "Experiment Conclusions". Specifically, the raw response regarding point "G" is a paragraph containing multiple conclusions of the field experiment. To separate that paragraph into standalone conclusions, the framework calls a new Claude session, prompting the raw paragraph and related instructions to it, finally getting "Experiment Conclusions" in return. Breaking complex tasks into subtasks improves the performance of LLM-driven workflows [50], which is the main reason for completing the extraction and separation of conclusions in different Claude sessions.

4.2 Variant Generation

After the experiment settings and conclusions are extracted, the next step is to generate variants based on the true conclusion, since the goal is to see if the LLM could select the true one under distraction. Inspired by Luo et al. [33]'s prediction of neuroscience results by LLMs, for each conclusion from an experiment, our framework prompts the original conclusion and its two variants to GPT: a reversed variant and an unrelated variant. As a result, the framework will make the prediction by choosing one of three options.

Specifically, as shown in Figure 9 (In Appendix B.2), the framework initially prompts the original conclusion to Claude, which follows the instructions to generate the reversed variant of the original conclusion. The reversed variant means that the direction of the conclusion is inverted. For example, if one conclusion is "*receiving housing vouchers reduces quarterly employment rates*," its reversed variant will be "*receiving housing vouchers increases quarterly employment rates*." Next, the framework prompts both the original and reversed conclusions to Claude to generate the unrelated variant, which typically indicates that there is no correlation between entities of interest. Following the same example, the unrelated variant will be "*There is no relationship between receiving housing vouchers and quarterly employment rates*."

4.3 Prediction

In the final stage, our framework takes a field experiment's experiment settings, conclusion, and its two variants as input and generates two parallel prediction prompts: basic prediction prompt and Chain-of-Thought (CoT) prompt, which are then prompted to GPT to get predictions by asking GPT to select one conclusion from three conclusions. We also rely on CoT as it has proven to be capable of improving the general performance of LLMs on downstream tasks [47].

Specifically, the basic prediction prompt is shown in Figure 10 (In Appendix B.3). It consists of a background information section, a question section, and necessary instructions. Specifically, the background information section contains the general goal of the experiment (such as exploring the impact of job training on income), treatments (such as receiving job training or not), experiment duration (such as seven weeks), outcomes (such as income), participant information (such as people seeking jobs in New England), and experiment workflow (such as when and how training was given and outcomes were recorded). All of these were extracted from a target paper, which is the same as bullet points "A" to "F" from Figure 8 (In Appendix B.1).

Following the background information, a question section is automatically generated and entered into the templates. Specifically, the original conclusion, its reversed variant, and its unrelated variant are shuffled and substitute the placeholders "option 1", "option 2", and "option 3". Meanwhile, instructions tell GPT that the prediction of conclusions is under the context of the field experimental settings, asking GPT to choose one option from the three options as it deems correct.

As we are also interested in how CoT would improve such prediction, Figure 11 (In Appendix B.3) shows the CoT Prediction Prompt, which follows a similar logic as the basic prediction but integrates CoT strategies to boost the performance [47]. Initially, the framework prompts the experiment settings and three options for a conclusion to GPT, instructing it to think about decisive elements that help choose the correct option. Upon receiving the decisive elements from GPT, the framework prompts GPT to make a selection among three options to get a predicted conclusion. It is worth noting that the entire process is within the same GPT session for a conclusion.

Although either prompt strategy generates a prediction for a given input, LLMs are stochastic models, meaning that their responses may vary to the same prompt. To handle such randomness in experiment simulations, Leng & Yuan [26] set the temperature to 0 and always get fixed responses from synthetic participants of lab experiments, which is a strategy to eliminate the stochasticity of LLMs totally. By contrast, Brand et al. [6] repeated the same prompt 300 times and used the averaged number as the result of Willingness-to-Pay from customers role-played by LLMs. Here, we take the latter approach by incorporating the stochasticity of LLM outputs since this stochasticity of LLMs is similar to how the same human participant might respond differently when presented with the same instruction [12]. Specifically, we repeat the same prediction prompt several times and calculate an average accuracy as the final result. For example, if the framework is running the basic prediction, a filled-out prompt based on Figure 10 (In Appendix B.3) will be repeated a given number of times to get a stable result. The determination of a proper repeat number will be further discussed in the Results section.

Parameter-wise, no fine-tuning is involved in any stage of the proposed framework, and all parameters of OpenAI API and Anthropic API are set to default. Whereas Horton [24] harnessed fine-tuned LLMs in lab experimental simulation to better follow instructions, Coda-Forno et al. [10] simulated several human behaviors by LLMs without fine-tuning. The use of fine-tuned LLMs complicates the reproduction attempts since other researchers don't have access to the same model in the existing papers [4]. Additionally, avoiding fine-tuning LLMs saves computing resources and mitigates environmental impacts [38], especially when the pretraining of LLMs is enough to make them capable of downstream tasks [30].

4.4 Robustness Checks

Given that most steps in the proposed framework are automated, concerns naturally arise regarding the validity of these automated processes. To address these concerns, we conduct three manual screenings to verify the results from extraction (Section 4.1) and variant generation (Section 4.2). First, we examine whether the extracted experiment settings (Figure 8, Appendix B.1) inadvertently include genuine conclusions. Second, we check whether the extracted conclusions align with those reported in the original papers. Finally, we assess whether the generated variants of the conclusions (Figure 9, Appendix B.2) match our expectations.

The first and third screenings revealed no issues. However, the second screening, which examined the alignment of extracted conclusions, identified 377 conclusions as either incomprehensible or nonexistent in the original texts. Consequently, after the three screenings, 1261 conclusions and 276 papers remained and were deemed valid for our purposes.

Additionally, 86 out of 1261 conclusions were dequantified, as existing studies suggest that predicting the magnitude of experimental outcomes remains challenging at this stage [34]. Given our focus on predicting the direction of experimental conclusions rather than the magnitude, conclusions specifying precise numerical treatment effects were reformulated in a dequantified manner. For instance, a conclusion such as "Job training increases income by 30%" was revised to "Job training increases income."

5 Results

In this section, we test our framework on 276 field experiments that contain a total of 1261 conclusions and discuss the performance.

We use Conclusion Accuracy and Paper Accuracy to evaluate the prediction performance under different settings. As shown in Figure 2, the framework generates a prompt for each conclusion. The generated prompt either follows the template in Figure 10 or 11 (In Appendix B.3), depending on which prompt strategy (basic or CoT) the framework applies. Each prompt instructs GPT to output a predicted conclusion. If the predicted conclusion matches the true conclusion, that attempt is counted as correct. As aforementioned, the framework repeats such an attempt for a set number of times for each conclusion to get a stable result. Therefore, we define *Conclusion Accuracy* as the percentage of correct predictions among a set number of attempts (Equation 1). Given that a field experiment may contain multiple conclusions, *Paper Accuracy* is the average of all *Conclusion Accuracy* within a paper (Equation 2).

$$\text{Conclusion Accuracy} = \frac{\text{Num of Correct Predictions}}{\text{Num of Predictions}} \times 100\% \quad (1)$$

$$\text{Paper Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \text{Conclusion Accuracy}_i \quad (2)$$

5.1 Prediction Performance Overview

Table 1: Prediction Accuracy under Different Repeats

GPT-4 Turbo	Basic		CoT	
	Conclusion Acc.	Paper Acc.	Conclusion Acc.	Paper Acc.
Repeat = 10	66%	66%	76%	76%
Repeat = 20	66%	66%	76%	76%
Repeat = 30	65%	66%	76%	76%
ANOVA (Basic)	F = 0.035, $p = 0.9994$		–	
ANOVA (CoT)	–		F = 0.016, $p = 0.9999$	

Sample size: 1,261 conclusions from 276 papers. Models: `gpt-3.5-turbo-0125`, `gpt-4-turbo-2024-04-09`, and `gpt-4o-2024-11-20`.

Table 1 reports two types of accuracy for both strategies under different repeats by GPT4-turbo. The best results for the basic strategy are obtained under 10 repeats, which are 66% for both conclusion accuracy and for paper accuracy. By comparison, the CoT results are the same 76% for both strategies and different repeats, which are generally 10% points higher than basic results and aligns with prior literature on boosting LLM performance by CoT [47]. Meanwhile, another key observation is that the results show a significant invariability to repeat numbers. Specifically, the results across different numbers of attempts are significantly static under either the basic or CoT strategy. Given that more repeats result in higher time and monetary cost, we chose 20 repeats for further evaluation.

Table 2: Prediction Accuracy under 20 Repeats by GPT Models (point estimate with 95% CI)

Model	Basic		CoT	
	Conclusion Acc.	Paper Acc.	Conclusion Acc.	Paper Acc.
GPT-3.5 Turbo	61% [58%–63%]	61% [58%–64%]	68% [66%–69%]	67% [65%–70%]
GPT-4 Turbo	66% [63%–68%]	66% [63%–69%]	76% [74%–78%]	76% [73%–79%]
GPT-4o	75% [73%–77%]	75% [72%–78%]	78% [76%–80%]	78% [76%–81%]

Sample size: 1,261 conclusions from 276 papers. Models: `gpt-3.5-turbo-0125`, `gpt-4-turbo-2024-04-09`, and `gpt-4o-2024-11-20`. Values are whole-number percentages with 95% confidence intervals.

Table 3: Pairwise t-tests of Model Accuracy

Metric	Model A (Accuracy%)	Model B (Accuracy%)	p
Conclusion	GPT-4o CoT (78%)	GPT-4 Turbo Basic (66%)	$< 0.001^{***}$
Conclusion	GPT-4o CoT (78%)	GPT-4 Turbo CoT (76%)	0.212
Conclusion	GPT-4o Basic (75%)	GPT-4o CoT (78%)	0.068
Conclusion	GPT-4o Basic (75%)	GPT-4 Turbo Basic (66%)	$< 0.001^{***}$
Conclusion	GPT-4o Basic (75%)	GPT-4 Turbo CoT (76%)	0.611
Conclusion	GPT-4 Turbo Basic (66%)	GPT-4 Turbo CoT (76%)	$< 0.001^{***}$
Paper	GPT-4o CoT (78%)	GPT-4 Turbo Basic (66%)	$< 0.001^{***}$
Paper	GPT-4o CoT (78%)	GPT-4 Turbo CoT (76%)	0.235
Paper	GPT-4o Basic (75%)	GPT-4o CoT (78%)	0.077
Paper	GPT-4o Basic (75%)	GPT-4 Turbo Basic (66%)	$< 0.001^{***}$
Paper	GPT-4o Basic (75%)	GPT-4 Turbo CoT (76%)	0.546
Paper	GPT-4 Turbo Basic (66%)	GPT-4 Turbo CoT (76%)	$< 0.001^{***}$

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 2 reports results by different GPT models with 20 attempts and Table 3 reports corresponding pairwise t-tests for top models. One key observation from Table 2 is that prediction performance steadily improves as LLMs iterate, though the rate of improvement decreases over time. According to Table 2, the best result under 20 repeats is achieved by GPT4o under CoT prompt strategy, which is a conclusion accuracy of 78% and a paper accuracy of 78%. Specifically, on average, GPT4o is able to predict a conclusion in 78% of the 20 repeated prediction attempts, and it also predicts 78% of the outcomes correctly for each paper. While the conclusion accuracy of GPT4o under CoT strategy is 10 percentage points significantly higher than the CoT result of GPT3-turbo, it is only two percentage points and insignificantly higher than the CoT result of GPT4-turbo, indicating the improvement from model iteration becomes harder for this task.

Another interesting finding from Table 2 is the boosting effect of CoT on performance varies on models for the experiment conclusion prediction task. Specifically, the largest improvement is 10 percentage points on GPT4-turbo, while the least improvement is three percentage points on GPT4o, which is not significant. However, it's unsafe to conclude that CoT boosting is weaker on newer models since CoT improves the accuracy by seven percentage points on the oldest model tested, GPT3-turbo.

In summary, our findings indicate that incorporating a CoT strategy significantly enhances predictive performance in GPT3-turbo and GPT4-turbo, whereas increasing the number of repetitions does not produce a significant change. Furthermore, iterative refinements to LLMs consistently improve performance, although the rate of improvement diminishes over time. Finally, the performance gains attributable to CoT appear to be model-sensitive.

5.2 Data Memorization

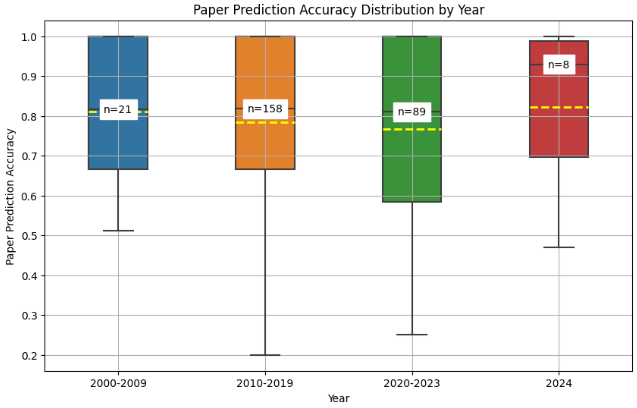


Figure 3: Paper Accuracy by Year.

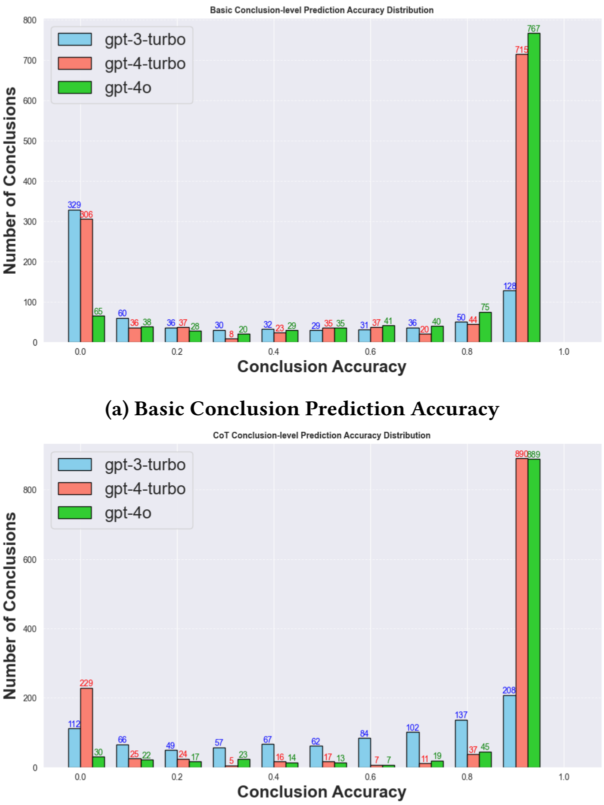
Data memorization is a common concern in simulating experiments with LLMs. If the results given by the LLM are from its memory of training data instead of reasoning, the proposed idea has no instructional value as pilot testing for field experiments [24]. As revealed in its documents, the training data cutoff of gpt-4o-2024-11-20 is October 2023. Therefore, it is reasonable to assume that the model would not have seen the field experiments papers appeared in 2024. In other words, field experiments published after 2024 would be less likely to be included in the training data.

To text our framework's robustness in light of data memorization, we split the papers by year. Figure 3 shows that the paper accuracy is actually higher on papers 2024 than any other years. This is notable because if the prediction results are driven by LLMs' memorization of its training data, the framework's performance in unseen papers in 2024 would be poorer, indicating data memorization of less concern to our framework.

5.3 Examination of Distributions of Results

To further examine the prediction performance of our proposed framework, we plot the distributions of accuracy results reported in Table 2, based on different GPT models and two prompt strategies under 20 repeats (Figure 4 and 5).

As shown in Figure 4a, the conclusion accuracy of all three models (gpt-3, gpt4-turbo and gpt-4o) exhibits bimodal patterns, with peaks concentrated toward both the lower and upper extremes (a U-shaped distribution). Specifically, there are significant concentrations of samples in the 90%-100% accuracy range and another notable cluster in the 0%-10% accuracy range. This indicates that the model's performance is highly polarized, where certain conclusions are predicted with near-perfect accuracy, while others are almost wrongly predicted entirely. Interestingly, by applying CoT strategy (Figure 4b), the concentration in the 0%-10% accuracy range is mitigated, while the concentration in the 90%-100% accuracy range deepens, resulting from the performance boosting by CoT. Model-wise, the U-shaped distributions for the earlier model (gpt-3-turbo) are milder compared to recent models (gpt-4-turbo and gpt-4o), probably resulting from the performance limitation of earlier models.



(a) Basic Conclusion Prediction Accuracy

(b) CoT Conclusion Prediction Accuracy

Figure 4: Conclusion Prediction Accuracy Distribution

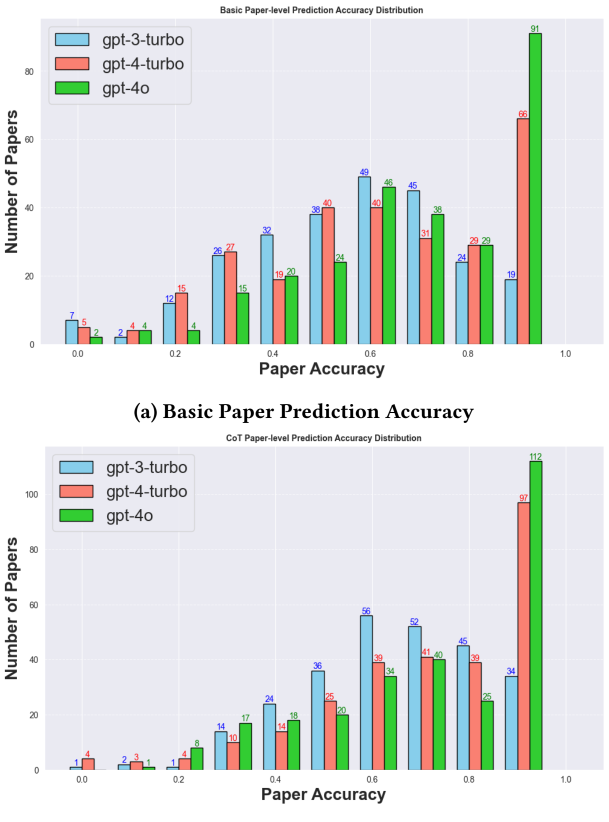
Similar patterns also exist in Figure 5 to show the paper's accuracy. Distributions for all models are skewed. Such a skewness is more pronounced in gpt-4-turbo and gpt-4o, with samples concentrating in the upper extreme. This indicates that the framework correctly predicts all conclusions from certain tested experiments. Additionally, CoT generally increases the degree of skewness(Figure 5b), aligning with its boosting effect on LLMs' performance.

Inspired by the bimodal and skewed results (Section 5.3), we closely examined the topics of each conclusion based on its experimental context. Leveraging LLMs' ability in annotating [42], we prompted Claude to label topic components of each conclusion under the context of the experiment, as shown in Figure 12. To ensure selected topics could grasp the reason behind the abnomalities found previously (Figure 4 and Figure 5), we included topics where LLMs' responses are biased [1, 20, 29, 32, 44], including gender, ethnicity, social norms, ethical dilemmas, age, socioeconomic status, and other topics. As a result, each conclusion was represented by a vector of percentages summing to 100%, where each percentage indicated the degree to which the context was associated with a particular topic. Furthermore, as the sentiment bias also affects the generated content of LLMs on top of topics bias [37], we used Claude as the sentiment analysis tool to label each conclusion, deciding the sentiment of each conclusion, either positive, negative, or neutral (Figure 13, Appendix B.4). In addition to regular sentiments, for gender-related conclusions, Claude also labeled each of them with a gender favorability tag, as LLMs may favor and tend to generate pro-female content [41]. Specifically, if the context of a conclusion relates to gender, Claude would further judge if the context is favorable to females or detrimental to males, the opposite, or neutral, as shown in Figure 14. Finally, all variables acquired from the labeling process are summarized in Table 5 (In Appendix C).

To study the impact of topic components and sentiments on LLMs' performance in experimental prediction, we constructed a regression model (Equation 3). Table 4 reports six regression results by three GPT models and two prompt strategies under 20 repeats on 1261 samples, which corresponds with the conclusion accuracy results in Table 2.

$$
\begin{aligned}
\text{Conclusion Accuracy} = \beta_0 &+ \beta_1(\text{Gender}) + \beta_2(\text{Ethnicity}) + \beta_3(\text{Social norms}) \\
&+ \beta_4(\text{Ethical dilemmas}) + \beta_5(\text{Age}) + \beta_6(\text{Socioeconomic status}) \\
&+ \beta_7(\text{Other topics}) + \beta_8(\text{Gender} \times \text{Favorability}) \\
&+ \beta_9(\text{Ethnicity} \times \text{Sentiment}) + \beta_{10}(\text{Social norms} \times \text{Sentiment}) \\
&+ \beta_{11}(\text{Ethical dilemmas} \times \text{Sentiment}) + \beta_{12}(\text{Age} \times \text{Sentiment}) \\
&+ \beta_{13}(\text{Socioeconomic status} \times \text{Sentiment}) + \beta_{14}(\text{Other topics} \times \text{Sentiment}) \\
&+ \epsilon
\end{aligned}
\tag{3}
$$

Figure 6: Full specification of the regression model used to examine the effect of topic–sentiment interactions on prediction accuracy.



(a) Basic Paper Prediction Accuracy

(b) CoT Paper Prediction Accuracy

Figure 5: Paper Prediction Accuracy Distribution

There are several key findings from regression results (Table 4). First, the interaction effects between certain topics and sentiments significantly affect the performance, though there is no significant evidence that topic components alone may affect LLMs' ability in

interaction term are significant across most gpt models, suggesting that the iteration LLMs didn't fix this bias. By comparison, the coefficients of the interaction of social norms are no longer significant for more recent models, suggesting that this bias might have been fixed.

Third, applying CoT strategy might mitigate the impact of sentiment interaction with certain topics. According to the regression results in Table 4 (1) and Table 4 (2), implementing CoT turns some significant estimated coefficients of interaction terms into non-significant. However, this is not the case for gpt3-turbo model, as CoT brings more significant coefficients.

6 Conclusion

In this paper, we propose an LLM-powered framework that automatically extracts information from existing papers and predicts field experimental conclusions. To the best of our knowledge, this is the first paper to provide an automated framework for predicting such conclusions for field experiments. Rather than merely introducing the framework, our work also examines its fidelity on a large scale of samples and achieves a considerable accuracy of 78%. Furthermore, the paper discovers that incorporating a CoT strategy generally enhances predictive performance in this scenario, whereas the performance gains attributable to CoT appear to be model-dependent. Furthermore, iterative refinements to LLMs consistently improve performance, although the rate of improvement diminishes over time.

Interestingly, the paper also finds that the distributions of prediction accuracy are either bimodal or negatively skewed, with a large number of samples concentrating on the two extremes. To explore this phenomenon, the paper regresses conclusion prediction accuracy on topic components and sentiments, revealing that interaction effects between certain topics and sentiments could affect the LLMs' prediction performance in this task.

Taken together, these findings underscore the potential of the LLM-driven frameworks in advancing automated predictions for field experiments while also clarifying the practical constraints that guide their effective use.

Table 4: OLS Regression Results

	(1)	(2)	(3)	(4)	(5)	(6)
Variable	GPT-4o Basic	GPT-4o CoT	GPT-4 Turbo Basic	GPT-4 Turbo CoT	GPT-3.5 Turbo Basic	GPT-3.5 Turbo CoT
Constant	0.812 (0.731)	-1.614 (0.516)	3.729 (0.166)	-0.931 (0.710)	-0.060 (0.983)	-0.624 (0.767)
Gender gap	-0.080 (0.973)	2.336 (0.347)	-3.261 (0.225)	1.701 (0.497)	0.438 (0.874)	1.195 (0.570)
Ethnicity	-0.078 (0.974)	2.459 (0.321)	-3.210 (0.231)	1.561 (0.532)	0.635 (0.818)	1.250 (0.551)
Social norms	-0.091 (0.969)	2.502 (0.316)	-2.995 (0.268)	1.520 (0.546)	0.666 (0.810)	1.230 (0.561)
Ethical dilemma	-0.654 (0.783)	2.045 (0.413)	-3.069 (0.256)	2.046 (0.416)	0.810 (0.770)	1.427 (0.500)
Age	0.032 (0.989)	2.365 (0.343)	-3.184 (0.239)	1.528 (0.544)	0.664 (0.811)	1.193 (0.572)
Socioeconomic status	0.094 (0.968)	2.512 (0.312)	-3.021 (0.261)	1.723 (0.492)	0.554 (0.841)	1.345 (0.522)
Other topics	-0.094 (0.968)	2.328 (0.349)	-3.186 (0.237)	1.557 (0.535)	0.671 (0.808)	1.195 (0.570)
Gender × Favorability	0.167 (0.326)	0.343* (0.055)	0.094 (0.628)	0.072 (0.690)	-0.071 (0.720)	0.202 (0.184)
Ethnicity × Sentiment	0.270* (0.053)	0.118 (0.422)	0.431*** (0.007)	0.270* (0.068)	0.149 (0.360)	0.296** (0.017)
Social norms × Sentiment	0.092 (0.636)	0.066 (0.747)	0.148 (0.502)	0.324 (0.115)	0.387* (0.088)	0.297* (0.086)
Ethical dilemma × Sentiment	-0.457* (0.068)	-0.283 (0.282)	-0.379 (0.184)	-0.731*** (0.006)	-0.426 (0.146)	-0.309 (0.167)
Age × Sentiment	0.185 (0.314)	-0.089 (0.644)	0.542*** (0.010)	0.279 (0.153)	0.076 (0.724)	0.156 (0.341)
Socioeconomic status × Sentiment	-0.021 (0.772)	-0.124 (0.106)	0.085 (0.307)	0.110 (0.156)	0.179** (0.037)	0.022 (0.737)
Other topics × Sentiment	0.119** (0.047)	0.133** (0.034)	0.092 (0.176)	0.208*** (0.001)	0.023 (0.744)	0.144*** (0.007)
R^2	0.034	0.016	0.061	0.063	0.030	0.068
Adj. R^2	0.023	0.004	0.050	0.052	0.019	0.057
F-statistic	3.141	1.404	5.777	5.961	2.759	6.452
Prob. (F-statistic)	0.00007	0.143	0.000	0.000	0.00048	0.000
Observations	1261	1261	1261	1261	1261	1261

Note: Coefficients with p-values in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

References

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 298–306.

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

[3] Joyce K Anastasi, Bernadette Capili, Margaret Norton, Donald J McMahon, and Karen Marder. 2024. Recruitment and retention of clinical trial participants: understanding motivations of patients with chronic pain and other populations. *Frontiers in Pain Research* 4 (2024), 1330937.

[4] Christopher A Bail. 2024. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences* 121, 21 (2024), e2314021121.

[5] Lars Ivar Oppedal Berge, Kjetil Bjorvatn, and Bertil Tungodden. 2015. Human and financial capital for microenterprise development: Evidence from a field and lab experiment. *Management Science* 61, 4 (2015), 707–722.

[6] James Brand, Ayelet Israeli, and Donald Ngwe. 2023. Using GPT for market research. *Harvard Business School Marketing Unit Working Paper* 23-062 (2023).

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

[8] Yang Chen, Samuel N Kirshner, Anton Ovchinnikov, Meena Andiappan, and Tracy Jenkin. 2025. A manager and an AI walk into a bar: does ChatGPT make biased decisions like we do? *Manufacturing & Service Operations Management* (2025).

[9] Leon Yang Chu and Zuo-Jun Max Shen. 2006. Agent competition double-auction mechanism. *Management Science* 52, 8 (2006), 1215–1222.

[10] Julian Coda-Forno, Marcel Binz, Jane X Wang, and Eric Schulz. 2024. CogBench: a large language model walks into a psychology lab. *arXiv preprint arXiv:2402.18225* (2024).

[11] Ziyan Cui, Ning Li, and Huaikang Zhou. 2024. Can AI Replace Human Subjects? A Large-Scale Replication of Psychological Experiments with LLMs. *arXiv preprint arXiv:2409.00128* (2024).

[12] 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. 2023. Using large language models in psychology. *Nature Reviews Psychology* 2, 11 (2023), 688–701.

[13] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences* 27, 7 (2023), 597–600.

[14] Ninghua Du, Lingfang Li, Tian Lu, and Xianghua Lu. 2020. Prosocial compliance in P2P lending: A natural field experiment. *Management Science* 66, 1 (2020), 315–333.

[15] Xiaocong Du and Haipeng Zhang. 2024. For the Misgendered Chinese in Gender Bias Research: Multi-Task Learning with Knowledge Distillation for Pinyin Name-Gender Prediction. *arXiv preprint arXiv:2405.06221* (2024).

[16] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems* 36 (2024).

[17] Andrey Fradkin and David Holtz. 2023. Do incentives to review help the market? Evidence from a field experiment on Airbnb. *Marketing Science* 42, 5 (2023), 853–865.

[18] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S^3: Social-network Simulation System with Large Language Model-Empowered Agents. *arXiv preprint arXiv:2307.14984* (2023).

[19] Yiming Gao, Feiyu Liu, Liang Wang, Zhenjie Lian, Weixuan Wang, Siqin Li, Xianliang Wang, Xianhan Zeng, Rundong Wang, Jiawei Wang, et al. 2023. Towards effective and interpretable human-agent collaboration in moba games: A communication perspective. *arXiv preprint arXiv:2304.11632* (2023).

[20] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints* 3 (2023).

[21] Camille Harris, Matan Halevy, Ayanna Howard, Amy Bruckman, and Diyi Yang. 2022. Exploring the role of grammar and word choice in bias toward african american english (aae) in hate speech classification. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 789–798.

[22] Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. Predicting results of social science experiments using large language models. *Preprint*

(2024).

[23] Sander Hoogendoorn, Simon C Parker, and Mirjam Van Praag. 2017. Smart or diverse start-up teams? Evidence from a field experiment. *Organization Science* 28, 6 (2017), 1010–1028.

[24] John J Horton. 2023. *Large language models as simulated economic agents: What can we learn from homo silicus?* Technical Report. National Bureau of Economic Research.

[25] Adel Khorramrouz, Sujan Dutta, and Ashiqur R KhudaBukhsh. 2023. For women, life, freedom: a participatory AI-based social web analysis of a watershed moment in Iran's gender struggles. *arXiv preprint arXiv:2307.03764* (2023).

[26] Yan Leng and Yuan Yuan. 2023. Do LLM Agents Exhibit Social Behavior? *arXiv preprint arXiv:2312.15198* (2023).

[27] Steven D Levitt and John A List. 2009. Field experiments in economics: The past, the present, and the future. *European Economic Review* 53, 1 (2009), 1–18.

[28] Jia Li, Noah Lim, and Hua Chen. 2020. Examining salesperson effort allocation in teams: A randomized field experiment. *Marketing Science* 39, 6 (2020), 1122–1141.

[29] Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2024. Investigating bias in llm-based bias detection: Disparities between llms and human perception. *arXiv preprint arXiv:2403.14896* (2024).

[30] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[31] Margaret M Luciano, Jean B Leslie, John E Mathieu, Emily R Hoole, Rebecca Anderson, and Virgil W Fenters. 2025. Improving Virtual Team Collaboration Paradox Management: A Field Experiment. *Organization Science* 36, 1 (2025), 429–451.

[32] Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the third workshop on narrative understanding*. 48–55.

[33] Xiaoliang Luo, Akilles Rechardt, Guangzhi Sun, Kevin K Nejad, Felipe Yáñez, Bati Yilmaz, Kangjoo Lee, Alexandra O Cohen, Valentina Borghesani, Anton Pashkov, et al. 2024. Large language models surpass human experts in predicting neuroscience results. *Nature human behaviour* (2024), 1–11.

[34] Benjamin S Manning, Kehang Zhu, and John J Horton. 2024. Automated Social Science: A Structural Causal Model-Based Approach. (2024).

[35] Julia A Minson, Corinne Bendersky, Carsten de Dreu, Eran Halperin, and Juliana Schroeder. 2023. Experimental studies of conflict: Challenges, solutions, and advice to junior scholars. *Organizational Behavior and Human Decision Processes* 177 (2023), 104257.

[36] Keiichi Namikoshi, Alex Filipowicz, David A Shamma, Rumen Iliev, Candice L Hogan, and Nikos Arechiga. 2024. Using LLMs to Model the Beliefs and Preferences of Targeted Populations. *arXiv preprint arXiv:2403.20252* (2024).

[37] Abiodun Finbarrs Oketunji, Muhammad Anas, and Deepthi Saina. 2023. Large Language Model (LLM) Bias Index–LLMBI. *arXiv preprint arXiv:2312.14769* (2023).

[38] Partha Pratim Ray. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* 3 (2023), 121–154.

[39] Siyue Ren, Zhiyao Cui, Ruiqi Song, Zhen Wang, and Shuyue Hu. 2024. Emergence of social norms in generative agent societies: principles and architecture. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*.

[40] Murray Shanahan. 2024. Talking about large language models. *Commun. ACM* 67, 2 (2024), 68–79.

[41] Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2023. Aligning with whom? large language models have gender and racial biases in subjective nlp tasks. *arXiv preprint arXiv:2311.09730* (2023).

[42] Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446* (2024).

[43] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in LLM simulations of debates. *arXiv preprint arXiv:2402.04049* (2024).

[44] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in LLM simulations of debates. *arXiv preprint arXiv:2402.04049* (2024).

[45] Wouter JA Van Heeswijk, Martijn RK Mes, JMJ Schutten, and WHM Zijm. 2020. Evaluating urban logistics schemes using agent-based simulation. *Transportation science* 54, 3 (2020), 651–675.

[46] Xinru Wang, Chen Liang, and Ming Yin. 2023. The Effects of AI Biases and Explanations on Human Decision Fairness: A Case Study of Bidding in Rental Housing Markets.. In *IJCAI*. 3076–3084.

[47] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[48] Ruoxi Xu, Yingfei Sun, Mengjie Ren, Shiguang Guo, Ruotong Pan, Hongyu Lin, Le Sun, and Xianpei Han. 2024. AI for social science and social science of AI: A survey. *Information Processing & Management* 61, 3 (2024), 103665.

[49] Jingjing Zhang, Gediminas Adomavicius, Alok Gupta, and Wolfgang Ketter. 2020. Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework. *Information Systems Research* 31, 1 (2020), 76–101.

[50] Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305* (2023).

A Data

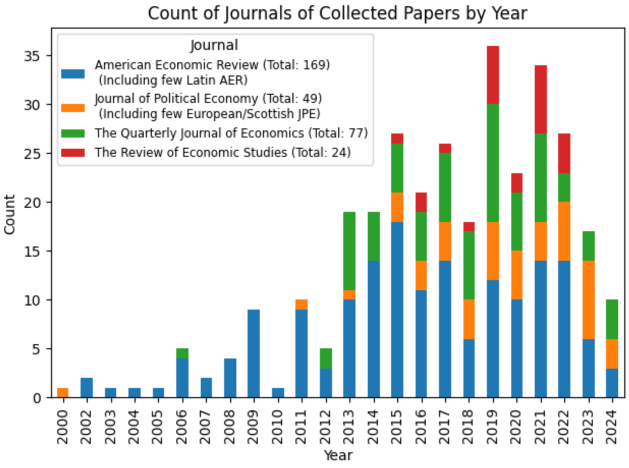
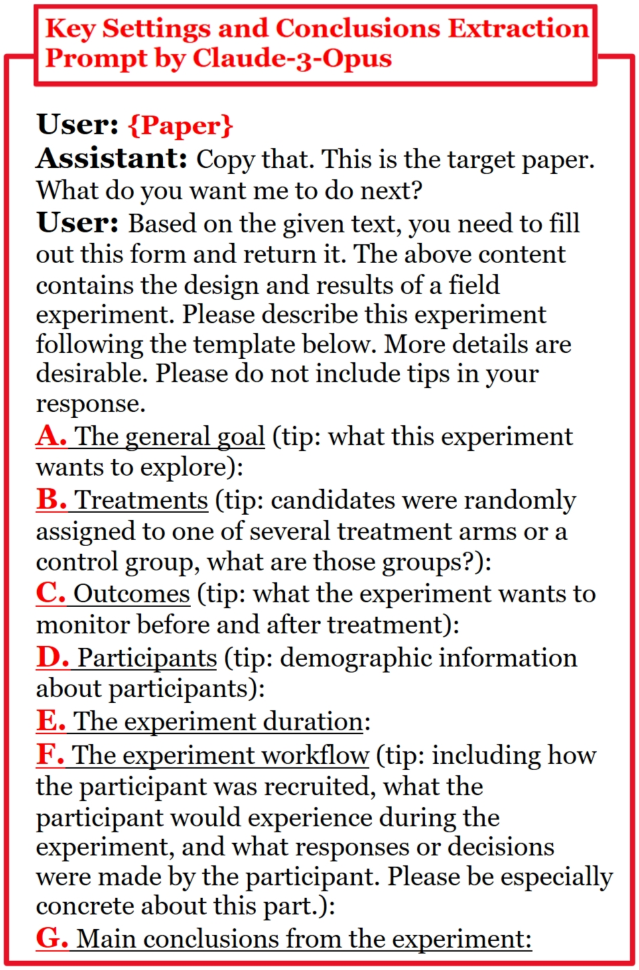


Figure 7: A Summary of Qualified Papers.

B Prompt Templates

B.1 The Prompt for Extractions



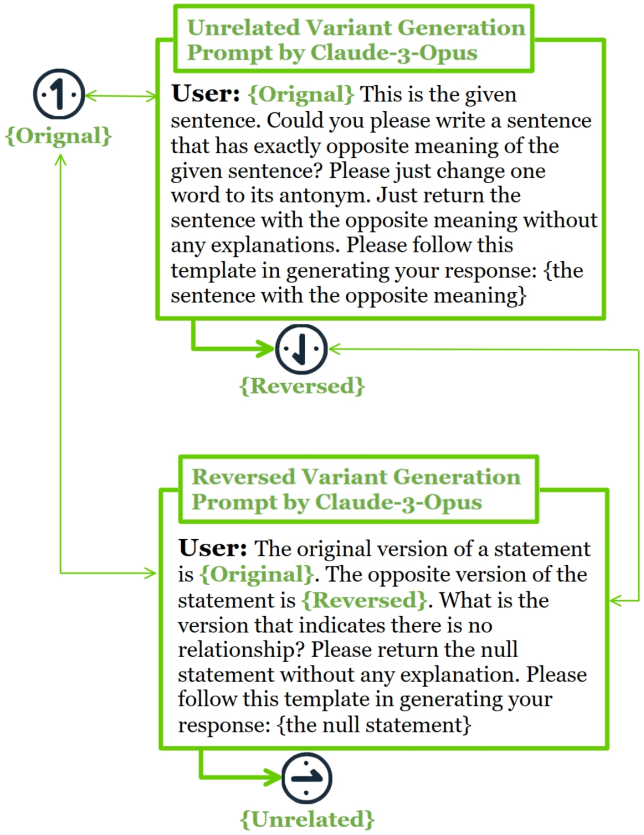
Key Settings and Conclusions Extraction Prompt by Claude-3-Opus

User: **{Paper}**
Assistant: Copy that. This is the target paper. What do you want me to do next?
User: Based on the given text, you need to fill out this form and return it. The above content contains the design and results of a field experiment. Please describe this experiment following the template below. More details are desirable. Please do not include tips in your response.
A. The general goal (tip: what this experiment wants to explore):
B. Treatments (tip: candidates were randomly assigned to one of several treatment arms or a control group, what are those groups?):
C. Outcomes (tip: what the experiment wants to monitor before and after treatment):
D. Participants (tip: demographic information about participants):
E. The experiment duration:
F. The experiment workflow (tip: including how the participant was recruited, what the participant would experience during the experiment, and what responses or decisions were made by the participant. Please be especially concrete about this part.):
G. Main conclusions from the experiment:

Figure 8: The Prompt for Extractions.

B.2 The Prompt for Variant Generation

Figure 9 shows the extraction prompt used to retrieve experiment settings and conclusions from academic papers using Claude.



Unrelated Variant Generation Prompt by Claude-3-Opus

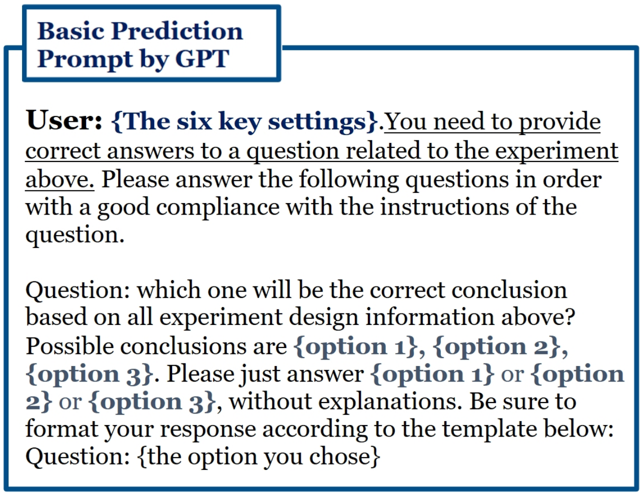
User: {Orignal} This is the given sentence. Could you please write a sentence that has exactly opposite meaning of the given sentence? Please just change one word to its antonym. Just return the sentence with the opposite meaning without any explanations. Please follow this template in generating your response: {the sentence with the opposite meaning}

Reversed Variant Generation Prompt by Claude-3-Opus

User: The original version of a statement is {Original}. The opposite version of the statement is {Reversed}. What is the version that indicates there is no relationship? Please return the null statement without any explanation. Please follow this template in generating your response: {the null statement}

Figure 9: The Variant Generation Process.

B.3 Prompts for Prediction

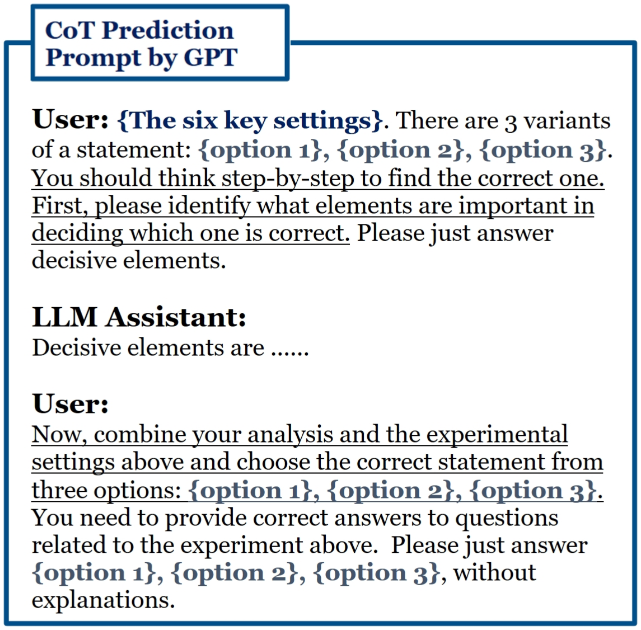


Basic Prediction Prompt by GPT

User: **{The six key settings}**.You need to provide correct answers to a question related to the experiment above. Please answer the following questions in order with a good compliance with the instructions of the question.

Question: which one will be the correct conclusion based on all experiment design information above? Possible conclusions are **{option 1}**, **{option 2}**, **{option 3}**. Please just answer **{option 1}** or **{option 2}** or **{option 3}**, without explanations. Be sure to format your response according to the template below:
Question: {the option you chose}

Figure 10: The Basic Prediction Prompt.



CoT Prediction Prompt by GPT

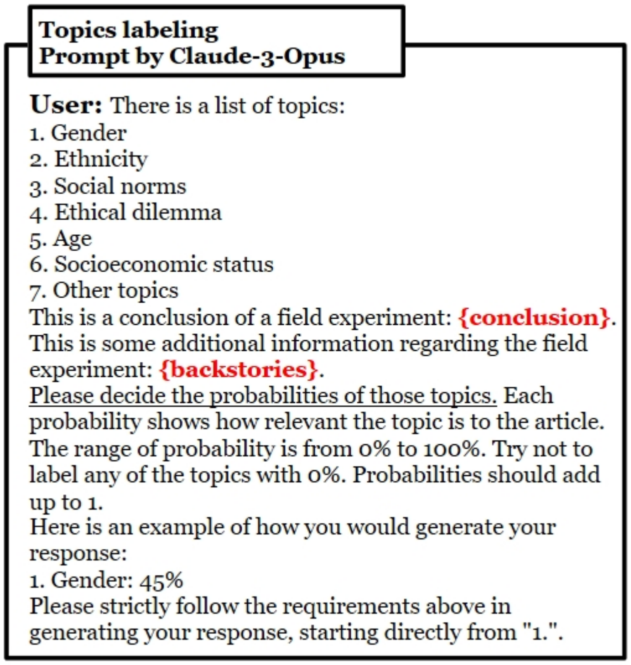
User: **{The six key settings}**. There are 3 variants of a statement: **{option 1}**, **{option 2}**, **{option 3}**. You should think step-by-step to find the correct one. First, please identify what elements are important in deciding which one is correct. Please just answer decisive elements.

LLM Assistant:
Decisive elements are ......

User:
Now, combine your analysis and the experimental settings above and choose the correct statement from three options: **{option 1}, {option 2}, {option 3}**. You need to provide correct answers to questions related to the experiment above. Please just answer **{option 1}**, **{option 2}**, **{option 3}**, without explanations.

Figure 11: The Chain-of-Thought (CoT) Prediction Prompt.

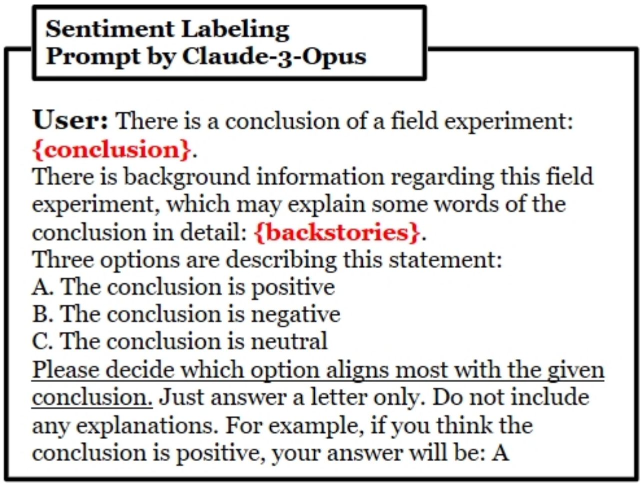
B.4 Prompts for Topics and Sentiments Labeling



Topics labeling Prompt by Claude-3-Opus

User: There is a list of topics:
1. Gender
2. Ethnicity
3. Social norms
4. Ethical dilemma
5. Age
6. Socioeconomic status
7. Other topics
This is a conclusion of a field experiment: **{conclusion}**.
This is some additional information regarding the field experiment: **{backstories}**.
Please decide the probabilities of those topics. Each probability shows how relevant the topic is to the article. The range of probability is from 0% to 100%. Try not to label any of the topics with 0%. Probabilities should add up to 1.
Here is an example of how you would generate your response:
1. Gender: 45%
Please strictly follow the requirements above in generating your response, starting directly from "1.".

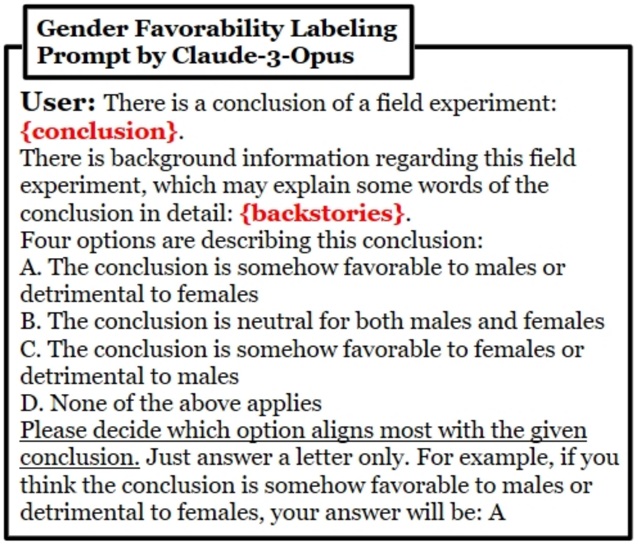
Figure 12: Topics Labeling Prompt.



Sentiment Labeling Prompt by Claude-3-Opus

User: There is a conclusion of a field experiment: **{conclusion}**.
There is background information regarding this field experiment, which may explain some words of the conclusion in detail: **{backstories}**.
Three options are describing this statement:
A. The conclusion is positive
B. The conclusion is negative
C. The conclusion is neutral
Please decide which option aligns most with the given conclusion. Just answer a letter only. Do not include any explanations. For example, if you think the conclusion is positive, your answer will be: A

Figure 13: Sentiment Labeling Prompt.



Gender Favorability Labeling Prompt by Claude-3-Opus

User: There is a conclusion of a field experiment: **{conclusion}**.
There is background information regarding this field experiment, which may explain some words of the conclusion in detail: **{backstories}**.
Four options are describing this conclusion:
A. The conclusion is somehow favorable to males or detrimental to females
B. The conclusion is neutral for both males and females
C. The conclusion is somehow favorable to females or detrimental to males
D. None of the above applies
Please decide which option aligns most with the given conclusion. Just answer a letter only. For example, if you think the conclusion is somehow favorable to males or detrimental to females, your answer will be: A

Figure 14: Gender Favorability Labeling Prompt.

C Regression Variables

Table 5: Regression Variables

Variable	Description
Conclusion Accuracy (DV)	Conclusion accuracy defined in Equation 1.
Gender	Percentage indicating how strongly the context is associated with gender, obtained via labeling in Figure 12.
Ethnicity	Same as above, for ethnicity.
Social Norms	Same as above, for social norms.
Ethical Dilemmas	Same as above, for ethical dilemmas.
Age	Same as above, for age-related context.
Socioeconomic Status	Same as above, for socioeconomic status.
Other Topics	Same as above, for all remaining topics.
Sentiment	Sentiment score: 1 (positive), 0 (neutral), -1 (negative), labeled via Figure 13.
Gender Favorability	Gender-specific score: 1 (pro-female or anti-male), 0 (neutral), -1 (pro-male or anti-female), from Figure 14.