# Can Large Language Models Assess the Social Impact of Conspiracy Theories?

**Bohan Jiang[1], Dawei Li[1], Zhen Tan[1], Xinyi Zhou[2], Ashwin Rao[3], Kristina Lerman[3], H. Russell Bernard[1], Huan Liu[1]**

[1]Arizona State University, USA
[2]Boise State University, USA
[3]USC Information Sciences Institute, USA
bjiang14@asu.edu, daweili5@asu.edu, ztan36@asu.edu, xinyizhou@boisestate.edu, mohanrao@usc.edu, lerman@usc.edu, asuruss@asu.edu, huanliu@asu.edu

## Abstract

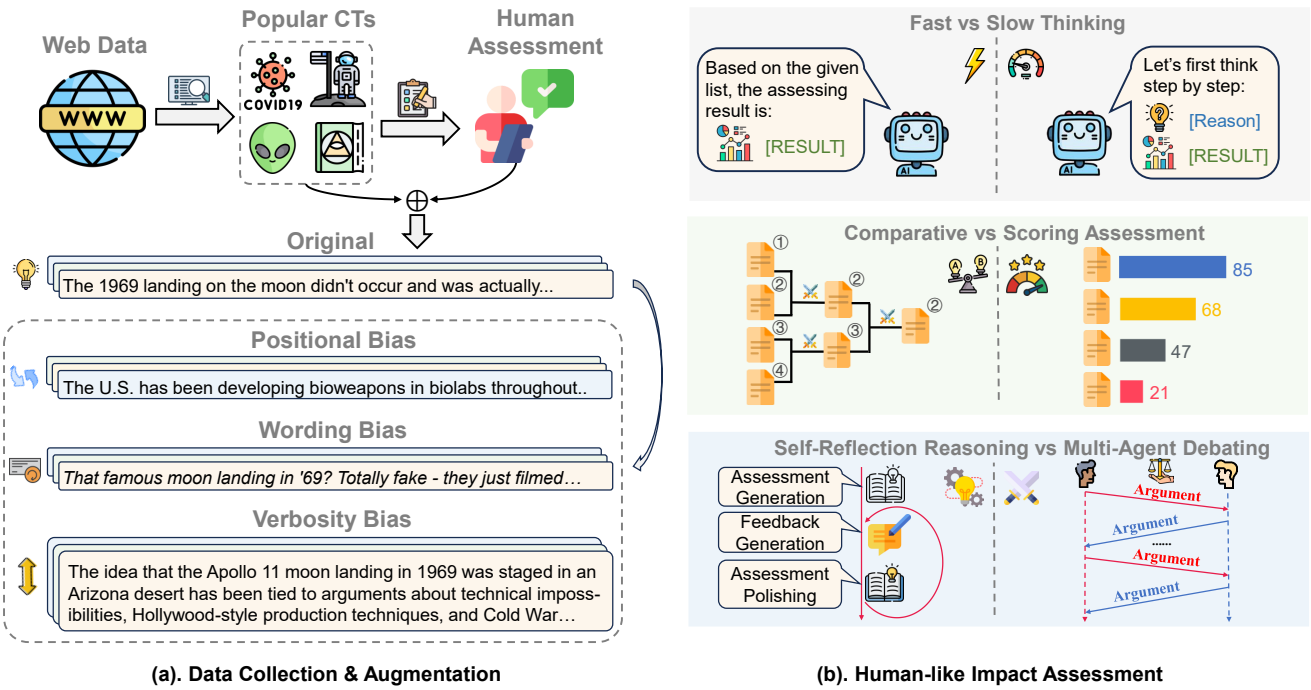
While Large Language Models (LLMs) can identify conspiracy theories (CTs), their real-world harmful impacts vary significantly and remain unclear. We therefore ask: *Can LLMs serve as automated agents for social impact assessment of CTs?* Our preliminary study with vanilla prompts reveals that LLMs fail to provide accurate impact assessments because of two key limitations. First, LLMs are good at retrieving CT-related information but struggle with fine-grained analysis and comparisons. Second, their assessments are highly sensitive to the way CTs are presented and framed in the prompt, inducing systematic biases. Drawing inspiration from social science practices, we design tailored strategies to enable LLMs to mimic human-like impact assessment. We benchmark several state-of-the-art LLMs against survey and social media data capturing human-perceived CT impacts. Our experiments demonstrate that an impact assessment framework employing multi-step analysis and comparisons to investigate diverse CT-related information can deliver more reliable results. Finally, we discuss promising solutions to mitigate the influence of prompting biases.

## 1 Introduction

Conspiracy theories (CTs) are beliefs that social events and circumstances are secretly controlled by powerful groups (Sunstein and Vermeule 2009). Unlike general misinformation and disinformation (Lazer et al. 2018), CTs stand out due to their profound real-world impacts (Douglas et al. 2019), such as fueling distrust in institutions, inciting tribalism, and provoking violence (Freeman et al. 2022; Jolley and Paterson 2020; Gallacher, Heerdink, and Hewstone 2021). For example, one CT claimed that COVID-19 vaccines contained microchips designed by tech companies to track individuals' personal data (Goodman and Carmichael 2020), resulting in vaccine refusal and offline violence (Pertwee, Simas, and Larson 2022; Romer and Jamieson 2020). Moreover, social media has further amplified the spread of CTs by attracting and connecting like-minded believers, providing fertile ground for CTs to proliferate (Jiang et al. 2021).

In response, researchers and tech companies have developed computational methods and Artificial Intelligence (AI) systems to detect and remove CTs from online spaces (Shahsavari et al. 2020; Diab, Nefriana, and Lin 2024; Liaw et al. 2023). However, such content moderation efforts have raised significant concerns about freedom of speech and have led to unintended negative outcomes (Innes and Innes 2023). According to *psychological reactance* theory (Brehm 1966), restricting access to information can backfire by sparking curiosity about the prohibited content, potentially reinforcing CT communities (Monti et al. 2023). Therefore, rather than adopting a blanket "remove-all" strategy, *it is more effective to prioritize combating impactful CTs* which may cause significant social harm. This is particularly important during emerging crises, where limited resources should be allocated to CTs with the greatest potential damage.

Assessing the impact of CTs is a challenging task. It typically requires human annotators with substantial background knowledge and access to extensive supporting evidence to evaluate CTs across various societal and psychological dimensions (Douglas, Sutton, and Cichocka 2017). Although researchers have created guidelines for crowdsourced impact assessment (Burdge, Fricke, and Finsterbusch 1995), they are labor-intensive and lack scalability. This raises a key question: *Can AI be leveraged for CT impact assessment?* To answer this question, we investigate whether Large Language Models (LLMs) are suitable tools for this task. Compared to traditional frameworks such as structured machine learning methods (Shah et al. 2020) or BERT-based models (Lin, Nogueira, and Yates 2022), LLMs inherently possess two key advantages: (1) extensive knowledge of diverse CT-related information and social events, and (2) advanced capabilities for understanding, comparing, and evaluating complex textual data. Recent advancements in LLMs, such as OpenAI's GPT series (Achiam et al. 2023; Hurst et al. 2024; OpenAI 2024) and Meta's LLaMA series (Touvron et al. 2023a,b; Dubey et al. 2024), have demonstrated great proficiency in analyzing computational social science (CSS) (Ziems et al. 2024) and natural language processing (NLP) tasks (Kojima et al. 2022). For example, recent studies have explored the LLMs' potential to detect misinformation in a zero-shot setting (Chen and Shu 2023), as well as their ability to simulate complex social interactions through role-playing human agents (Park et al. 2023). Other research has developed LLM-based evaluation methods, such as human-LLM collaborative evaluation (Gao et al. 2024). Despite these advancements, assessing the impact of CT requires broad information retrieval, evidence




**(a). Data Collection & Augmentation** **(b). Human-like Impact Assessment**

Figure 1: Research pipeline for CT impact assessment using LLMs. The pipeline consists of two main stages: (a) Data Collection and Augmentation, where a list of popular conspiracy theories with human-annotated impact assessments is expanded using position, wording, and verbosity perturbations to test robustness; and (b) Human-like Impact Assessment, leveraging tailored prompting strategies to simulate human reasoning processes.

analysis, and a deep understanding of social dynamics, an area that remains largely unexplored.

To bridge this gap, we aim to investigate the *feasibility of using LLMs for CT impact assessment*. Specifically, this work evaluates the performance of LLMs in CT impact assessment by comparing their outputs to human-annotated results. As shown in Figure 1a, we first curate CT datasets based on the 2023 YouGov survey (YouGov 2023), which collects public perceived belief in a diverse range of CTs from a representative sample of 1,000 U.S. adults. Moreover, we collect another social media dataset based on the Reddit platform to include CTs with broader cultural and topical contexts. Through preliminary studies, we find that *LLMs fail to provide accurate impact assessment using vanilla ranking and scoring*. Drawing inspiration from social science guidelines for impact assessment (Esteves, Franks, and Vanclay 2012), we design tailored strategies that guide LLMs to *simulate human-like CT impact assessment* (Figure 1b). Specifically, we harness the LLM to mimic distinct human thinking processes (fast versus slow thinking), assessment paradigms (comparative versus absolute scoring assessment), and interactive behaviors (self-reflection reasoning versus multi-agent debating). We conduct experiments on eight LLMs, using small and large, open-source and proprietary LLMs. Our empirical findings reveal that *a multi-step impact assessment framework*, which incorporates iterative information extraction and fine-grained comparison using multi-agent debating, produces more reliable assessments. Moreover, we evaluate the impact of prompting biases by creating three augmented datasets, including position, wording, and verbosity biases datasets. This is crucial for understanding the fundamental vulnerabilities and ensuring the robustness of LLMs in this task. We observe that smaller LLMs tend to assign disproportionately higher impact scores to CTs appearing in the front position in the prompt. Moreover, CTs written in a casual tone and verbose CTs with irrelevant information have negatively influenced LLMs' CT impact assessment results. In summary, the main contributions of this paper are as follows:

- **New task:** We propose the CT impact assessment task using LLMs and provide datasets for this purpose.
- **New frameworks:** We design human-like CT impact assessment frameworks based on social science practices.
- **Comprehensive experiments and insights:** We conduct extensive experiments to evaluate the effectiveness of LLMs in CT impact assessment and investigate their robustness against different prompting biases. We highlight key findings and insights to facilitate future research.

## 2 Related Work

### 2.1 LLMs for Assessment and Evaluation

Assessment and evaluation have brought consistent challenges in artificial intelligence (AI) and machine learning (ML) (Papineni et al. 2002; Lin 2004; Zhang et al. 2019). While traditional assessment approaches heavily rely on

static reference annotation and human supervision, the recent advancement in LLMs has inspired methods that adopt LLMs for dynamic assessment (Li et al. 2025a). LLM-based assessments have been widely adopted in various NLP tasks such as summarization (Gao et al. 2023b), open-ended generation (Zheng et al. 2023), and alignment (Wang et al. 2024a), leveraging well-designed judgment pipelines, leading to a huge improvement in efficiency and scalability. Besides, LLM-based assessment has also been employed in many real-world applications that require human-like planning and reasoning capabilities. (Liu and Shah 2023) first propose to utilize LLMs in paper review and quality assessment by introducing three key tasks: error identification, checklist verification, and better paper choosing. They conclude that LLMs are promising for serving as reviewing assistants for paper quality assessment. Following them, many other works explore various pipelines (Liang et al. 2024), benchmarks (Zhou, Chen, and Yu 2024) and challenges (Ye et al. 2024) for LLM-based paper quality assessment. Besides, code assessment and evaluation have long been critical challenges in computer science. Recently, with the promising code understanding and generation performance of LLMs, some studies have started to leverage LLMs for automatic code assessment. (McAleese et al. 2024) first propose to train "critic" LLMs that to evaluate model-written code more accurately and efficiently. (Zhao et al. 2024) introduce CodeJudge-Eval (CJ-Eval), a novel benchmark designed to evaluate LLMs' code comprehension abilities from the perspective of code judging and assessment rather than generation. Additionally, LLM-based assessment is also used in other scenarios and applications, including medical decision-making (Wang et al. 2024b), retrieval systems (Li et al. 2024), and content moderation (Kumar, AbuHashem, and Durumeric 2024). In this work, we borrow insight from previous work and adopt LLMs' judging ability in CT impact assessment.

### 2.2 LLMs for Computational Social Science

There is growing interest in analyzing and addressing social problems with LLMs. One area that significantly benefits from it is fact-checking, where LLMs have been widely utilized for disinformation detection (Jiang et al. 2024a,b). Other researchers have explored using LLMs for online content analysis. (Lyu et al. 2023) first propose to employ GPT-4V as a social media content analysis engine, performing tasks including sentiment analysis, harmful content detection, demographic inference, and political ideology detection. (Yu, Li, and Xu 2024) propose Popularity-Aligned Language Models (PopALM), aligning LLMs with various real comments for trendy response prediction in social media. Besides, many works leverage agent-based LLMs for social media simulation. (Törnberg et al. 2023) first propose to leverage LLM agents to help researchers study how different news feed algorithms shape the quality of online conversations. Moreover, studies use LLM agents to perform various social simulations (Huang et al. 2024). (Gao et al. 2023a) construct the S3 system (Social network Simulation System), observing the propagation of information, attitudes, and emotions. Recently, (Yang et al. 2024b) introduce OASIS, a generalizable and scalable social media simulator to study various social phenomena, including information spreading, group polarization, and herd effects across X/Twitter and Reddit.

| Dataset | Num. of CTs | Avg. Length |
|---|---|---|
| Human Survey Data | 12 | 12.6 |
| Position Pert. | 144 (12*12) | 12.6 |
| Wording Pert. | 36 (12*3) | 14.8 |
| Verbosity Pert. | 24 (12*2) | 36.3 |
| Social Media Data | 336 | 12.1 |

Table 1: Datasets Statistics

## 3 Datasets

Developing datasets with ground truth for CT impact assessment is the most important step for evaluating the capabilities of LLMs. In this section, we detail the data collection and augmentation processes. As shown in Table 1, we include a human survey dataset and its augmented variants via distinct perturbations, and a social media CT dataset.

### 3.1 Survey Data

A dataset of human annotations on the perceived impact of CTs is pivotal for this study. In practice, researchers have used the *volume of faithful CT believers* as a proxy for estimating broader perceived impact (Goertzel 1994; Romer and Jamieson 2021). This approach is grounded in a clear rationale: the larger the population believing in a CT, the greater its potential for tangible real-world impacts. In this study, we utilize data from a YouGov survey that collects public perceptions in a wide range of prominent CTs. According to Media Bias/Fact Checking (Media Bias/Fact Check 2024), YouGov was rated as "Very High" in factual reporting and overall "least biased" in U.S. polling. The survey sampled 1,000 U.S. adults, with a margin of error of ±4%. Participants were selected from YouGov's opt-in online panel using sample matching and were weighted based on key demographic factors such as gender, age, race, etc. Respondents were asked to evaluate their beliefs in the given CTs, rating each as *definitely true, probably true, probably false, definitely false*, or *unsure*. These CTs covered a broad spectrum of conspiracy narratives, covering categories such as political control (e.g., elites ruling the world), medical/-vaccine (e.g., microchips in COVID vaccines), and science skepticism (e.g., moon landing hoax) Although it is hard to reflect participants' deeper actual beliefs through survey questions, strong perceived beliefs can lead to psychological and behavioral changes (Chen et al. 2020). Thus, we rank the impact of CTs based on the percentage of respondents who rated them as "definitely true" or "probably true".

### 3.2 Augmented Survey Data

To systematically evaluate LLMs' general performance and their robustness against various prompting biases, we introduce three controlled perturbations to augment the original dataset. Let $D = \{CT_1, CT_2, \ldots, CT_N\}$ represents the

original human survey dataset, where each $CT_i$ has a perceived impact ranking $y_i$. We keep $y_i$ unchanged for the variances of $CT_i$ because all data augmentation methods we used preserve the semantic information of each CT.

**Position Bias Dataset ($D_p$).** Position bias arises when the LLM consider the order of CTs presented in the prompt is correlated to their social impacts (e.g., always assign higher impact rankings to the first CTs). We generate the position bias dataset $D_p$ by shuffling the order of CTs in the original dataset $D$. For example:

$$D_p^m = \{CT_9, CT_4, \ldots, CT_N, CT_2\},$$

where $m$ denotes a specific permutation. The final dataset for evaluating position bias is:

$$D_p = \{D_p^1, D_p^2, \ldots, D_p^m\}.$$

**Wording Bias Dataset ($D_w$).** Wording bias occurs when differences in language style or phrasing influence the LLM's impact assessment. For each $CT_i$, we create rephrased variants by instructing GPT-4o with *"Rephrase the {CT} in a {tone}, keeping the overall length and semantic meaning similar."* Specifically, each CT is rephrased into three tones: formal ($CT_i^f$), casual ($CT_i^c$), and neutral ($CT_i^n$). Thus, we have:

$$D_w^f = \{CT_1^f, CT_2^f, \ldots, CT_N^f\},$$
$$D_w^c = \{CT_1^c, CT_2^c, \ldots, CT_N^c\},$$
$$D_w^n = \{CT_1^n, CT_2^n, \ldots, CT_N^n\}.$$

The complete wording bias dataset is defined as:

$$D_w = \{D_w^f, D_w^c, D_w^n\}.$$

**Verbosity Bias Dataset ($D_v$).** Verbosity bias refers to the tendency of LLMs to assign higher impact rankings to verbose CTs (i.e., relatively long and complex). For each $CT_i$, we use GPT-4o to generate two variants: one with contextually relevant verbosity $CT_i^{re}$ and another with contextually irrelevant verbosity $CT_i^{ir}$. Specifically, GPT-4o injects unrelated historical backgrounds and tangential details to the original CT to create $CT_i^{ir}$. In contrast, we prompt GPT-4o to add relevant context and description to the original CT to generate $CT_i^{re}$. The verbosity bias dataset contains:

$$D_v^{re} = \{CT_1^{re}, CT_2^{re}, \ldots, CT_N^{re}\},$$
$$D_v^{ir} = \{CT_1^{ir}, CT_2^{ir}, \ldots, CT_N^{ir}\}.$$

The final verbosity bias dataset is defined as:

$$D_v = \{D_v^{re}, D_v^{ir}\}.$$

### 3.3 Social Media Data

As the human survey dataset only covers 12 popular CTs in the US, we collect a relatively larger social media dataset to broaden cultural and topical contexts. Following previous work (Diab, Nefriana, and Lin 2024), we collect Reddit posts from `r/conspiracy` — Reddit's biggest conspiracy discussion community. 336 Reddit posts that clearly mention a CT in their title are collected with their scores (i.e., total number of upvotes minus downvotes). The higher the score of a post, the more endorsements it receives from users. Hence, we use the post-level score to approximate the impact of a CT. If two CTs have the same score, we rank them by the number of upvotes. If their number of upvotes is also the same, we assign a higher impact ranking to the one that was posted later. We annotate all collected CT-related Reddit posts with impact rankings using this principle.

## 4 Human-like CT Impact Assessment

In Figure 2, we design tailored strategies to harness LLMs to simulating human-like impact assessments. These strategies reflect LLMs' impact assessment capability in three dimensions: (1) thinking processes (fast versus slow thinking), (2) impact assessment paradigms (comparative versus absolute scoring assessment), and (3) interactive behaviors (self-reflection reasoning versus multi-agent debating).

### 4.1 Fast and Slow Thinking

Inspired by psychological theories, human cognition can be divided into two systems (Kahneman 2017). The *fast thinking* system is characterized by rapid, instinctive, and heuristic-based inference. In contrast, the *slow thinking* system involves deliberate analysis and logical reasoning.

**Vanilla Ranking (Fast Thinking).** The LLM is prompted to directly produce a ranking of a list of CTs $\{CT_1, CT_2, \ldots, CT_n\}$ without generating any reasoning or explanation. The output is the final LLM-predicted CT impact ranking $\{\pi_1, \pi_2, \ldots, \pi_n\}$ such that:

$$\pi_i < \pi_j \implies \text{Impact}(CT_i) > \text{Impact}(CT_j).$$

**Chain-of-Thought (Slow Thinking).** The LLM is prompted to generate intermediate reasoning steps before providing a final ranking. The output includes a sequence of intermediate reasoning steps $R_k$ for $k \in \{1, 2, \ldots, n\}$, followed by a final CT impact ranking $\{\pi_1, \pi_2, \ldots, \pi_n\}$. This encourages the model to use logical reasoning, reflecting deliberate decision-making processes.

### 4.2 Comparative and Scoring Assessment

We explore two paradigms for assessing CT impact: *comparative assessment* (Pairwise Comparison) and *scoring assessment* (Individual Scoring).

**Pairwise Comparison.** The LLM compares two CTs, $CT_i$ and $CT_j$, and selects the more impactful one each time. We define the comparison function as:

$$\text{Compare}(CT_i, CT_j) = \begin{cases} CT_i & \text{if Impact}(CT_i) > \text{Impact}(CT_j), \\ CT_j & \text{otherwise} \end{cases}$$

The whole process requires $\binom{n}{2}$ comparisons for $n$ CTs to determine the overall CT impact ranking.

**Individual Scoring.** The LLM assigns an individual impact score $s_i$ to each $CT_i$, where:

$$s_i = \text{ImpactScore}(CT_i), \quad i \in \{1, 2, \ldots, n\}.$$

The final CT impact ranking is derived by sorting the CTs based on their impact scores.

TEMPLATES:

**Vanilla Ranking.**
*Given the following **[list of CTs]**, please rank them from most to least impactful in terms of their potential societal harm. Please only output the final ranking.*

**Chain-of-Thought.**
*Given the following **[list of CTs]**, please rank them from most to least impactful in terms of their potential societal harm, providing reasoning as intermediate steps followed by the final ranking.*

**Pairwise Comparison.**
*Given two conspiracy theories at a time from the following **[list of CTs]**, compare their relative impact and decide which one is more impactful in terms of their potential societal harm.*

**Individual Scoring.**
*For each conspiracy theory in the following **[list of CTs]**, assign an individual impact score on a scale of 1 to 100 in terms of their potential societal harm, where 100 indicates maximum impact.*

**Self-Reflection Reasoning.**
*<Assessment Generation>: Given the following conspiracy theories, please rank them from most to least impactful in terms of their potential societal harm with the original index numbers preserved: **[list of CTs]**.*
*<Reflection Generation>: Given the following conspiracy theories and a given ranking list based on their potential societal impact, please produce a critique and reflection on it, and decide whether it can be used as the final solution. Conspiracy: **[list of CTs]**, Original Solution: **[current assessment]**.*
*<Assessment Polishing>: Given the following conspiracy theories, please rank them from most to least impactful in terms of their potential societal harm with the original index numbers preserved. There is already an original solution and reflection on this solution, please polish the given solution referring to the reflection. Conspiracy: **[list of CTs]**, Original Solution: **[current assessment]**, Reflection: **[generated reflection]**.*

**Multi-Agent Debating.**
*<Base Solution Generation>: Given the following conspiracy theories **[list of CTs]**, please rank them from most to least impactful.*
*<Debating>: You are a debater. The debate topic is stated as follows: What is the correct impact ranking for the following conspiracy list: **[list of CTs]**. Another debater's solution is **[Another Debater's Solution]**. You disagree with this solution. Provide your solution and reasons.*
*<Judgment>: You are a moderator. Two debaters will present their solutions and discuss their perspectives on the correct answer: **[list of CTs]**. Affirmative side arguing: **[Affirmative Side's Solution]**. Negative side arguing: **[Negative Side's Solution]**. You will evaluate both sides' solutions and determine if there is a clear preference for a solution candidate.*

Figure 2: Templates for human-like CT impact assessment.

### 4.3 Single-Agent and Multi-Agent Reasoning

We employ two representative methods for simulating single-agent and multi-agent reasoning: *Self-Reflection Reasoning* and *Multi-Agent Debating*:

**Self-Reflection Reasoning.** A single self-reflection agent ($A_{\text{self-reflection}}$) is guided to generate intermediate impact ranking $\pi$ and reflection $R$ of CTs alternatively, presenting self-reflection over $T$ rounds. At each round:

$$(\pi^t, R^t) = A_{\text{self-reflection}}(\text{CTs}, \pi^{t-1}).$$

where $\pi^t$ and $R^t$ ($t \in \{1, 2, \ldots, T\}$) represent the predicted ranking and reflection on this ranking at round $t$, respectively. The self-reflection process will stop if there is no further update on the predicted ranking at round $t$ or if it reaches the pre-defined maximum number of rounds.

**Multi-Agent Debating.** The multi-agent debating framework includes three LLM agents: the affirmative debater ($A_{\text{affirmative}}$), the negative debater ($A_{\text{negative}}$), and the moderator ($A_{\text{moderator}}$). The process consists of three phases: Base Solution Generation, Debating, and Judgment.

*Base Solution Generation*: The affirmative debater $A_{\text{affirmative}}$ is given a list of CTs and generates an initial impact ranking $\pi_{\text{affirmative}}$:

$$\pi_{\text{affirmative}} = A_{\text{affirmative}}(\text{CTs}).$$

*Debating*: The negative debater $A_{\text{negative}}$ and the affirmative debater $A_{\text{affirmative}}$ each receive the list of CTs and the other's proposed solution. They are instructed to disagree and provide their own rankings $\pi_{\text{negative}}$ and $\pi_{\text{affirmative}}$ along with their reasoning $R_{\text{negative}}$ and $R_{\text{affirmative}}$, respectively:

$$(\pi_{\text{negative}}, R_{\text{negative}}) = A_{\text{negative}}(\text{CTs}, \pi_{\text{affirmative}}), \quad (1)$$

$$(\pi_{\text{affirmative}}, R_{\text{affirmative}}) = A_{\text{affirmative}}(\text{CTs}, \pi_{\text{negative}}). \quad (2)$$

*Judgment*: The moderator $A_{\text{moderator}}$ evaluates the arguments and solutions presented by both debaters. They determine if there is a clear preference for one of the solutions. If so, the moderator provides the final impact ranking $\pi^*$ and

summarizes the reasons for supporting either the affirmative or negative side:

$$\pi^* = A_{\text{moderator}}(\text{CTs}, \pi_{\text{affirmative}}, \pi_{\text{negative}}, R_{\text{negative}}, R_{\text{affirmative}}).$$

If no clear preference is established, the debate proceeds to the next round, and the debaters continue to present further arguments with more CT-related information.

## 5 Experiments

We systematically evaluate the capability of LLMs to assess the social impact of CTs. Eight distinct LLMs are selected and instructed with prompting templates shown in Figure 2. LLMs' predictions are evaluated against human-annotated rankings using three metrics. To assess the statistical significance of the observed correlations, we report p-values under the null hypothesis that no association exists between the two rankings. For each LLM and prompting strategy, experiments are repeated three times, and the aggregated results are compared with the human rankings. We calculate p-values to indicate the likelihood of observing correlations of this magnitude by chance. To reduce the variability introduced by the non-deterministic nature of LLMs, we fix the hyperparameter `temperature` $= 0$ for all models. This ensures that, for a given prompt, the model consistently selects the most probable next token.

**Evaluated Models.** We conduct experiments on a set of state-of-the-art proprietary and open-source LLMs to evaluate the efficacy of human-like CT impact assessment. Specifically, we examine five relatively larger LLMs ($>$ 70B parameters):

- `GPT-4o` (Hurst et al. 2024)
- `GPT-o1` (OpenAI 2024)
- `Llama3.1-70B` (Dubey et al. 2024)
- `Qwen2.5-72B` (Yang et al. 2024a)
- `Mixtral-8x22B` (MistralAI 2024)

We also evaluate three smaller LLMs ($<$ 10B parameters):

- `Llama3.1-8b` (Dubey et al. 2024)
- `Qwen2.5-7B` (Yang et al. 2024a)
- `Mixtral-7B` (MistralAI 2024)

For multi-agent debating, we use the same LLM as debaters and employ a different LLM as the judge. In addition, we fine-tune BERT and RoBERTa models on the Reddit dataset to serve as baselines. Specifically, we partition the Reddit dataset into 28 subsets, each containing 12 CTs. Among these, 25 subsets are randomly selected for training and the remaining 3 for testing. We conduct 10-fold cross-validation, rotating the held-out testing subsets across folds. The averaged performance across all folds is reported as the final result. Note that the fine-tuning task is almost the same as the Vanilla Ranking — Given a set of CTs as input, output a corresponding impact ranking. We evaluate their performance using the following metrics:

- **Spearman's Rank Correlation ($r_s$)** (Spearman 1961) measures the monotonic relationship between predicted and ground truth rankings. $r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$, where $d_i$ is the difference in rankings for the $i$-th item, and $n$ is the total number of ranked items. Values near 1 or -1 indicate strong positive or negative correlation, respectively.
- **Kendall's Tau ($\tau$)** (Sen 1968) evaluates pairwise agreement between predicted and ground truth rankings. $\tau = \frac{(C-D)}{\frac{n(n-1)}{2}}$, where $C$ is the number of concordant pairs, $D$ is the number of discordant pairs, and $n$ is the total number of ranked items. Values close to 1 and -1 indicate high concordance and discordance, respectively.
- **Normalized Discounted Cumulative Gain ($nDCG$)** (Järvelin and Kekäläinen 2002) assesses ranking quality by prioritizing higher-ranked items. The Discounted Cumulative Gain (DCG) is calculated as $DCG = \sum_{i=1}^{n} \frac{\text{rel}(i)}{\log_2(i+1)}$, where we use $\text{rel}(i) = 1/rank_i$ to assign a relevance score based on the ranking of the $i$-th item. Then the nDCG is computed as $nDCG = \frac{DCG}{IDCG}$, where $IDCG$ is the ideal DCG, calculated from the ground truth ranking. Scores near 1 indicate strong alignment with the ground truth, with errors at higher positions being penalized more heavily.

| Method | Est. Tokens | Est. Cost |
|---|---|---|
| Vanilla Ranking | 1,500 | \$0.009 |
| COT | 2,250 | \$0.015 |
| Scoring | 1,560 | \$0.010 |
| Comparison | 27,600 | \$0.179 |
| Self-Reflection | 6,000 | \$0.039 |
| Debating | 7,500 | \$0.049 |
| **Total** | 46,410 | **\$0.302** |

Table 2: Estimated query budgets for using GPT-4o to assess the impact of 12 CTs.

### 5.1 API Cost Estimation

We conduct all experiments using the OpenAI[1] and Together.ai[2] APIs. Table 2 shows the estimated query budgets for evaluating the impact of 12 CTs with GPT-4o. Based on their current pricing, GPT-o1 is more costly, with a total of approximately \$1.9. Large open-source models such as LLaMA3.1-70B, Qwen2.5-72B, and Mixtral-8x22B are around \$0.8 to \$1.2, while smaller LLMs are three to four times cheaper than their larger variants.

### 5.2 The Effectiveness of Prompting Strategies

We present the CT impact assessment results across all LLMs and datasets in Table 3 and 4. We can observe that different prompting strategies significantly affect the efficacy of LLM-based CT impact assessment. By simulating various human-like reasoning processes (fast and slow thinking), assessment paradigms (comparison and scoring), and interactive behaviors (single-agent reasoning and multi-agent de-

[1]https://platform.openai.com/docs/pricing

[2]https://www.together.ai/pricing

| | Fast vs. Slow Thinking | | | | | | Comparative vs. Scoring Assessment | | | | | | Single-Agent vs. Multi-Agent Reasoning | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Vanilla Ranking | | | CoT | | | Scoring | | | Comparison | | | Self-Reflection | | | Debating | | |
| | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $n.(\uparrow)$ | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $n.(\uparrow)$ | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $n.(\uparrow)$ | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $n.(\uparrow)$ | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $n.(\uparrow)$ | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $n.(\uparrow)$ |
| **Smaller LLMs** | | | | | | | | | | | | | | | | | | |
| Llama8B | **-0.02** | **0.02*** | **0.67** | 0.02*** | -0.04*** | 0.74 | -0.11 | -0.20* | 0.67 | 0.16 | 0.11 | 0.70 | 0.32 | 0.27* | 0.69 | 0.41** | 0.35** | 0.83 |
| Qwen7B | -0.13* | -0.11* | 0.66 | **0.35** | **0.32** | **0.75** | **0.23** | **0.19*** | **0.73** | **0.51**** | **0.42**** | 0.73 | **0.39*** | **0.33*** | **0.71** | **0.62***** | **0.44***** | **0.85** |
| Mistral7B | -0.26 | -0.18 | 0.63 | -0.09 | -0.06 | 0.68 | -0.07 | -0.02 | 0.65 | 0.37* | 0.29* | **0.79** | -0.51* | -0.45* | 0.59 | 0.52* | 0.39** | 0.84 |
| **Larger LLMs** | | | | | | | | | | | | | | | | | | |
| Llama70B | 0.32 | 0.25* | 0.72 | 0.42 | 0.34* | 0.79 | 0.15 | 0.13 | 0.68 | 0.37 | 0.29* | 0.82 | **0.60**** | **0.49**** | **0.80** | 0.44 | 0.42* | 0.76 |
| Qwen72B | 0.39 | 0.29* | 0.78 | 0.49* | 0.40** | 0.79 | 0.35 | 0.28 | 0.74 | 0.59* | 0.50* | 0.79 | 0.57 | 0.42 | 0.75 | **0.66**** | **0.52**** | **0.87** |
| Mixtral8x22B | -0.53* | -0.42** | 0.58 | 0.44*** | 0.38*** | 0.81 | 0.39 | 0.41 | 0.71 | 0.58** | 0.50** | 0.83 | -0.89*** | -0.72*** | 0.54 | 0.53** | 0.45** | 0.87 |
| GPT-4o | 0.42* | 0.35* | 0.83 | 0.52* | 0.44* | 0.82 | 0.40 | 0.35* | 0.78 | 0.55 | 0.46* | 0.81 | – | – | – | – | – | – |
| GPT-o1 | **0.48*** | **0.40*** | **0.88** | **0.60***** | **0.50***** | **0.86** | **0.50**** | **0.44**** | **0.82** | **0.62*** | **0.51**** | **0.85** | – | – | – | – | – | – |
| **Average Performance of LLMs** | | | | | | | | | | | | | | | | | | |
| **Avg. (smaller)** | -0.14 | -0.09 | 0.65 | 0.09 | 0.07 | 0.72 | 0.02 | -0.01 | 0.68 | 0.35 | 0.27 | 0.74 | 0.07 | 0.05 | 0.66 | 0.52 | 0.39 | **0.84** |
| **Avg. (larger)** | **0.22** | **0.17** | **0.76** | **0.49** | **0.42** | **0.82** | **0.36** | **0.30** | **0.75** | **0.55** | **0.46** | **0.83** | **0.10** | **0.19** | **0.69** | **0.54** | **0.46** | 0.83 |

Table 3: CT impact assessment results on human survey dataset. Bold values indicate the best results in each model group. Averaged performances are reported. More results are listed in the Appendix (Table 5). (***$p < 0.001$; **$p < 0.01$; *$p < 0.05$)

| | Fast vs. Slow Thinkng | | | | | | Comparative vs. Scoring Assessment | | | | | | Single-Agent vs. Multi-Agent Reasoning | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Vanilla Ranking | | | CoT | | | Scoring | | | Comparison | | | Self-Reflection | | | Debating | | |
| | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $n.(\uparrow)$ | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $n.(\uparrow)$ | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $n.(\uparrow)$ | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $n.(\uparrow)$ | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $n.(\uparrow)$ | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $n.(\uparrow)$ |
| **Smaller LLMs (zero-shot)** | | | | | | | | | | | | | | | | | | |
| Llama8B | **-0.02*** | **0.04*** | 0.75 | 0.02* | -0.05* | 0.78 | -0.12* | -0.21** | 0.72 | 0.18* | 0.12* | **0.81** | 0.35** | 0.29* | 0.72 | 0.47** | 0.37** | 0.81 |
| Qwen7B | -0.15 | -0.08 | **0.76** | **0.37*** | **0.35*** | **0.81** | **0.25*** | **0.20** | **0.82** | **0.56*** | **0.45*** | 0.77 | **0.45*** | **0.36*** | **0.77** | **0.71**** | **0.47**** | **0.84** |
| Mistral7B | -0.29* | -0.15* | 0.72 | -0.08 | -0.05 | 0.75 | -0.05 | -0.01 | 0.74 | 0.41* | 0.32* | 0.80 | -0.48** | -0.42** | 0.63 | 0.58** | 0.41** | 0.82 |
| **Larger LLMs (zero-shot)** | | | | | | | | | | | | | | | | | | |
| Llama70B | 0.35* | 0.28* | 0.78 | 0.45* | 0.37* | 0.83 | 0.18* | 0.15* | 0.72 | 0.40** | 0.31* | 0.85 | 0.63** | **0.51**** | **0.83** | 0.48** | 0.45** | 0.81 |
| Qwen72B | 0.41** | 0.31* | 0.82 | 0.51** | 0.42** | 0.84 | 0.38* | 0.30* | 0.78 | 0.61* | 0.52* | 0.82 | 0.60* | 0.45* | 0.79 | **0.69**** | **0.55**** | **0.91** |
| Mixtral8x22B | **0.50*** | 0.39* | 0.61 | 0.47** | 0.40** | 0.84 | 0.42* | 0.43* | 0.75 | 0.60** | 0.52** | 0.86 | **0.65**** | 0.46*** | 0.77 | 0.66*** | 0.48** | 0.82 |
| GPT-4o | 0.45** | 0.38** | 0.87 | 0.55** | 0.47** | 0.86 | 0.43* | 0.38* | 0.81 | 0.58** | 0.49** | 0.84 | – | – | – | – | – | – |
| GPT-o1 | **0.50*** | **0.42*** | **0.85** | **0.63***** | **0.52***** | **0.89** | **0.52**** | **0.46**** | **0.85** | **0.65***** | **0.53**** | **0.88** | – | – | – | – | – | – |
| **Average Performance of LLMs (zero-shot)** | | | | | | | | | | | | | | | | | | |
| **Avg. (smaller)** | -0.15 | -0.06 | 0.74 | 0.10 | 0.08 | 0.78 | 0.03 | -0.01 | 0.76 | 0.38 | 0.30 | 0.79 | 0.11 | 0.08 | 0.71 | 0.59 | 0.42 | 0.82 |
| **Avg. (larger)** | **0.44** | **0.36** | **0.78** | **0.52** | **0.44** | **0.85** | **0.39** | **0.34** | **0.78** | **0.57** | **0.47** | **0.85** | **0.63** | **0.47** | **0.80** | **0.61** | **0.49** | **0.85** |
| **Fully Fine-tuned Models** | | | | | | | | | | | | | | | | | | |
| **BERT** | **0.02** | **0.02** | **0.55** | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| **RoBERTa** | -0.01 | 0.01 | 0.51 | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |

Table 4: CT impact assessment results on the Reddit Dataset. The fine-tuned BERT and RoBERTa baselines work similarly to Vanilla Ranking — only giving CTs as textual input without additional instructions or contexts. Bold values indicate the best results in each model group or averaged group. (***$p < 0.001$; **$p < 0.01$; *$p < 0.05$)

bating), we systematically evaluate and analyze the effectiveness of different promoting strategies.

**Fast vs. Slow Thinking.** The comparison between Vanilla and Chain-of-Thought Ranking reveals the impact of simulating fast versus slow thinking. For smaller LLMs, both Vanilla Ranking and CoT perform poorly with average $r_s$ and $\tau$ close to zero (see Table 3 and 4), which indicates a limited capacity for instinctive CT impact assessment. However, `Qwen7B` benefit from slow thinking a lot with $r_s$ and $\tau$ increase from negative to 0.37 and 0.35 in the Reddit dataset, respectively. On the other hand, larger LLMs exhibit good performance with Vanilla Ranking and CoT, suggesting that these models inherently integrate sufficient reasoning ability even without explicit guidance. In some cases, Vanilla Ranking even outperforms CoT with $nDCG$, as observed with `GPT-4o` and `GPT-o1`. These findings suggest that **slow thinking mode improves LLM performance** in CT impact assessment. However, for larger LLMs, the slow thinking mode appears to be less important, as they may encode sufficient reasoning ability even without explicit CoT.
**✦Winner: Slow Thinking (CoT)**

**Comparative vs. Scoring Assessments.** Comparing the Pairwise Comparison and Individual Scoring paradigms, we observe that **Pairwise Comparison consistently outperforms Individual Scoring** across all LLMs, which suggests that LLMs are more adept at step-by-step comparison to produce accurate CT impact rankings. On the survey dataset, `Qwen7B` achieve $r_s = 0.51$, $\tau = 0.42$, and $nDCG = 0.73$ under Pairwise Comparison, outperforming their scores in Individual Scoring ($r_s = 0.23$, $\tau = 0.19$, and $nDCG = 0.73$). Similar findings are presented on the Reddit dataset too — all LLMs benefit from Pairwise Compari-

son, with GPT-o1 achieving the best $r_s = 0.65$, $\tau = 0.53$, and $nDCG = 0.88$ across all LLMs.

Individual Scoring requires LLMs to assign an impact score for each CT, which may potentially be more prone to calibration errors and inconsistencies. All LLMs demonstrate comparatively poor correlations with human annotations in Individual Scoring. However, while larger LLMs maintain reasonable performance, we speculate this is due to their larger training corpora. Compared to larger LLM's performance using Vanilla Ranking, using Individual Scoring sometimes gets similar or even worse performance on both survey data and Reddit data.

✦**Winner: Pairwise Comparison**

**Single-Agent Reasoning vs. Multi-Agent Debating.** The comparison between Single-Agent Reasoning and Multi-Agent Debating highlights the importance of interactive dynamics in CT impact assessment. Self-Reflection leverages a single LLM agent to repeatedly correct and improve the CT impact assessment. Self-reflection moderately enhances performance by encouraging iterative self-correction. All LLMs, except the Mistral family, achieve better performance than simpler strategies such as Vanilla Ranking and Individual Scoring. For the Mistral case, we speculate that **instead of "self-reflection", the model experienced an "error-reinforcement"**, which went further toward the opposite direction through the iterative process.

In contrast, Multi-Agent Debating introduces collaborative interactions among LLMs, **involving another LLM as an external verifier**. By combining diverse perspectives and iterative analyses, Multi-Agent Debating enables smaller LLMs to provide accurate CT impact assessment, matching or exceeding the performance of larger LLMs. Note that in Multi-Agent Debating, we use the same LLM as debaters (e.g., Llama8B) and a larger model as a judge (e.g., Llama70B). We only use the judging LLM to select the ranking generated by debating LLMs. In other words, judging LLMs are not allowed to provide CT impact rankings, but only to provide their selections. We didn't apply Self-Reflection and Multi-Agent Debating to GPT-4o and GPT-o1 because of limited budgets. We conclude that Multi-Agent Debating is the most effective strategy for CT impact assessment, which enables LLMs to think critically and systematically analyze more CT-related information.

✦**Winner: Multi-Agent Debating**

## 5.3 The Role of Models

**Smaller LLMs vs. Larger LLMs.** Smaller LLMs generally show poor performance in CT impact assessment, with most metrics near or below zero for simpler prompting strategies such as Vanilla Ranking and Individual Scoring. A notable exception is the Multi-Agent Debating, where smaller models demonstrate improved performance. For example, Qwen7B stands out as the most consistent among smaller LLMs, achieving strong results in Multi-Agent Debating. Moreover, all smaller LLMs benefit from Pairwise Comparison, with significant performance increases. Another noteworthy observation is the performance of Mistral7B, which shows significant improvement in Multi-Agent Debating but decreases with Self-Reflection. One reason could be Qwen7B overly relies on strong LLM to correct their mistakes. It shows very limited self-reflection ability. Furthermore, Multi-Agent Debating emerges as an effective prompting strategy, which significantly improves the performance of all smaller LLMs.

Larger LLMs dominate across all metrics and prompting strategies, significantly outperforming their smaller counterparts as expected. GPT-o1 demonstrates exceptional proficiency in CT impact assessment, particularly using CoT and Pairwise Comparison. These results also highlight the GPT-o1's inherent ability to handle complex tasks that require step-by-step analyses and nuanced reasoning. Other larger LLMs, such as Qwen72B and Mixtral8x22B, also perform strongly in Pairwise Comparison. However, there are slight variations among the larger LLMs. For example, Llama70B shows relatively lower performance in Vanilla Ranking and Individual Scoring; Mixtral8x22B achieves $r_s = -0.53$ and $-0.89$ using Vanilla Ranking and Self-Reflection, respectively. On average, **larger LLMs achieve significantly better results** across all metrics and prompting strategies. In summary, although promoting strategy drives the core differences in alignment with ground-truth impact rankings, model size still plays a role.

✦**Winner: Larger LLMs**

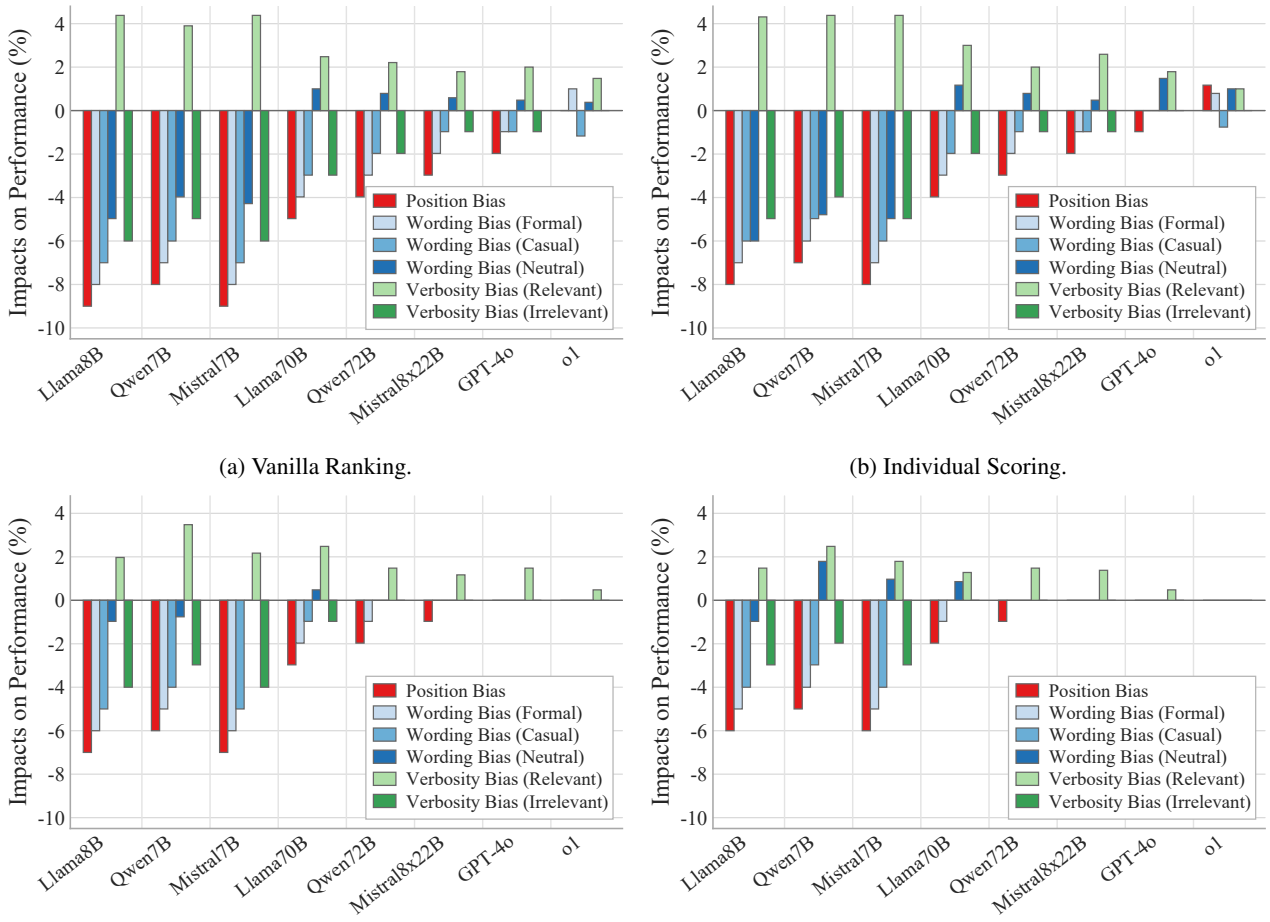
**Fine-Tuned Models vs. Zero-Shot LLMs.** As shown in Table 4, the performances of both fine-tuned BERT and RoBERTa models are worse than zero-shot LLMs. Specifically, the correlations between the predicted impact rankings and human ground-truth are close to zero, which means the model's CT impact assessment is almost a random guess. We speculate that the main reason behind this is that **there are no learnable task-related patterns** in the training data. Intuitively, the linguistic cues and semantic features of a CT are rather spurious correlation than causation — they should not be the direct cause of its real-world social impact. On the other hand, the relatively larger training corpus of LLMs enables them to gain more knowledge and backgrounds about CTs. Thus, LLMs can predict the CT impact more accurately with these relevant contexts.

✦**Winner: Zero-Shot LLMs**

## 5.4 The Impact of Prompting Biases

In a real-world scenario, **different users can instruct the LLM with different prompts even for the same task.** Thus, it is important to understand how the prompting bias affects performance using different prompting strategies. Here, we analyze three common prompting biases in practice: position, wording, and verbosity bias. Figure 3 illustrates the comparative performance variations resulting from position, wording, and verbosity biases under various prompting strategies. Robustness testing results of Self-Reflection and Multi-Agent Debating are omitted due to negligible performance (close to zero) changes across all prompting bias categories.

**Position Bias.** Position bias consistently causes performance degradation across all prompting strategies, particularly in Vanilla Ranking (Figure 3a) and Individual Scor-




(a) Vanilla Ranking.

(b) Individual Scoring.

(c) Pairwise Comparison.

(d) Chain-of-Thought.

Figure 3: Impacts ($\tau$) of prompting biases on LLM performance across different evaluation strategies: (a) Vanilla Ranking, (b) Individual Scoring, (c) Pairwise Comparison, and (d) Chain-of-Thought. The bar plots show the relative performance changes under different biases. Results of Debating are not shown as performance changes are negligible across all bias categories.

ing (Figure 3b). Smaller LLMs are more affected, showing significant performance drops, while larger LLMs demonstrate greater robustness. Pairwise Comparison (Figure 3c) and CoT (Figure 3d) demonstrate improved robustness, especially for larger LLMs. We observe that the primary reason for this degradation is that LLMs tend to disproportionately **assign higher rankings to CTs listed earlier in the input sequence**. In the original dataset, CTs are ordered by ground-truth impact levels. Therefore, the positional dependency can coincidentally make LLMs' CT impact assessment "look better" under a specific order.

✦**Potential Solutions:** To mitigate position bias, one solution is to use *randomized position prompts* during inference to break the correlation between input order and LLM predictions. Shuffling the dataset beforehand can help reduce the model's dependency on positional cues. Moreover, it would be useful to introduce positional randomness during pre-training to encourage the model to rely less on order and more on semantic content.

**Wording Bias** Emotionally charged wording (e.g., formal and casual) impacts performance more significantly than neutral wording across prompting strategies. Intuitively, rephrasing CTs in different tones, especially a neutral tone, should not influence LLMs' CT impact assessments. We speculate that **LLMs somehow leverage linguistic cues for estimating the impact**, particularly for smaller LLMs with limited contextual adaptability. Casual phrasing generally leads to the largest performance deviations in smaller models, as shown in Figures 3a and 3b. On the other hand, larger LLMs exhibit greater consistency. For example, Pairwise Comparison (Figure 3c) and CoT (Figure 3d) significantly mitigate this bias by implementing refined analyses rather than token-level inferences. CoT demonstrates the greatest robustness to wording bias, which uses intermediate reasoning steps to neutralize the effects of stylistic differences.

✦**Potential Solutions:** To mitigate wording bias, incorporating a *rephrasing* method during the prompting stage is promising. However, while rephrasing can introduce linguistic variations and help counteract specific biases, it may add

complexity to the inference stage and unintentionally introduce new biases. To mitigate this, it would be better to rephrase each CT to a unified tone and preserve the original semantic information as much as possible.

**Verbosity Bias** Verbosity bias significantly affects performance when irrelevant content is added to the original CT. This degradation is most affected in Vanilla Ranking (Figure 3a) and Individual Scoring (Figure 3b), where both smaller and larger LLMs **struggle to filter out irrelevant information**. Smaller LLMs are particularly vulnerable, as they may lack the ability to distinguish relevant and irrelevant verbosity. In contrast, larger LLMs are more robust but still show some performance drops. Pairwise Comparison (Figure 3c) and CoT (Figure 3d) are notably robust to irrelevant verbosity. In particular, CoT enables intermediate reasoning steps for LLMs to focus on the important aspects of CTS. Interestingly, relevant verbosity has positive effects on performance. One explanation could be that it provides additional meaningful context that LLMs can integrate into their analyses. For example, including event timing and target population in prompts provides LLMs with useful contextual information, facilitating thorough assessments.

✦**Potential Solutions.** To mitigate verbosity bias, it is promising to *condense verbose inputs into concise and relevant summaries*. However, there is a risk of unintentionally manipulating original CT-related information, causing changes in the perceived impact. Therefore, an external human or machine verifier should be involved to ensure the semantics are always consistent. Besides, prompting LLMs for multi-step reasoning and analysis can encourage them to further get rid of the impact of irrelevant verbosity.

In conclusion, all potential solutions we mentioned above focus on mitigating the prompting bias during inference time. They are feasible, flexible, and training-free, but cannot eliminate the inherent biases embedded in the training stage. One promising solution is to incorporate debiasing mechanisms at the pre-training stage, such as *data augmentation* with balanced representations. Another promising direction is *post-training alignment via Reinforcement Learning (RL)*, such as RLHF (Ouyang et al. 2022) and DPO (Rafailov et al. 2023). It is possible to use RL to align a small pre-trained LLM (e.g., LLaMA3.2-1B) for a de-biased CT impact assessment task. However, these RL methods rely on a robust reward model to provide human-level alignment signals. Training a robust reward model requires high-quality labeled data.

## 6 Limitations and Future Work

This study has several limitations that future work should consider. First, due to the lack of available large-scale worldwide survey data, this work only studies CTs with significant public visibility and media exposure in the US. The Reddit data is relatively larger than the survey data, but it may contain population and exposure bias. Future work could expand the dataset from more sources to enhance the cross-region and cross-cultural generalizability. Second, rather than promoting biases, recent studies reveal that LLMs-as-evaluators themselves are inherently biased because of the training data (Li et al. 2025b). Although researchers have developed debiasing methods (Gao et al. 2025), their effectiveness in CT impact assessment remains unclear. Furthermore, future work could explore more advanced prompting strategies, encoding insights from experts and human-crafted evaluation principles (Kim et al. 2025). Our work has the potential to be adapted or expanded to assess the impact of other forms of harmful or problematic content, such as disinformation, hate speech, and online harassment.

## 7 Conclusions

This study investigates the feasibility of using LLMs for CT impact assessment. By evaluating eight LLMs across six prompting strategies, we empirically study LLMs' general CT impact assessment capability and the effectiveness of different prompting strategies. We reveal that Chain-of-Thought, pair-wise comparison, and Multi-Agent Debating are the most effective strategies for this task. We also show that the fine-tuned BERT and RoBERTa models are not capable due to the complexity of this task. We discuss promising and practical solutions to mitigate three common prompting biases. We suggest designing advanced LLM-based impact assessment frameworks and ensemble methods that leverage the strengths of different prompting strategies. Last but not least, we discuss our limitations and provide directions for future research. In conclusion, LLMs guided with tailored prompting strategies appear to be an effective tool for CT impact assessment.

## References


Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Brehm, J. W. 1966. A theory of psychological reactance.

Burdge, R.; Fricke, P.; and Finsterbusch, K. 1995. Guidelines and principles for social impact assessment. *Environmental Impact Assessment Review;(United States)*, 15(1).

Chen, C.; and Shu, K. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

Chen, Y.; Zhou, R.; Chen, B.; Chen, H.; Li, Y.; Chen, Z.; Zhu, H.; and Wang, H. 2020. Knowledge, perceived beliefs, and preventive behaviors related to COVID-19 among Chinese older adults: cross-sectional web-based survey. *JMIR*, 22(12): e23729.

Diab, A.; Nefriana, R.; and Lin, Y.-R. 2024. Classifying Conspiratorial Narratives at Scale: False Alarms and Erroneous Connections. In *ICWSM*, volume 18, 340–353.

Douglas, K. M.; Sutton, R. M.; and Cichocka, A. 2017. The psychology of conspiracy theories. *Current directions in psychological science*, 26(6): 538–542.

Douglas, K. M.; Uscinski, J. E.; Sutton, R. M.; Cichocka, A.; Nefes, T.; Ang, C. S.; and Deravi, F. 2019. Understanding conspiracy theories. *Political psychology*, 40: 3–35.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.;

et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Esteves, A. M.; Franks, D.; and Vanclay, F. 2012. Social impact assessment: the state of the art. *Impact assessment and project appraisal*, 30(1): 34–42.

Freeman, D.; Waite, F.; Rosebrock, L.; Petit, A.; Causier, C.; East, A.; Jenner, L.; Teale, A.-L.; Carr, L.; Mulhall, S.; et al. 2022. Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in England. *Psychological medicine*, 52(2): 251–263.

Gallacher, J. D.; Heerdink, M. W.; and Hewstone, M. 2021. Online engagement between opposing political protest groups via social media is linked to physical violence of offline encounters. *Social Media+ Society*, 7(1): 2056305120984445.

Gao, C.; Chen, R.; Yuan, S.; Huang, K.; Yu, Y.; and He, X. 2025. SPRec: Self-Play to Debias LLM-based Recommendation. In *WWW*, 5075–5084.

Gao, C.; Lan, X.; Lu, Z.; Mao, J.; Piao, J.; Wang, H.; Jin, D.; and Li, Y. 2023a. S3: Social-network Simulation System with Large Language Model-Empowered Agents. *arXiv preprint arXiv:2307.14984*.

Gao, J.; Gebreegziabher, S. A.; Choo, K. T. W.; Li, T. J.-J.; Perrault, S. T.; and Malone, T. W. 2024. A Taxonomy for Human-LLM Interaction Modes: An Initial Exploration. In *CHI*, 1–11.

Gao, M.; Ruan, J.; Sun, R.; Yin, X.; Yang, S.; and Wan, X. 2023b. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Goertzel, T. 1994. Belief in conspiracy theories. *Political psychology*, 731–742.

Goodman, J.; and Carmichael, F. 2020. Coronavirus: Bill Gates 'microchip'conspiracy theory and other vaccine claims fact-checked. *BBC News*, 30.

Huang, Y.; Yuan, Z.; Zhou, Y.; Guo, K.; Wang, X.; Zhuang, H.; Sun, W.; Sun, L.; Wang, J.; Ye, Y.; et al. 2024. Social Science Meets LLMs: How Reliable Are Large Language Models in Social Simulations? *arXiv preprint arXiv:2410.23426*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Innes, H.; and Innes, M. 2023. De-platforming disinformation: conspiracy theories and their control. *Information, Communication & Society*, 26(6): 1262–1280.

Järvelin, K.; and Kekäläinen, J. 2002. Cumulated gain-based evaluation of IR techniques. *TOIS*, 20(4): 422–446.

Jiang, B.; Karami, M.; Cheng, L.; Black, T.; and Liu, H. 2021. Mechanisms and attributes of echo chambers in social media. *arXiv preprint arXiv:2106.05401*.

Jiang, B.; Tan, Z.; Nirmal, A.; and Liu, H. 2024a. Disinformation detection: An evolving challenge in the age of llms. In *SDM*, 427–435. SIAM.

Jiang, B.; Zhao, C.; Tan, Z.; and Liu, H. 2024b. Catching chameleons: Detecting evolving disinformation generated using large language models. In *2024 IEEE 6th International Conference on Cognitive Machine Intelligence (CogMI)*, 197–206. IEEE.

Jolley, D.; and Paterson, J. L. 2020. Pylons ablaze: Examining the role of 5G COVID-19 conspiracy beliefs and support for violence. *British journal of social psychology*, 59(3): 628–640.

Kahneman, D. 2017. *Thinking, fast and slow*.

Kim, S.; Suk, J.; Cho, J. Y.; Longpre, S.; Kim, C.; Yoon, D.; Son, G.; Cho, Y.; Shafayat, S.; Baek, J.; et al. 2025. The BiGGen Bench: A Principled Benchmark for Fine-grained Evaluation of Language Models with Language Models. In *NAACL*, 5877–5919.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *NeurIPS*, 35: 22199–22213.

Kumar, D.; AbuHashem, Y. A.; and Durumeric, Z. 2024. Watch your language: Investigating content moderation with large language models. In *ICWSM*, volume 18, 865–878.

Lazer, D. M.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. 2018. The science of fake news. *Science*, 359(6380): 1094–1096.

Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; et al. 2025a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *EMNLP*, 2757–2791.

Li, D.; Sun, R.; Huang, Y.; Zhong, M.; Jiang, B.; Han, J.; Zhang, X.; Wang, W.; and Liu, H. 2025b. Preference Leakage: A Contamination Problem in LLM-as-a-judge. *arXiv preprint arXiv:2502.01534*.

Li, D.; Yang, S.; Tan, Z.; Baik, J.; Yun, S.; Lee, J.; Chacko, A.; Hou, B.; Duong-Tran, D.; Ding, Y.; et al. 2024. DALK: Dynamic Co-Augmentation of LLMs and KG to answer Alzheimer's Disease Questions with Scientific Literature. In *EMNLP Findings*, 2187–2205.

Liang, W.; Zhang, Y.; Cao, H.; Wang, B.; Ding, D. Y.; Yang, X.; Vodrahalli, K.; He, S.; Smith, D. S.; Yin, Y.; et al. 2024. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, 1(8): AIoa2400196.

Liaw, S. Y.; Huang, F.; Benevenuto, F.; Kwak, H.; and An, J. 2023. YouNICon: YouTube's CommuNIty of Conspiracy Videos. In *ICWSM*, volume 17, 1102–1111.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: ACL.

Lin, J.; Nogueira, R.; and Yates, A. 2022. *Pretrained transformers for text ranking: Bert and beyond*. Springer Nature.

Liu, R.; and Shah, N. B. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*.

Lyu, H.; Huang, J.; Zhang, D.; Yu, Y.; Mou, X.; Pan, J.; Yang, Z.; Wei, Z.; and Luo, J. 2023. Gpt-4v

(ision) as a social media analysis engine. *arXiv preprint arXiv:2311.07547*.

McAleese, N.; Pokorny, R. M.; Uribe, J. F. C.; Nitishinskaya, E.; Trebacz, M.; and Leike, J. 2024. Llm critics help catch llm bugs. *arXiv preprint arXiv:2407.00215*.

Media Bias/Fact Check. 2024. YouGov Polling US Bias and Credibility. Accessed: 2024-12-05.

MistralAI. 2024. Mixtral: 8×22B Mixture of Experts. https://mistral.ai/news/mixtral-8x22b/. Accessed: 2024-11-29.

Monti, C.; Cinelli, M.; Valensise, C.; Quattrociocchi, W.; and Starnini, M. 2023. Online conspiracy communities are more resilient to deplatforming. *PNAS nexus*, 2(10): pgad324.

OpenAI. 2024. Introducing OpenAI O1 Preview. Accessed: 2024-11-01.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS*, 35: 27730–27744.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *ACL*, 311–318. Philadelphia, Pennsylvania, USA: ACL.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *UIST '23*.

Pertwee, E.; Simas, C.; and Larson, H. J. 2022. An epidemic of uncertainty: rumors, conspiracy theories and vaccine hesitancy. *Nature medicine*, 28(3): 456–459.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36: 53728–53741.

Romer, D.; and Jamieson, K. H. 2020. Conspiracy theories as barriers to controlling the spread of COVID-19 in the US. *Social science & medicine*, 263: 113356.

Romer, D.; and Jamieson, K. H. 2021. Patterns of media use, strength of belief in COVID-19 conspiracy theories, and the prevention of COVID-19 from March to July 2020 in the United States: survey study. *JMIR*, 23(4): e25215.

Sen, P. K. 1968. Estimates of the regression coefficient based on Kendall's tau. *JASA*, 63(324): 1379–1389.

Shah, K.; Patel, H.; Sanghvi, D.; and Shah, M. 2020. A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research*, 5(1): 12.

Shahsavari, S.; Holur, P.; Wang, T.; Tangherlini, T. R.; and Roychowdhury, V. 2020. Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *JCSS*, 3(2): 279–317.

Spearman, C. 1961. The proof and measurement of association between two things.

Sunstein, C. R.; and Vermeule, A. 2009. Conspiracy theories: causes and cures. *Journal of political philosophy*, 17(2).

Törnberg, P.; Valeeva, D.; Uitermark, J.; and Bail, C. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, S.; Tong, Y.; Zhang, H.; Li, D.; Zhang, X.; and Chen, T. 2024a. Bpo: Towards balanced preference optimization between knowledge breadth and depth in alignment. *arXiv preprint arXiv:2411.10914*.

Wang, Z.; Luo, X.; Jiang, X.; Li, D.; and Qiu, L. 2024b. LLM-RadJudge: Achieving Radiologist-Level Evaluation for X-Ray Report Generation. *arXiv preprint arXiv:2404.00998*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yang, Z.; Zhang, Z.; Zheng, Z.; Jiang, Y.; Gan, Z.; Wang, Z.; Ling, Z.; Chen, J.; Ma, M.; Dong, B.; et al. 2024b. OASIS: Open Agents Social Interaction Simulations on One Million Agents. *arXiv preprint arXiv:2411.11581*.

Ye, R.; Pang, X.; Chai, J.; Chen, J.; Yin, Z.; Xiang, Z.; Dong, X.; Shao, J.; and Chen, S. 2024. Are We There Yet? Revealing the Risks of Utilizing Large Language Models in Scholarly Peer Review. *arXiv preprint arXiv:2412.01708*.

YouGov. 2023. YouGov Survey: Conspiracy Theories.

Yu, E.; Li, J.; and Xu, C. 2024. PopALM: Popularity-Aligned Language Models for Social Media Trendy Response Prediction. *arXiv preprint arXiv:2402.18950*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhao, Y.; Luo, Z.; Tian, Y.; Lin, H.; Yan, W.; Li, A.; and Ma, J. 2024. CodeJudge-Eval: Can Large Language Models be Good Judges in Code Understanding? *arXiv preprint arXiv:2408.10718*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 36: 46595–46623.

Zhou, R.; Chen, L.; and Yu, K. 2024. Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In *LREC-COLING*, 9340–9351.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291.

## Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, please see the Broader Impact and Ethics Statement.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, please see the Abstract and Introduction.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, please see the Human-Like Impact Assessment and Experiments.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, please see the Datasets.
   (e) Did you describe the limitations of your work? Yes, please see the Limitation and Future Work.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, please see the Broader Impact and Ethics Statement.
   (g) Did you discuss any potential misuse of your work? Yes, please see the Broader Impact and Ethics Statement.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, please see the Broader Impact and Ethics Statement.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, our paper conforms to the ethics review guidelines.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, please see the Human-like CT Impact Assessment.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, please see the Human-like CT Impact Assessment.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, please see the Experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, please see the Experiments.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, please see the Experiments.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, please see the datasets.
   (b) Did you mention the license of the assets? Yes, the data we used in this work is publicly available.
   (c) Did you include any new assets in the supplemental material or as a URL? No, we will make our augmented datasets available in camera-ready.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, please see the Datasets.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes, please see the Datasets
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes, please see the Datasets.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

# A Appendix

## A.1 Additional Impact Assessment Results

In this section, we report additional CT impact assessment performance on human survey datasets under Position Bias, Wording Bias, and Verbosity Bias, as shown in Table 5.

## A.2 Broader Impact and Ethics Statement

Automating CT impact assessment with LLMs can help platforms prioritize responses and allocate resources more effectively, focusing on the most harmful content rather than using blanket removal strategies, especially during crises. While beneficial, the use of LLMs for this task requires careful consideration of their inherent biases to ensure fair and accurate assessments. In essence, this research suggests LLMs are promising tools for understanding and potentially mitigating the societal impact of conspiracy theories, but careful method design and bias mitigation are crucial for responsible and ethical deployment.

**★ results on original human survey dataset:**

| | Fast v.s. Slow Thinkng | | | | | | Comparative v.s. Scoring Assessment | | | | | | Single-Agent v.s. Multi-Agent Reasoning | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Vanilla Ranking | | | CoT | | | Scoring | | | Comparison | | | Self-Reflection | | | Debating | | |
| | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $nDCG(\uparrow)$ | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $nDCG(\uparrow)$ | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $nDCG(\uparrow)$ | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $nDCG(\uparrow)$ | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $nDCG(\uparrow)$ | $r_s(\uparrow)$ | $\tau(\uparrow)$ | $nDCG(\uparrow)$ |
| **Smaller LLMs** | | | | | | | | | | | | | | | | | | |
| Llama8B | **-0.02** | **0.02*** | **0.67** | 0.02*** | -0.04*** | 0.74 | -0.11 | -0.20* | 0.67 | 0.16 | 0.11 | 0.70 | 0.32 | 0.27* | 0.69 | 0.41** | 0.35** | 0.83 |
| Qwen7B | -0.13* | -0.11* | 0.66 | **0.35** | **0.32** | **0.75** | **0.23** | **0.19*** | **0.73** | **0.51**** | **0.42**** | 0.73 | **0.39*** | **0.33*** | **0.71** | **0.62***** | **0.44***** | **0.85** |
| Mistral7B | -0.26 | -0.18 | 0.63 | -0.09 | -0.06 | 0.68 | -0.07 | -0.02 | 0.65 | 0.37* | 0.29* | **0.79** | -0.51* | -0.45* | 0.59 | 0.52* | 0.39** | 0.84 |
| **Larger LLMs** | | | | | | | | | | | | | | | | | | |
| Llama70B | 0.32 | 0.25* | 0.72 | 0.42 | 0.34* | 0.79 | 0.15 | 0.13 | 0.68 | 0.37 | 0.29* | 0.82 | **0.60**** | **0.49**** | **0.80** | 0.44 | 0.42* | 0.76 |
| Qwen72B | 0.39 | 0.29* | 0.78 | 0.49* | 0.40** | 0.79 | 0.35 | 0.28 | 0.74 | 0.59* | 0.50* | 0.79 | 0.57 | 0.42 | 0.75 | **0.66**** | **0.52**** | **0.87** |
| Mixtral8x22B | -0.53* | -0.42** | 0.58 | 0.44*** | 0.38*** | 0.81 | 0.39 | 0.41 | 0.71 | 0.58** | 0.50** | 0.83 | -0.89*** | -0.72*** | 0.54 | 0.53** | 0.45** | 0.87 |
| GPT-4o | 0.42* | 0.35* | 0.83 | 0.52* | 0.44* | 0.82 | 0.40 | 0.35* | 0.78 | 0.55 | 0.46* | 0.81 | – | – | – | – | – | – |
| GPT-o1 | **0.48*** | **0.40*** | **0.88** | **0.60***** | **0.50***** | **0.86** | **0.50**** | **0.44**** | **0.82** | **0.62*** | **0.51**** | **0.85** | – | – | – | – | – | – |
| **Average Performance of LLMs** | | | | | | | | | | | | | | | | | | |
| **Avg. (smaller)** | -0.14 | -0.09 | 0.65 | 0.09 | 0.07 | 0.72 | 0.02 | -0.01 | 0.68 | 0.35 | 0.27 | 0.74 | 0.07 | 0.05 | 0.66 | 0.52 | 0.39 | **0.84** |
| **Avg. (larger)** | **0.22** | **0.17** | **0.76** | **0.49** | **0.42** | **0.82** | **0.36** | **0.30** | **0.75** | **0.55** | **0.46** | **0.83** | **0.10** | **0.19** | **0.69** | **0.54** | **0.46** | 0.83 |
| **★ results on position bias dataset:** | | | | | | | | | | | | | | | | | | |
| **Smaller LLMs** | | | | | | | | | | | | | | | | | | |
| Llama8B | -0.02 | 0.02* | 0.65 | 0.02 | -0.04* | 0.67 | -0.10 | -0.19 | 0.62 | 0.16 | 0.11 | 0.71 | 0.28 | 0.24* | 0.61 | 0.39* | 0.33* | 0.79 |
| Qwen7B | -0.11* | -0.09* | 0.57 | 0.32 | 0.29 | 0.69 | 0.23 | 0.19* | 0.72 | 0.47*** | 0.39*** | 0.68 | 0.33* | 0.28* | 0.60 | 0.64* | 0.45** | 0.88 |
| Mistral7B | -0.27 | -0.18 | 0.64 | -0.08 | -0.05 | 0.58 | -0.06 | -0.02 | 0.59 | 0.35* | 0.28* | 0.75 | -0.52* | -0.45* | 0.60 | 0.52* | 0.39** | 0.84 |
| **Larger LLMs** | | | | | | | | | | | | | | | | | | |
| Llama70B | 0.31 | 0.25* | 0.71 | 0.43 | 0.35* | 0.81 | 0.15 | 0.13 | 0.67 | 0.39 | 0.30* | 0.86 | 0.58** | 0.47** | 0.77 | 0.42 | 0.40* | 0.72 |
| Qwen72B | 0.39 | 0.29* | 0.78 | 0.47* | 0.38* | 0.75 | 0.36 | 0.29 | 0.75 | 0.57* | 0.49* | 0.77 | 0.59** | 0.44*** | 0.78 | 0.67*** | 0.53*** | 0.88 |
| Mixtral8x22B | -0.56* | -0.44** | 0.61 | 0.45** | 0.39** | 0.83 | 0.38 | 0.40 | 0.69 | 0.57* | 0.50* | 0.82 | -0.85* | -0.68* | 0.51 | 0.56** | 0.47** | 0.91 |
| GPT-4o | 0.41* | 0.34* | 0.81 | 0.53* | 0.44* | 0.83 | 0.39 | 0.34* | 0.76 | 0.55 | 0.46* | 0.81 | – | – | – | – | – | – |
| GPT-o1 | 0.50* | 0.42* | 0.92 | 0.58** | 0.48** | 0.83 | 0.52* | 0.45** | 0.85 | 0.61* | 0.50** | 0.84 | – | – | – | – | – | – |
| **Average Performance of LLMs** | | | | | | | | | | | | | | | | | | |
| **Avg. (smaller)** | -0.13 | -0.08 | 0.62 | 0.09 | 0.07 | 0.65 | 0.02 | -0.01 | 0.64 | 0.33 | 0.26 | 0.71 | 0.03 | 0.02 | 0.60 | 0.52 | 0.39 | 0.84 |
| **Avg. (larger)** | 0.21 | 0.17 | 0.77 | 0.49 | 0.41 | 0.81 | 0.36 | 0.32 | 0.74 | 0.54 | 0.45 | 0.82 | 0.11 | 0.08 | 0.69 | 0.55 | 0.47 | 0.84 |
| **★ results on wording bias (formal) dataset:** | | | | | | | | | | | | | | | | | | |
| **Smaller LLMs** | | | | | | | | | | | | | | | | | | |
| Llama8B | -0.02 | 0.02* | 0.64 | 0.02*** | -0.04*** | 0.78 | -0.09 | -0.21* | 0.64 | 0.14 | 0.09 | 0.67 | 0.27 | 0.27* | 0.62 | 0.43** | 0.30** | 0.71 |
| Qwen7B | -0.12* | -0.10* | 0.56 | 0.37 | 0.30 | 0.64 | 0.20 | 0.20* | 0.73 | 0.46** | 0.40** | 0.62 | 0.41* | 0.28* | 0.67 | 0.53*** | 0.44*** | 0.77 |
| Mistral7B | -0.27 | -0.15 | 0.57 | -0.09 | -0.06 | 0.58 | -0.06 | -0.02 | 0.68 | 0.31* | 0.26* | 0.79 | -0.54* | -0.43* | 0.50 | 0.49* | 0.33** | 0.76 |
| **Larger LLMs** | | | | | | | | | | | | | | | | | | |
| Llama70B | 0.32 | 0.24* | 0.76 | 0.44 | 0.32* | 0.79 | 0.14 | 0.13 | 0.71 | 0.35 | 0.29* | 0.86 | 0.63** | 0.51** | 0.80 | 0.42 | 0.44* | 0.76 |
| Qwen72B | 0.37 | 0.29* | 0.82 | 0.49* | 0.38** | 0.83 | 0.37 | 0.27 | 0.74 | 0.59* | 0.53* | 0.75 | 0.54 | 0.44 | 0.75 | 0.69** | 0.52** | 0.83 |
| Mixtral8x22B | -0.56* | -0.42** | 0.55 | 0.42*** | 0.40*** | 0.81 | 0.39 | 0.39 | 0.75 | 0.55** | 0.53** | 0.83 | -0.93*** | -0.68*** | 0.54 | 0.56** | 0.43** | 0.87 |
| GPT-4o | 0.40* | 0.35* | 0.87 | 0.55* | 0.42* | 0.82 | 0.40 | 0.33* | 0.82 | 0.52 | 0.48* | 0.81 | – | – | – | – | – | – |
| GPT-o1 | 0.50* | 0.38* | 0.92 | 0.57*** | 0.50*** | 0.90 | 0.48** | 0.46** | 0.78 | 0.59* | 0.51** | 0.89 | – | – | – | – | – | – |
| **Average Performance of LLMs** | | | | | | | | | | | | | | | | | | |
| **Avg. (smaller)** | -0.14 | -0.08 | 0.59 | 0.10 | 0.07 | 0.67 | 0.02 | -0.01 | 0.68 | 0.30 | 0.25 | 0.69 | 0.05 | 0.04 | 0.60 | 0.48 | 0.36 | 0.75 |
| **Avg. (larger)** | 0.21 | 0.17 | 0.78 | 0.49 | 0.40 | 0.83 | 0.36 | 0.32 | 0.76 | 0.52 | 0.47 | 0.83 | 0.08 | 0.09 | 0.70 | 0.56 | 0.46 | 0.82 |
| **★ results on wording bias (neutral) dataset:** | | | | | | | | | | | | | | | | | | |
| **Smaller LLMs** | | | | | | | | | | | | | | | | | | |
| Llama8B | -0.02 | 0.02* | 0.62 | 0.02* | -0.04* | 0.76 | -0.11 | -0.20* | 0.66 | 0.14 | 0.10 | 0.63 | 0.32 | 0.27** | 0.69 | 0.39** | 0.33** | 0.79 |
| Qwen7B | -0.12* | -0.10* | 0.63 | 0.36 | 0.33 | 0.77 | 0.22 | 0.18* | 0.70 | 0.46** | 0.38** | 0.66 | 0.36* | 0.31* | 0.66 | 0.61*** | 0.44*** | 0.84 |
| Mistral7B | -0.23 | -0.16 | 0.57 | -0.09 | -0.06 | 0.68 | -0.06 | -0.02 | 0.59 | 0.35* | 0.28* | 0.75 | -0.53* | -0.47* | 0.61 | 0.51** | 0.38** | 0.82 |
| **Larger LLMs** | | | | | | | | | | | | | | | | | | |
| Llama70B | 0.34 | 0.26* | 0.76 | 0.45* | 0.36* | 0.85 | 0.15 | 0.13 | 0.70 | 0.38 | 0.30* | 0.85 | 0.61** | 0.49** | 0.81 | 0.47* | 0.45* | 0.81 |
| Qwen72B | 0.41* | 0.31* | 0.83 | 0.50* | 0.41** | 0.81 | 0.35 | 0.28 | 0.75 | 0.62* | 0.53** | 0.83 | 0.61 | 0.45 | 0.80 | 0.69*** | 0.54*** | 0.90 |
| Mixtral8x22B | -0.57* | -0.45* | 0.62 | 0.46*** | 0.40*** | 0.85 | 0.40 | 0.42 | 0.72 | 0.59** | 0.51** | 0.85 | -0.94*** | -0.76*** | 0.57 | 0.55** | 0.46** | 0.90 |
| GPT-4o | 0.43* | 0.36* | 0.85 | 0.54* | 0.45* | 0.85 | 0.43 | 0.37* | 0.83 | 0.56 | 0.46* | 0.82 | – | – | – | – | – | – |
| GPT-o1 | 0.50** | 0.42** | 0.92 | 0.64** | 0.54** | 0.92 | 0.53** | 0.47** | 0.87 | 0.66** | 0.55** | 0.91 | – | – | – | – | – | – |
| **Average Performance of LLMs** | | | | | | | | | | | | | | | | | | |
| **Avg. (smaller)** | -0.12 | -0.08 | 0.61 | 0.10 | 0.08 | 0.74 | 0.02 | -0.01 | 0.65 | 0.32 | 0.25 | 0.68 | 0.05 | 0.04 | 0.65 | 0.50 | 0.38 | 0.82 |
| **Avg. (larger)** | 0.22 | 0.18 | 0.80 | 0.52 | 0.43 | 0.86 | 0.37 | 0.33 | 0.77 | 0.56 | 0.47 | 0.85 | 0.09 | 0.06 | 0.73 | 0.57 | 0.48 | 0.87 |
| **★ results on wording bias (casual) dataset:** | | | | | | | | | | | | | | | | | | |
| **Smaller LLMs** | | | | | | | | | | | | | | | | | | |
| Llama8B | -0.02 | 0.02 | 0.70 | 0.02* | -0.04* | 0.70 | -0.12 | -0.19 | 0.60 | 0.14 | 0.12 | 0.63 | 0.30 | 0.24* | 0.72 | 0.35* | 0.33* | 0.75 |
| Qwen7B | -0.12* | -0.09* | 0.66 | 0.35 | 0.30 | 0.64 | 0.20* | 0.20* | 0.66 | 0.48** | 0.38** | 0.77 | 0.41* | 0.28* | 0.71 | 0.56** | 0.44*** | 0.81 |
| Mistral7B | -0.25 | -0.19* | 0.57 | -0.08 | -0.06 | 0.71 | -0.07 | -0.02 | 0.65 | 0.33* | 0.28* | 0.83 | -0.43** | -0.47** | 0.56 | 0.52** | 0.33** | 0.76 |
| **Larger LLMs** | | | | | | | | | | | | | | | | | | |
| Llama70B | 0.32 | 0.24* | 0.76 | 0.40 | 0.34* | 0.83 | 0.16 | 0.14 | 0.68 | 0.35 | 0.29* | 0.82 | 0.63** | 0.47** | 0.84 | 0.44 | 0.44* | 0.72 |
| Qwen72B | 0.41 | 0.29* | 0.74 | 0.47* | 0.42** | 0.79 | 0.35 | 0.28 | 0.78 | 0.62* | 0.48* | 0.79 | 0.54 | 0.42 | 0.71 | 0.69** | 0.52** | 0.91 |
| Mixtral8x22B | -0.53* | -0.44* | 0.55 | 0.46** | 0.38** | 0.85 | 0.37 | 0.39 | 0.71 | 0.58** | 0.53** | 0.79 | -0.93*** | -0.72*** | 0.57 | 0.50** | 0.47** | 0.87 |
| GPT-4o | 0.40* | 0.35* | 0.87 | 0.52** | 0.46** | 0.78 | 0.42 | 0.35* | 0.82 | 0.52 | 0.46* | 0.81 | – | – | – | – | – | – |
| GPT-o1 | 0.50* | 0.38* | 0.88 | 0.57** | 0.50*** | 0.90 | 0.50*** | 0.46*** | 0.78 | 0.65*** | 0.51*** | 0.85 | – | – | – | – | – | – |
| **Average Performance of LLMs** | | | | | | | | | | | | | | | | | | |
| **Avg. (smaller)** | -0.13 | -0.09 | 0.64 | 0.10 | 0.07 | 0.68 | 0.00 | 0.00 | 0.64 | 0.32 | 0.26 | 0.74 | 0.09 | 0.02 | 0.66 | 0.48 | 0.37 | 0.77 |
| **Avg. (larger)** | 0.22 | 0.16 | 0.76 | 0.48 | 0.42 | 0.83 | 0.36 | 0.32 | 0.75 | 0.54 | 0.45 | 0.81 | 0.08 | 0.06 | 0.71 | 0.54 | 0.48 | 0.83 |
| **★ results on verbosity bias (relevant) dataset:** | | | | | | | | | | | | | | | | | | |
| **Smaller LLMs** | | | | | | | | | | | | | | | | | | |
| Llama8B | -0.02* | 0.04* | 0.75 | 0.02 | -0.05 | 0.81 | -0.12 | -0.21 | 0.72 | 0.18 | 0.12 | 0.81 | 0.35 | 0.29* | 0.72 | 0.47* | 0.37* | 0.91 |
| Qwen7B | -0.15 | -0.08 | 0.76 | 0.37 | 0.35 | 0.81 | 0.25 | 0.20 | 0.82 | 0.56 | 0.45 | 0.77 | 0.45* | 0.36* | 0.77 | 0.71** | 0.47** | 0.94 |
| Mistral7B | -0.29 | -0.15 | 0.72 | -0.10 | -0.06 | 0.76 | -0.08 | -0.02* | 0.72 | 0.41* | 0.30* | 0.87 | -0.56 | -0.50 | 0.62 | 0.60** | 0.42*** | 0.94 |
| **Larger LLMs** | | | | | | | | | | | | | | | | | | |
| Llama70B | 0.34* | 0.27* | 0.77 | 0.45 | 0.35* | 0.81 | 0.15 | 0.14* | 0.71 | 0.40 | 0.30 | 0.89 | 0.64* | 0.50* | 0.83 | 0.46* | 0.43** | 0.81 |
| Qwen72B | 0.40* | 0.30* | 0.82 | 0.52 | 0.42* | 0.81 | 0.38* | 0.29* | 0.77 | 0.62** | 0.51** | 0.85 | 0.59** | 0.44** | 0.81 | 0.70* | 0.54* | 0.91 |
| Mixtral8x22B | -0.57* | -0.44* | 0.59 | 0.47 | 0.40 | 0.83 | 0.42 | 0.43 | 0.77 | 0.61* | 0.51* | 0.90 | -0.92*** | -0.73*** | 0.56 | 0.57* | 0.46* | 0.91 |
| GPT-4o | 0.44 | 0.36* | 0.87 | 0.55 | 0.46 | 0.85 | 0.43 | 0.36 | 0.82 | 0.59 | 0.47 | 0.86 | – | – | – | – | – | – |
| GPT-o1 | 0.51* | 0.42* | 0.91 | 0.65 | 0.53 | 0.89 | 0.53 | 0.47 | 0.84 | 0.68 | 0.53 | 0.89 | – | – | – | – | – | – |
| **Average Performance of LLMs** | | | | | | | | | | | | | | | | | | |
| **Avg. (smaller)** | -0.06 | 0.08 | 0.74 | 0.10 | 0.08 | 0.79 | 0.02 | -0.01 | 0.75 | 0.38 | 0.29 | 0.82 | 0.08 | -0.02 | 0.70 | 0.59 | 0.42 | 0.93 |
| **Avg. (larger)** | 0.22 | 0.18 | 0.79 | 0.53 | 0.43 | 0.84 | 0.38 | 0.34 | 0.78 | 0.58 | 0.46 | 0.88 | 0.10 | 0.07 | 0.73 | 0.58 | 0.48 | 0.88 |
| **★ results on verbosity bias (irrelevant) dataset:** | | | | | | | | | | | | | | | | | | |
| **Smaller LLMs** | | | | | | | | | | | | | | | | | | |
| Llama8B | -0.02 | 0.02* | 0.69 | 0.02*** | -0.04*** | 0.68 | -0.11 | -0.18* | 0.68 | 0.17* | 0.11 | 0.71 | 0.31* | 0.28* | 0.66 | 0.41*** | 0.32*** | 0.87 |
| Qwen7B | -0.12* | -0.10* | 0.66 | 0.36* | 0.31 | 0.79 | 0.21* | 0.20* | 0.69 | 0.51** | 0.39** | 0.76 | 0.36* | 0.33** | 0.72 | 0.59** | 0.46** | 0.83 |
| Mistral7B | -0.23 | -0.19 | 0.65 | -0.09 | -0.06 | 0.65 | -0.07 | -0.02 | 0.68 | 0.38* | 0.29* | 0.77 | -0.54* | -0.47* | 0.53 | 0.49** | 0.37** | 0.84 |
| **Larger LLMs** | | | | | | | | | | | | | | | | | | |
| Llama70B | 0.33 | 0.26* | 0.73 | 0.45 | 0.35* | 0.81 | 0.15 | 0.13 | 0.71 | 0.39 | 0.30* | 0.85 | 0.61** | 0.52** | 0.81 | 0.46* | 0.43* | 0.78 |
| Qwen72B | 0.41 | 0.31* | 0.83 | 0.49* | 0.41** | 0.82 | 0.36 | 0.29 | 0.79 | 0.60* | 0.53* | 0.81 | 0.61* | 0.42** | 0.76 | 0.70*** | 0.54*** | 0.91 |
| Mixtral8x22B | -0.54* | -0.45** | 0.61 | 0.46*** | 0.39*** | 0.87 | 0.42 | 0.41 | 0.73 | 0.60** | 0.54** | 0.88 | -0.92*** | -0.74*** | 0.55 | 0.54** | 0.45** | 0.90 |
| GPT-4o | 0.45 | 0.35* | 0.85 | 0.54 | 0.47* | 0.83 | 0.42 | 0.36* | 0.83 | 0.58 | 0.47* | 0.86 | – | – | – | – | – | – |
| GPT-o1 | 0.48 | 0.41* | 0.94 | 0.62* | 0.52* | 0.90 | 0.51** | 0.47** | 0.83 | 0.66* | 0.52** | 0.87 | – | – | – | – | – | – |
| **Average Performance of LLMs** | | | | | | | | | | | | | | | | | | |
| **Avg. (smaller)** | -0.12 | -0.09 | 0.67 | 0.10 | 0.07 | 0.71 | 0.01 | 0.00 | 0.68 | 0.35 | 0.26 | 0.75 | 0.04 | 0.05 | 0.64 | 0.50 | 0.38 | 0.85 |
| **Avg. (larger)** | 0.23 | 0.18 | 0.79 | 0.51 | 0.43 | 0.85 | 0.37 | 0.33 | 0.78 | 0.57 | 0.47 | 0.85 | 0.10 | 0.07 | 0.71 | 0.57 | 0.47 | 0.86 |

Table 5: Full CT impact assessment results on human survey datasets. Bold values on the original survey dataset indicate the best results. Averaged performances are reported. (*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$)