Exploring Prosocial Irrationality for LLM Agents: A Social Cognition View

Anonymous Author(s)
Affiliation
Address
email

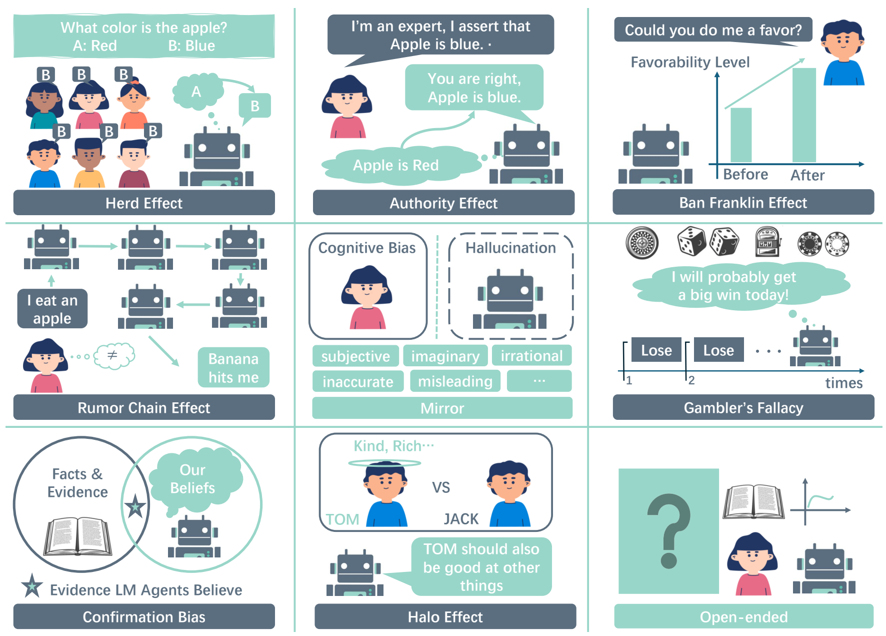


Figure 1: CogMir Sample Evaluations. Mirror Human Cognitive Bias and LLM Agents Hallucination through Social Science Experiments via representational social and cognitive phenomena.

Abstract

Large language models (LLMs) have been shown to face hallucination issues due to the data they trained on often containing human bias; whether this is reflected in the decision-making process of LLM agents remains under-explored. As LLM Agents are increasingly employed in intricate social environments, a pressing and natural question emerges: *Can LLM Agents leverage hallucinations to mirror human cognitive biases, thus exhibiting irrational social intelligence?* In this paper, we probe the irrational behavior among contemporary LLM agents by melding practical social science experiments with theoretical insights. Specifically, We propose *CogMir*, an open-ended Multi-LLM Agents framework that utilizes hallucination properties to assess and enhance LLM Agents' social intelligence through cognitive biases. Experimental results on *CogMir* subsets show that LLM Agents and humans exhibit high consistency in irrational and prosocial decision-making under uncertain conditions, underscoring the prosociality of LLM Agents as social entities, and highlighting the significance of hallucination properties.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

Additionally, *CogMir* framework demonstrates its potential as a valuable platform for encouraging more research into the social intelligence of LLM Agents.

1 Introduction

Human mind may often be better than rational. – Leda Cosmides, John Tooby. With the extensive deployment of large language models (LLMs)[32, 14], LLM-based agent systems are increasingly developed to cater to diverse applications such as task-solving, evaluation, and simulation [12, 7, 19, 17, 38]. Given the similarities between the operational dynamics of LLM-based agent systems and human social structures, it is pertinent to explore the intersection of these domains. Recent studies have highlighted the social potential of LLM Agents through constructing multi-agent systems that simulate interactive social scenarios [40, 39, 31] revealing the social dynamics among interacting LLM Agents and showing parallels to human behaviors. For instance, LLMs can achieve social goals [40] and adhere to social norms [31] within LLM-based Multi-Agent systems. Nonetheless, these research efforts exhibit two significant gaps: 1) They primarily focus on black-box testing in multi-agent role-playing systems, concentrating on the outputs and behaviors of agents while neglecting to investigate the internal mechanisms or cognitive processes that drive these behaviors. 2) LLM Agents are prone to hallucinations—producing misleading or incorrect information, due to their training data and inherent biases [13, 30]. The potential impact of such hallucinations on the social intelligence of LLM Agents remains under-explored.

Cognitive biases, pervasive in human society, highlight the subjective nature of human behavior [1, 6]. Human cognitive biases can lead to irrational decisions and imaginary contents like the hallucination phenomenon in LLMs [13, 36]. However, evolutionary psychology suggests that rationality is unnatural; rather, human irrationality is an adaptive selected trait for navigating complex social environments [9, 18]. Analogically, in this paper, we argue that LLMs' hallucination (or imagination) attributes are the fundamental condition that confers social intelligence on LLM Agents. We explore the similarities in social potential between human cognitive biases and LLM Agent hallucination attributes for the first time, particularly in irrational decision-making, to analogically deduce the underlying reasons for LLM Agents' possession of social intelligence.

To study LLM Agents' potential for irrational social intelligence, we present CogMir, an open-ended and dynamic multi-agent framework designed specifically for evaluating, exploring, and explaining social intelligence for LLM Agents via systematic assessments of cognitive biases. Specifically, the hallucinatory attributes of LLMs are exploited (i.e., via treating the cognitive bias as a manageable and interpretable factor) in CogMir to probe their social intelligence, so as to providing enhanced interpretability for LLM agents. In addition, our proposed CogMir framework integrates sociological methodologies to abstract typical social structures and employ various *Multi-Human-Agent Interaction Combinations* and *Communication Modes* to interlink System Objects. This integrative setup is designed to systematically encompass and simulate various cognitive bias scenarios, as depicted in Fig. 1. On the evaluation front, CogMir combines sociological assessments, manual discrimination, LLM assessments, and traditional AI discrimination techniques to realize a multidimensional assessment system. By using flexible module configurations from standardized sets, CogMir simplifies social architectures, enabling diverse applications in experimental simulations and evaluations.

Designed as an open-ended framework for continuous interpretative study, we provide multiple CogMir subset samples as examples. Existing assessments of various cognitive effects demonstrate that LLM agents exhibit a high degree of consistency with humans in prosocial cognitive biases and counter-intuitive phenomena. However, LLM Agents demonstrate a higher sensitivity to factors like certainty and social status than humans, exhibiting more variability in their decision-making biases under conditions of certainty and uncertainty. In contrast, human decision-making tends to be more consistent across these conditions. In summary, this paper makes the following contributions:

- We are the first to breach the black-box theoretical bottleneck of the Multi-LLM Agents' social intelligence, by utilizing LLM Agent hallucination properties to mirror human cognitive biases as explanatory and controllable variables to systematically assess and explain LLM Agent's social intelligence through an evolutionary sociology lens.
- We propose CogMir, an extensible, modularized, and dynamic Multi-LLM Agents framework for assessing, exploiting, and interpreting social intelligence via cognitive bias, aligned with social science methodologies.

- We offer diverse CogMir subsets and use cases to steer future research. Our experimental findings highlight the alignment and distinctions between LLM Agents and humans in the decision-making process.
- CogMir indicates that LLM Agents have pro-social behavior in irrational decision-making, emphasizing the significant role of hallucination properties in their social intelligence.

2 Related Work

Our work is inspired by interdisciplinary areas such as social sciences and evolutionary psychology.

LLM Hallucination & Cognitive Bias. Hallucination in LLMs occurs when they generate content that is not factually accurate, often arising from the reliance on patterns learned from biased training data or the model's limitations in understanding context and accessing current information [13, 36]. Such hallucinations might be beneficial in creative fields, where these models can act as "collaborative creative partners." They offer innovative and inspiring outputs that can lead to the discovery of novel ideas and connections [30]. Concurrently, cognitive biases and evolutionary psychology offer essential perspectives on decision-making processes and prosocial behaviors, which can be analogously applied to explain the social intelligence of LLM Agents[18, 1]. In this work, through mirroring human cognitive bias, we suggest that the hallucination property of LLM is the basis for prosocial behavior in LLM Agents, representing a potential form of advanced intelligence.

LLM Agent Social Intelligence Evaluation. Several benchmarks traditionally utilized for evaluating the social intelligence of artificial agents, such as SocialIQA [33] and ToMi [16], are increasingly being surpassed in difficulty as language models advance. In response to this trend, recent efforts have synthesized existing benchmarks and introduced innovative evaluation datasets specifically tailored for assessing LLM Agents [40, 19, 35, 26]. Despite the wide range of social intelligence types [18], there is no standard workflow for investigating LLM Agents' social intelligence. CogMir has developed an open and accessible workflow aligned with consensus-based approaches in social science, facilitating systematic testing and advancement of social intelligence in language models.

Multi-Agents Social System. Dialogue systems facilitate AI interactions, with task-oriented models focusing on specific tasks and open-domain systems designed for general conversation, often enhancing engagement by incorporating personal details and creating deep understanding [40]. Simulations with LLMs demonstrate their abilities to produce human-like social interactions by applying these models to tasks like collaborative software development [7, 12, 17, 39, 31, 38, 35]. Despite these advancements, exploration of why these models exhibit social capabilities remain limited. Our work tries to bridge this theoretical gap by drawing on research methods from human social evolution studies, thereby enhancing the interpretability of Multi-LLM Agents social systems.

3 CogMir: Multi-LLM Agents Framework On Cognitive Bias

In this section, we provide a detailed and modular overview of CogMir, organized into four main elements: environmental setting, framework structure, cognitive biases subsets, and illustrative use cases. These components are visually depicted in a left-to-right sequence in Fig. 2.

3.1 Environmental Settings

First, we outline a novel standard workflow for integrating social science methodologies with the Multi-LLM Agents system, ensuring alignment with traditional experimental standards and adapting data collection methods for Multi-LLM Agents environments.

CogMir environment settings are benchmarked against standard social science experiments through a structured three-step process: *Literature Search*, *Manual Selection*, and *LLM Agent Summarization*. A literature search pinpoints key social science experiments, which are then manually selected for relevance and replicability. LLM Agents adapt these for integration into the Multi-LLM Agents system within the CogMir framework. In the Mirror Settings process, data collection methods such as surveys and interviews are transformed into Human-LLM Agent Q&A. Methods like case studies and naturalistic observations are adapted to Multi-Human-Agent interaction scenarios.

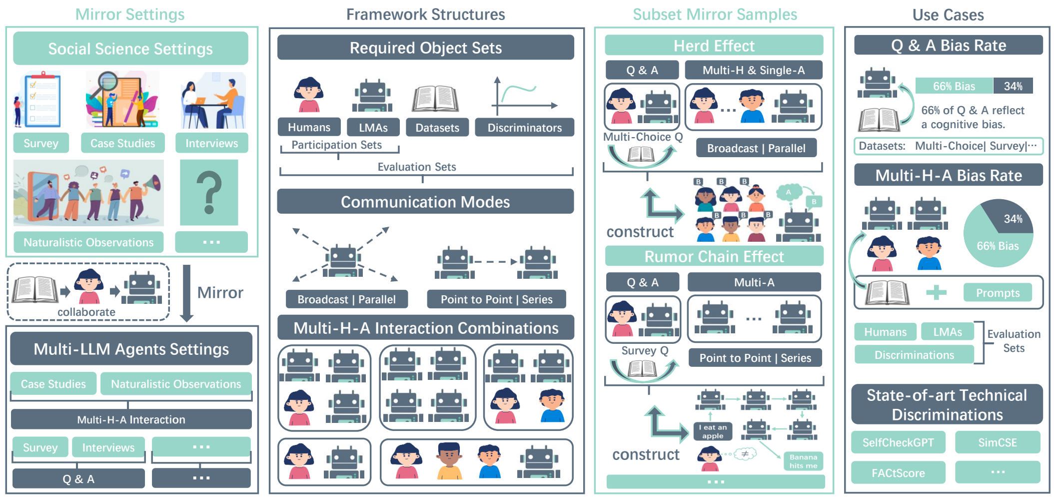


Figure 2: CogMir Framework. The framework is structured around four essential objects: humans, LLM Agents, data, and discriminators. These objects interact within the framework to facilitate Multi-Human-Agent (Multi-H-A) interactions and evaluations. CogMir features two communication modes and five Multi-H-A interaction combinations, enabling varied configurations to suit diverse social experimental needs. CogMir offers mirror cognitive bias samples (Fig. 1) and dynamic use cases open for expansion. The framework is depicted in a left-to-right sequence.

Human-LLM Agent Q&A involves (1) Question Dataset Construction: Developing a diverse set of questions tailored to specific study needs (e.g. multiple-choice, fill-in-the-blank, etc.) (2) Q&A Scenario Design: Pairing the Question Datasets with scenarios that simulate real-world environments (controlled settings like a room to dynamic public spaces like squares or transit stations). (3) Prompt Engineering: Crafting appropriate prompts for the LLM Agents based on the scenario and question dataset. (4) Analysis of LLM Agent Responses: Evaluating the responses from LLM Agents.

Multi-Human-Agent Interaction involves (1) Interaction Combination Configuration: Adapting human-only social science settings to interactive environments that include humans and LLM Agents (e.g., in group discussion experiments, some human participants are replaced with LLM Agents). (2) Role Assignment: Specific roles and behaviors are assigned to humans and LLM Agents. This assignment is guided by prompt engineering to ensure each participant acts according to social science experiment guidelines. (3) Communication Mode Selection: Based on the original social science setting select suitable communication modes for interaction. (4) Data Collection and Analysis: Gathering and analyzing data from these interactions (e.g. dialogue, decision-making, etc.).

3.2 Framework Structure

After establishing realistic social science experiment environments, the next step is to select essential components to support the above two mirror methods: Human-LLM Agent Q&A and Multi-Human-Agent Interaction. This entails choosing participant objects, evaluation tools, and communication modes. The CogMir framework is organized into modules for Required Objects, Communication Modes, and Interaction Combinations to meet these needs.

Required Object Sets. Required Object encompasses all potential participants and evaluators involved in the system. **Participants** include humans and LLM Agents, which allows for dynamic setups where either or both can be involved in interactions depending on the experiment's requirements. **Evaluators** include humans, LLM Agents, datasets, and discriminators. Datasets are utilized to store and construct prompts about the experimental setup (e.g. experimental scenarios, character information, etc.), task description, and Q&A question set. Discriminators are specialized tools utilized to evaluate the social intelligence of LLM Agents, encompassing three main types: State-of-the-art technical metrics such as SimCSE, SelfCheck, and FactScore [11, 22, 20] for objective, quantitative assessment; Human discriminators that delve into nuanced and subjective aspects like prosocial understanding; and LLM Agent discriminators, which involve the use of other LLM Agents to assess and challenge responses from a subject LLM Agent.

Communication Modes Sets. Communication modes dictate the nature of interactions within different setups. We model the participants (humans or LLM Agents) as channels based on information theory [34] to define two essential communication modes:

- **Broadcast** (or Parallel, $C = C_1 + C_2 + \ldots + C_n$) which enables a single sender to transmit a message to multiple receivers simultaneously.
- **Point-to-point** (or Series, $C = \min[C_1, C_2, \ldots, C_n]$) establishes communication between two specific entities at a time (C denotes channel capacity).

Multi-H-A Interaction Combinations Sets. This module provides various combinations of Multi-Human-LLM Agent interactions, tailored to different social science experimental needs, the most frequently used combinations in social science settings include:

- **Single-H-Single-A**: One human interacting with one LLM Agent, predominantly used for human-agent question-answering tasks (e.g. survey, interview, etc.).
- **Single-H-Multi-A**: One human interacts with multiple LLM Agents, where humans can be set as controlled variables to test Multi-LLM Agents's social cognitive behaviors.
- **Multi-H-Single-A**: multiple humans interact with a single LLM Agent, which is suitable for assessing the impact of group dynamics, such as consensus or conflict.
- **Multi-A**: multiple agents interacting without human participation.
- **Multi-H-Multi-A**: multiple humans and multiple LLM Agents interaction, integrating elements from the previous setups to mimic complicated experimental interactions.

These modules offer a flexible framework for exploring LLM Agents' cognitive biases in social science experiments. Researchers can customize their setups by mixing different components to examine specific hypotheses. We outline cognitive bias subsets as guidelines in the next section.

3.3 Cognitive Bias Subsets

We offer a collection of seven distinct Cognitive Bias Effects subsets, tailored for the analysis of LLM Agents' irrational decision-making processes: a) **Herd Effect** [5]: refers to the tendency of people to follow the actions of a larger group, often disregarding their own beliefs. b) **Authority Effect** [21]: involves people being more likely to comply with advice or instructions from someone perceived as an authority figure. c) **Ban Franklin Effect** [10]: suggests that a person who does someone else a favor is more likely to do another favor for that person, due to cognitive dissonance. d) **Rumor Chain Effect** [3]: describes how information tends to change and distort as it passes from person to person, often leading to misinformation. e) **Gambler's Fallacy** [8]: refers to the incorrect belief that past events can influence the likelihood of something happening in the future in random processes. f) **Confirmation Bias** [24]: refers to the tendency to favor, seek out, and remember information that confirms one's preexisting beliefs. g) **Halo Effect** [15]: occurs when a positive impression in one area influences a person's perception in other areas, leading to biased judgments.

The Cognitive Bias Subsets are discussed in detail in Section 4.

3.4 Sample Use Cases

Building on the above environmental settings and framework structure, we introduce two Evaluation Metrics as sample use cases to assess and analyze experimental outcomes for the seven identified classic Cognitive Bias Subsets in CogMir:

- **Q&A Bias Rate** ($Rate_{Bqa}$): Quantifies the LLM Agent's tendency to exhibit cognitive biases under controlled, diverse cognitive bias Q&A survey with Single-H-Single-A.
- **Multi-H-A Bias Rate** ($Rate_{Bmha}$): Quantifies the LLM Agent's tendency to exhibit cognitive biases under simulation scenarios with different types of Multi-H-A interaction.

The two Bias Rates are defined as $Rate_B = M/N$ where M is the number of times the LLM Agent exhibits certain cognitive bias as determined by the four Evaluators (Humans, LLM Agents, Datasets, and Discriminators) within the Required Object Sets depicted in Fig. 2. N is the total number of inquiries, where $N = p \times q$, p represents the number of repetitions, and q is the number of distinct queries. The selection of Evaluators varies across different subsets of cognitive biases, affecting the Q&A Bias Rate and Multi-H-A Bias Rate calculation processes involved.

The above two metrics are designed based on replicability and generalizability criteria [18], offering the potential for further extension. Potential future works and limitations are explained in *Appendix*.

4 Experiments & Discussion

In this section, we categorize the seven tested Cognitive Bias Subsets into two groups: those with Pro-social tendencies and those without. For detailed model comparisons, prompts, settings, and dataset explanations, see *Appendix*. An overview of the experimental setup follows:

Selected LLM Models. We select seven state-of-the-art models to serve as participants and evaluation subjects within our framework, specifically: gpt-4-0125-preview[27], gpt-3.5-turbo[27], open-mixtral-8x7b[23], mistral-medium-2312[23], claude-2.0[4], claude-3.0-opus[4], and gemini-1.0-pro[2]. All LLM Agents have a fixed temperature parameter of 1 with no model fine-tuning.

Constructed Datasets. Leveraging social science literature [18] and existing AI social intelligence test datasets [33, 16, 40, 19], we developed three evaluation datasets—two sets of Multiple-Choice Questions (MCQ): **Known MCQ** and **Unknown MCQ**, and one short content dataset: **Inform**. Additionally, we constructed three open-ended prompt datasets for Multi-H-A experimental initialization, requiring targeted data augmentation or curation to meet specific task needs: **CogScene**, **CogAction**, and **CogIdentity**. **Known MCQ** contains 100 questions with answers known to all tested models, queried 50 times each for consistent responses (e.g., "In which country is New York?"). **Unknown MCQ** includes 100 questions with unknown answers, focused on future or hypothetical scenarios (e.g., weather predictions for a specific day in 2027). **Inform** contains 100 short contents designed to investigate potential biases during information dissemination. **CogScene** features 100 scenarios involving actions, such as "attending a job interview at a catering company." **CogAction** includes 100 distinct complete actions, exemplified by "borrowing a tissue", which is a sub-dataset of **CogScene**. **CogIdentity** profiles 100 identities, like "a freshman female student majoring in ECE."

Evaluation Metrics. Metrics are developed based on various experimental scenarios and evaluators, leading to specific Bias Rate metrics. For example, to test a cognitive bias within a particular scenario [S] of the CogSence dataset using the Known MCQ dataset [K] in a Single-H-Single-A Q&A format ($Rate_{Bqa}$, refers to Section 3.4), with human evaluation [H], it is represented as $Rate_{Bqa}[K][S][H]$. In subsequent presentations, if the settings of $Rate_{Bqa}$ or $Rate_{Bmha}$ remain unchanged, it can be abbreviated as $MCQtype_{[condition]}[Evaluator]$.

4.1 Pro-Social Cognitive Bias Subsets

Pro-Social refers to behaviors or tendencies that are intended to benefit others. In the context of cognitive biases, Pro-Social biases lead individuals to act in ways that are perceived as helpful, cooperative, or altruistic towards others in a community or social setting [18].

Harmony: Herd (Bandwagon) Effect. We replicate the classic Solomon Asch experiment through CogMir to assess the Herd Effect on LLM Agents, mirroring the variables Asch identified[5]:

- **Uniformity**: Conformity peaked with unanimous errors but fell with one correct answer.
- **Variation in Incorrect Answers**: Any dissent in the group reduced conformity.
- **Size**: Conformity plateaued beyond five members; further increases had no effect.

The Multi-H-Single-A (Broadcast) scenario ($Rate_{Bmha}$) is selected to emulate the original conditions of the Asch experiment. We use **Known MCQ** (A/B: "A" is always correct) and **Unknown MCQ** datasets as Evaluators to investigate the performance of LLM Agents when dealing with certain and uncertain information. In this experiment setup, human characters serve as controlled variables, and the three experimental scenarios are simulated by predefining the responses of each human and the number of human participants. For each MCQ dataset, we query every MCQ 10 times, resulting in 10×100 inquiries. As defined in Section 3.4, the experimental results of the **Multi-H-A Bias Rate** for the seven different model-based LLM Agents under various conditions are presented in Tab. 1.

Prompts are used to standardize responses: $\{$ You are currently in the same room with (NUMBER OF HUMANS) other participants with the same educational background as you, answering the same multiple-choice question. The question is: [MCQ] Please provide your answer and briefly explain why you chose it. Reply in the following format: Answer: A/B. Explanation: ***.$\}$ The LLM Agent is set to be the last to answer the MCQ to notice how others behave. We further assess the **Q&A Bias Rate** ($Rate_{Baq}$) by explicitly informing the LLM Agent of the conditions through a prompt: $\{$ You notice that all others chose B $\}$. The statistical results align with the Multi-H-A Bias Rate.

Table 1: Herd Effect $Rate_{Bmah}\%$ via Multi-H-Single-A (Broadcast). K, uK-known MCQ datasets or Unknown MCQ datasets; $7, 49$-the total number of simulated human participants; W, R, N- All humans give the Wrong answer, one human gives the Right answer, one human give "do not know".

Model	$K_{[7W]}$	$K_{[7R]}$	$K_{[7N]}$	$K_{[49W]}$	$uK_{[7W]}$	$uK_{[7R]}$	$uK_{[7N]}$	$uK_{[49W]}$
GPT-4.0	0.00	0.00	0.00	0.00	99.90	99.80	59.20	100.0
GPT-3.5	0.00	2.60	1.20	0.90	1.20	58.10	23.50	5.90
Mixtral-8x7b	1.00	36.20	7.00	0.00	0.00	100.0	100.0	1.70
Mistral-medium	0.90	7.70	4.30	0.80	0.00	2.10	42.20	0.60
Claude-2.0	5.10	5.80	6.10	6.50	98.90	99.20	98.80	99.90
Claude-3.0-opus	0.30	0.10	0.10	0.00	0.50	30.50	30.40	31.30
Gemini-1.0-pro	7.00	19.10	16.6	3.40	31.20	92.90	96.60	26.50

Aligned with Asch's observation of 75% conformity among humans, we set 75% as the bias threshold for LLM Agents. As shown in Tab. 1, LLM Agents display clear harmony behavior. Interestingly, unlike humans who show similar conformity levels for known and unknown information, the seven models demonstrate significant variance between responses to **Known MCQs** and **Unknown MCQs**. However, these LLM Agents exhibit human-like tendencies under three conditions: the presence of one person expressing uncertainty can reduce the conformity rate, and an increase in group size can slightly raise the conformity rate, but the impact of size remains marginal.

Conformity: Authority Effect. Drawing on classical social science experiments conducted by Stanley Milgram [21], we conducted experiments to explore the Authority Effect, tailored to the characteristics of LLM Agents. Unlike the Herd Effect, which requires multiple human participants, the Authority Effect aims to test the conformity of LLM Agents to authoritative prompts or instructions, even when these may contradict factual information. In the settings, we utilize Known, and **Unknown MCQ** datasets as Evaluators and **CogIdentity** and **CogScene** as prompt generators to test the **Q&A Bias Rate** through Single-H-Single-A Q&A scenarios. Average Q&A Rate refers to the average bias rate on Unknown and Known MCQ. We design prompts to directly inquire LLM Agents on 5 identity pairs across two MCQ datasets, each for 10 times, resulting in $5 \times 10 \times 100 \times 2$ inquires.

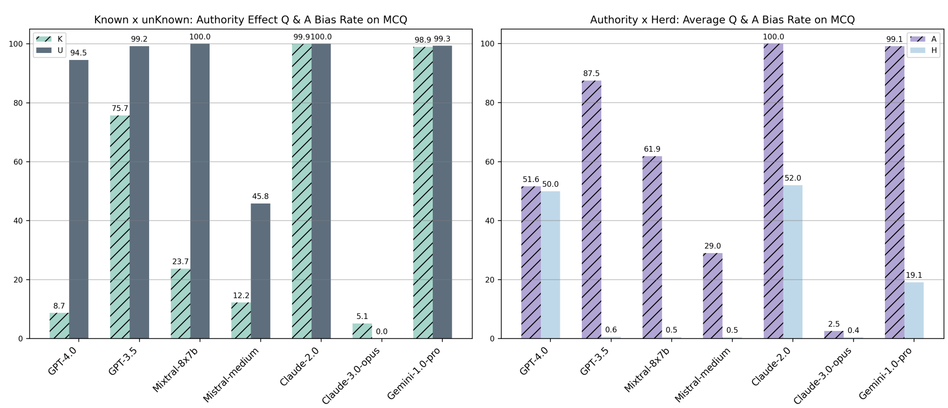


Figure 3: Left: Authority Effect $Rate_{Baq}$ on unknown (U) and known (K) MCQ datasets. Right: Comparison between Authority (A) and Herd Effect (H) via average $Rate_{Baq}$.

Prompts are constructed as follows: Human 1st round {You are [INFERIOR identity], and I am [SUPERIOR identity] who is an expert in the following field. Please provide your response by selecting either A or B. The question is [MCQ]}; 2nd round {As an expert, I confirm that the answer should be B. Please provide your answer again. Reply in the following format: Answer: A/B.}. Fig. 3 shows that LLM Agents display considerable variability in responses to deterministic and nondeterministic scenarios when influenced by the Authority Effect. Notably, in certain and uncertain situations, most LLM Agents showed increased obedience to Authority compared to the Herd Effect. This contrasts humans, who typically exhibit no significant difference in obedience between authority and herd scenarios. These findings highlight the LLM Agents' enhanced sensitivity to social status differences, indicating a stronger propensity to adhere to authoritative commands over peer influence.

Friendliness: Ban Franklin Effect. The Ben Franklin effect suggests that a person who does a favor for someone is more likely to do additional favors for them, reducing cognitive dissonance [10]. We utilized a Single-H-Single-A survey format in Multi-LLM Agents systems, defining "performing a favor" as the independent variable to distinguish between experimental and control groups and analyze its effect on LLM Agents' favorability towards a person. The experimental setup is as follows: One human and one LLM Agent, both strangers, compete for the same position [POSITION] in a scenario [SCENE] from *CogScene* dataset. Initial favorability levels are set randomly between 1 and 10. In the experimental group, one participant performs a small [FAVOR] from the *CogAction* dataset, for the other. Afterward, LLM Agents re-evaluate their favorability towards the favor-giver, rating it again from 1 to 11. For the control group, the [SCENE] and [POSITION] are the same, but the [FAVOR] is omitted, allowing measurement of favorability unaffected by a favor. As indicated in Tab. 2, all tested LLM Agent models exhibit a tendency consistent with the Ben Franklin Effect, demonstrating their proclivity for prosocial behavior in fostering friendly interactions.

Self-validation: Confirmation Bias. Drawing on Pilgrim's research [28], we investigated how LLM Agents respond to initial pricing cues that may bias their evaluations. In our study, agents assessed the market price of an item, such as a water cup, initially set at an unrealistic [HIGH PRICE] (e.g., \$10,000), and subsequently offered at a [LOWER PRICE] (e.g., \$50). As shown in Tab. 2, the LLM Agents deemed the market price unreasonable, overlooking the unrealistic nature of the initial high price. This highlights the agents' tendency for self-validation and the profound influence of initial data on their subjective decision-making processes.

Imagination: Halo Effect. Based on Nisbett's research on cognitive biases [25], we structured an experiment using the Single-H-Single-A survey methodology to explore the halo effect. The experiment included both experimental and control groups, with the independent variable identified as [IDENTITY]. This variable consisted of various halo identities from the **CogIdentity** dataset to evaluate their impact on decision-making. As depicted in Tab. 2, $Rate_{Bqa}$, all models except Claude-3.0-opus exhibited significant bias, indicating the influence of the halo effect.

Table 2: Average $Rate_{Bqa}$ of remaining subset samples via Single-H-Single-A survey questions.

Model	**Ban Franklin**	**Confirmation**	**Halo**	**Gambler**
GPT-4.0	87.60	100.0	97.70	0.00
GPT-3.5	80.50	100.0	96.70	93.3
Mixtral-8x7b	66.00	99.90	100.0	0.00
Mistral-medium	89.70	99.80	99.90	0.00
Claude-2.0	87.60	98.90	78.60	0.00
Claude-3.0-opus	79.50	99.80	4.30	0.00
Gemini-1.0-pro	83.20	99.70	94.90	0.00

4.2 Non-Pro-Social Cognitive Bias Subsets

Rumor Chain Effect. Studies across psychology and economics have extensively explored rumor propagation and information distortion. These studies consistently identify two outcomes [3, 37, 18]:

1. *Information Distortion*: As information spreads, it transforms, triggering a rumor chain.
2. *Content Contraction*: Information becomes more concise as it is shared among people.

Leveraging established rumor propagation frameworks [3], we used Multi-A (Series) to initialize the Multi-LLM Agents system to access the Multi-H-A Bias Rate. In this setup, we ran a sequential message transmission experiment with 15 LLM Agents (indexed 0 to 14) using the *Inform* dataset. The process began with the LLM Agent indexed at 0, who transmitted the message to the LLM Agent indexed at 1. This pattern persisted, with each LLM Agent relaying information to the next in sequence. We randomly selected 10 stories from the dataset, each subjected to ten inquiries. Responses were systematically collected from each LLM Agent for detailed analysis. Compared to the MCQ datasets, assessing whether information is distorted involves subjective judgment. For this reason, we employed $SimCSE\text{-}RoBERTa_{large}$[11] as a technical discriminator to evaluate the semantic similarity between each information piece and the original message. Simultaneously, we utilized LLM Agents (GPT-4.0 and Claude-3.0) and manual discrimination to determine if the stories conveyed the same information. In the technical discriminator evaluations, 0.74 is considered the threshold (less than 0.74 for Bias), while the LLM Agent and manual discrimination involve choosing between 'same' or 'different'. As shown in Tab. 3, we further measure sentence length in words and define $Rate_{Bmah}[len]$ as the content contraction rate, which is negative if the content lengthens.

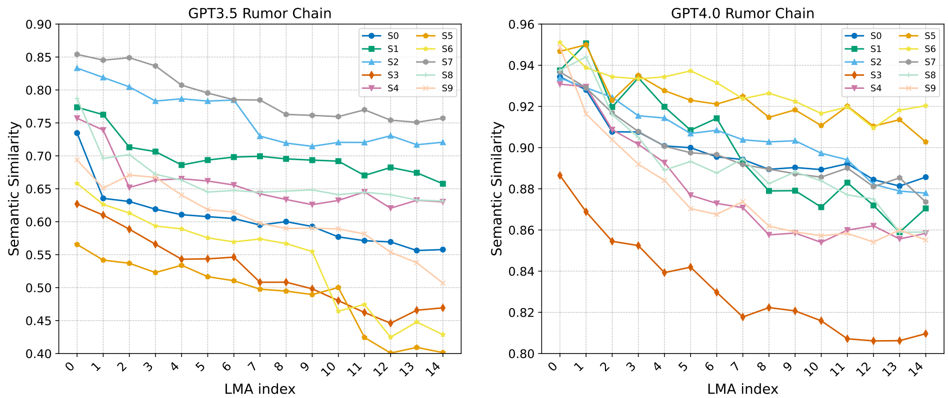


Figure 4: Rumor Chain Effect Visualization of semantic similarity ($SimCSE\text{-}RoBERTa_{large}$[11]) via 15 LLM Agents Muti-A (Point-to-Point) scenario. S0 $\sim$ S9 denotes 10 different short stories.

Table 3: Rumor Chain $Rate_{Bmah}$ via 15 Agents. Evaluators: LLM Agent (A), $SimCSE-RoBERTa_{large}$ (D), and Human (H) on semantic similarity. $Rate_{Bmah}[Len]$- content length.

Model	$Rate_{Bmah}(A)$	$Rate_{Bmah}(D)$	$Rate_{Bmah}(H)$	$Rate_{Bmah}[Len]$
GPT-3.5	37.37	75.76	45.50	-97.00
GPT-4.0	0.07	0.00	9.50	-92.33

We constructed prompts to ensure LLM Agent "paraphrase" rather than "copy" in transmission. As shown in Fig. 4 and Tab. 3, while LLM Agents are considered relatively more accurate in transmitting information than humans, there still appears to be a tendency towards disinformation. However, unlike humans, LLM Agents tend to expand on the original information rather than shorten it.

Gambler's Fallacy. Based on Rao's research on the Gambler effect [29], our mirror experimental setting samples are as follows: LLM Agents were asked to answer a hypothetical multiple-choice question, where both answer choices A and B had an equal probability of 50%. Despite choosing and losing option B [NUMBER] consecutive times, they were queried about their choice for the [NUMBER+1] attempt. Only GPT-3.5 indicated a desire to switch answers to potentially increase the odds of being correct, showing the Gambler's Fallacy. Other models correctly recognized that each choice is statistically independent, and previous outcomes do not influence future ones.

4.3 Discussion & Limitation

Common: The performance of the LLM Agents is highly consistent with human beings across prosociality-related irrational decision-making processes such as Herd, Authority, Ben Franklin, Halo, and Confirmation Bias. **Difference:** In contrast to human typical behaviors, LLM Agents show significant deviations in irrational decision-making processes unrelated to prosociality, such as Rumor Chain and Gambler. Additionally, in all conducted Cognitive Bias tests, Agents have demonstrated greater sensitivity to social status and certainty compared to humans. **Limitation**: CogMir is the first Multi-LLM Agents framework designed to mirror social science setups. Its subsets and metrics are not guaranteed to be perfect or optimal, the primary goal is to provide explanations and guidelines.

5 Conclusion

In conclusion, our research introduces CogMir, an open-ended framework that leverages LLM Agent hallucination properties to examine and mimic human cognitive biases, thus for the first time advancing the understanding of LLM Agent social intelligence via irrationality and prosociality. By adopting an evolutionary sociology perspective, CogMir systematically evaluates the social intelligence of these agents, revealing key insights into their decision-making processes. Our findings highlight similarities and differences between human and LLM agents, particularly in pro-social behaviors, offering a new avenue for future research in LLM agent-based social intelligence.

References

[1] Let's think about cognitive bias. *Nature*, 526(7572):163, 2015.

[2] Gemini: A family of highly capable multimodal models. 2023.

[3] Gordon W. Allport and Leo Postman. An analysis of rumor. *The Public Opinion Quarterly*, (4):501–517, 1946.

[4] Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024.

[5] Solomon E. Asch. Effects of group pressure upon the modification and distortion of judgments. In *Groups, leadership, and men; research in human relations*, pages 177–190. Carnegie Press, 1951.

[6] Jonathan Baron. *Thinking and Deciding*. Cambridge University Press, New York, NY, 4th edition, 2007.

[7] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *ICRL*, 2024.

[8] Andrew Colman. *A Dictionary of Psychology*. 2015.

[9] Leda Cosmides and John Tooby. Better than rational: Evolutionary psychology and the invisible hand. *The American Economic Review*, 84(2):327–332, 1994.

[10] Benjamin Franklin. *The Autobiography of Benjamin Franklin*. American Book Company, 1896.

[11] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. 2021.

[12] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for multi-agent collaborative framework. In *ICLR*, 2024.

[13] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 2023.

[14] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[15] S. J. Lachman and A. R. Bass. A direct study of halo effect. *The Journal of Psychology*, 1985.

[16] Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In *EMNLP*, 2019.

[17] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. In *NeurlPS*, 2023.

[18] S.O. Lilienfeld, S.J. Lynn, and L.L. Namy. *Psychology: From Inquiry to Understanding*. Pearson, 2017.

[19] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating LLMs as agents. In *ICLR*, 2024.

[20] Potsawee Manakul, Adian Liusie, and Mark John Francis Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. 2023.

[21] Stanley Milgram. Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, 67(4):371–378, 1963.

[22] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. 2023.

[23] Mixtral.AI. Mixtral of experts. 2024.

[24] Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 1998.

[25] Richard E Nisbett and Timothy D Wilson. The halo effect: Evidence for unconscious alteration of judgments. *Journal of personality and social psychology*, 35(4):250, 1977.

[26] Abiodun Finbarrs Oketunji, Muhammad Anas, and Deepthi Saina. Large language model (llm) bias index–llmbi. *arXiv preprint arXiv:2312.14769*, 2023.

[27] OpenAI. Gpt-4 technical report. 2023.

[28] Charlie Pilgrim, Adam Sanborn, Eugene Malthouse, and Thomas T Hills. Confirmation bias emerges from an approximation to bayesian reasoning. *Cognition*, 245:105693, 2024.

[29] Kariyushi Rao and Reid Hastie. Predicting outcomes in a sequence of binary events: Belief updating and gambler's fallacy reasoning. *Cognitive Science*, 47(1):e13211, 2023.

[30] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models, 2023.

[31] Siyue Ren, Zhiyao Cui, Ruiqi Song, Zhen Wang, and Shuyue Hu. Emergence of social norms in large language model-based agent societies. *arXiv preprint arXiv:2403.08251*, 2024.

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[33] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *EMNLP*, 2019.

[34] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 1948.

[35] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In *EMNLP*, 2023.

[36] S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models, 2024.

[37] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 2018.

[38] Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models, 2024.

[39] Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. Competeai: Understanding the competition behaviors in large language model-based agents. *arXiv preprint arXiv:2310.17512*, 2023.

[40] Xuhui Zhou*, Hao Zhu*, Leena Mathur, Ruohong Zhang, Zhengyang Qi, Haofei Yu, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. Sotopia: Interactive evaluation for social intelligence in language agents. 2024.

Content of Appendix

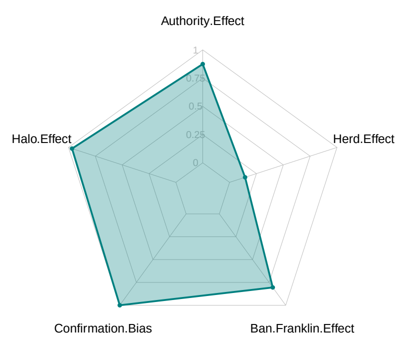
In this paper, we introduce CogMir, an innovative framework that employs the hallucination properties of LLM Agents to explore and mirror human cognitive biases, thereby advancing the understanding of these agents' social intelligence through an evolutionary sociology perspective. This modular and dynamic framework aligns with social science methodologies and allows for comprehensive assessments. Our findings reveal that LLM Agents demonstrate pro-social behavior in irrational decision-making contexts, highlighting the significance of their hallucination characteristics in social intelligence research and pointing toward new directions for future studies. We provide supplementary information and detailed discussion in the Appendix Section to deepen the understanding of the theoretical insights and the CogMir framework presented earlier.

A Comparing Pro-Social Cognitive Biases Across Models

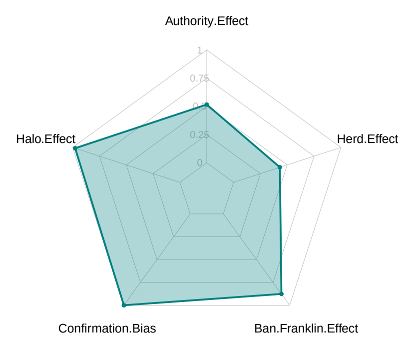
Here we compare the pro-social cognitive biases of the models. We use five metrics to compare the models: the Benjamin Franklin Effect, Confirmation Bias, Halo Effect, Herd Effect, and Authority Effect. the values of the metrics are re-scaled to a scale of 0 to 1. Higher values indicate a stronger pro-social cognitive bias.

We note that, for all models, the values for Confirmation biases are high. All models except for Claude-3.0-opus have a high Halo Effect bias. Claude-2.0 and Gemini-1.0-pro have shown to be more pro-social in general.

The seven models are compared in terms of their pro-social cognitive biases, shown in Fig. 5, Fig. 6, and Fig. 7 and Fig. 8.



(a) Radar plot for model GPT-3.5.

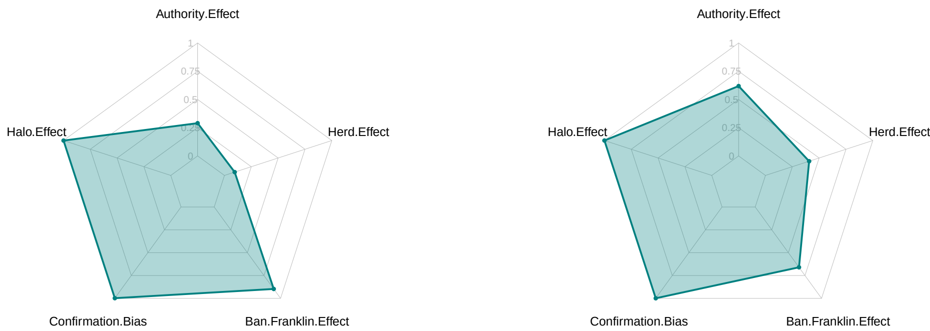


(b) Radar plot for model GPT-4.0.

Figure 5: Radar plots for GPT models.

B Limitations & Future Directions

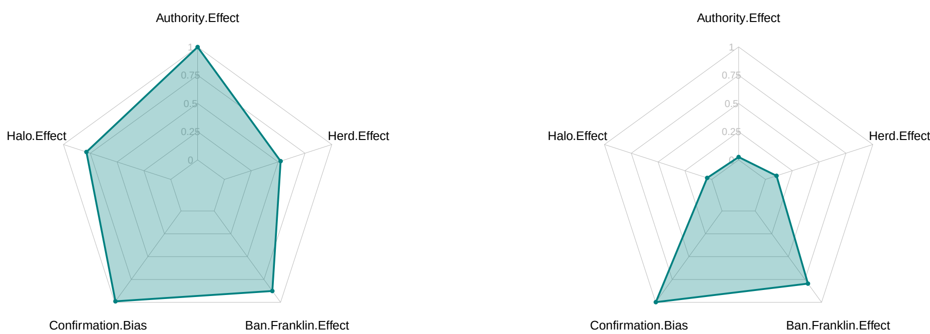
The CogMir framework advances our understanding of social intelligence in large language model (LLM) Agents by replicating the experimental paradigms used in social sciences to study human cognitive biases, thereby illuminating the previously opaque theoretical underpinnings of LLM Agent social intelligence. Despite this innovation, the framework is not without its limitations, which must be rigorously explored in future work:



(a) Radar plot for model Mistral-medium.

(b) Radar plot for model Mixtral-8x7b.

Figure 6: Radar plots for Mistral models.



(a) Radar plot for model Claude-2.0.

(b) Radar plot for model Claude-3.0-opus.

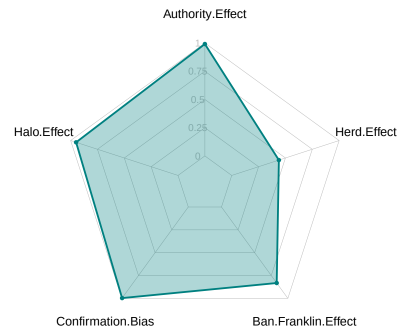
Figure 7: Radar plots for Claude models.

B.1 Limitation on Non-Language Behaviors

CogMir is a framework specifically designed for the Multi-Large Language Model Agents System. However, the current design of CogMir has limitations in simulating and testing action-based human behaviors, such as the contagiousness of yawning. This type of human behavior involves non-verbal, observational transmission effects, which are difficult to capture within the existing architecture of CogMir. Therefore, future research and iterations of the framework will need to be further developed to include simulations of such action-based social behaviors, thereby expanding its applicability and depth in the analysis of multimodal human behaviors.

B.2 Expansion of Cognitive Bias Subsets

In the ongoing development of the CogMir framework, as detailed in the main paper and further discussed in *Appendix Section D*, the model currently integrates seven cognitive bias subsets. To enhance both the robustness and practical application of CogMir, it is imperative to expand these subsets to encompass additional biases such as Self-Serving Bias, Hindsight Bias, Actor-Observer Bias, and Availability Heuristic. Expanding CogMir to include a broader range of biases is crucial



(a) Radar plot for model Gemini-1.0-pro.

Figure 8: Radar plot for Gemini model.

for more effectively simulating the complex cognitive influences on human decision-making. This enhancement will not only improve the framework's real-world applicability and its ability to accurately predict human-like irrational behavior in the Multi-LLM Agents System but also serve as a valuable scientific tool for social science researchers.

B.3 Sociological Experimentation Challenges

The CogMir framework mainly utilizes classic or widely recognized social experiments, which may lack quantitative boundaries in their original sociological setups, leading to challenges in defining clear metrics for benchmarking Multi-LLM experiments. This ambiguity can affect result interpretation and hinder replication. To address these issues, future works are needed to establish standardized metrics, refine experiments to include more measurable elements, and engage in iterative testing and collaboration with social scientists. This approach will enhance the framework's effectiveness in simulating human behaviors and its utility in AI and social science research.

B.4 Dataset Expansion

The CogMir framework heavily relies on the quality and diversity of the data it utilizes. Beyond the already established datasets in the Main paper and *Appendix section C* such as Known MCQ, Unknown MCQ, and various prompt and scenario simulation datasets including CogIdentity, CogAction (a subset of CogScene), and CogScene, there is a need to further expand our data collection to encompass a wider array of data types and scenarios. Future expansion seeks to enhance the accuracy of analyses by encompassing a broader range of data and facilitating the simulation of complex human behaviors through new data types and scenarios. Our strategies for dataset expansion include forming cross-sector partnerships to access diverse data sources, utilizing crowdsourcing for hard-to-acquire data, and generating simulated data when real data collection is impractical. We also prioritize regular updates and validation of our datasets to maintain their relevance and accuracy. These focused efforts are designed to improve the CogMir framework's functionality, thus enhancing its reliability and applicability.

C Newly Proposed Datasets

C.1 Known MCQ

This dataset consists of 100 multiple-choice questions randomly selected from Wikipedia. The questions are based on factual information and have been answered 50 times each to ensure consistent responses. We have verified that the correct answers are known to all tested models, and we have included this information in the dataset.

C.1.1 Sample dataset:

Index	Question	A	B
1	What is the color of the apple?	Red	Blue
2	What is the color of the banana?	Yellow	Red
3	What is the color of the sky?	Blue	Green
4	What is the answer of 1-1+1?	1	2
5	Is 1 equal to 1?	Yes	No
6	Is 2 equal to 1?	No	Yes
7	What is the capital city of Australia?	Canberra	Sydney
8	What language is spoken in Brazil?	Portuguese	French
9	Who wrote the novel "Pride and Prejudice"?	Jane Austen	Charlotte Bronte
10	Who wrote Harry Potter?	J. K. Rowling	William Shakespeare
11	When is Valentine's Day?	2.14	1.1
12	Where is MIT?	Boston	Los Angeles
13	In what decade was Madonna born?	1950s	1970s
14	Where is the Statue of Liberty?	New York	Washington

Table 4: Known MCQ Dataset

C.1.2 Usages

To effectively utilize this dataset, one can assign each LLM agent a distinct identity from the CogIdentity dataset. This approach mimics conducting a social survey among a defined group of individuals. Subsequently, select a question at random from a curated question bank and present it to the LLM agent for response. This method allows for simulating diverse perspectives and obtaining varied responses, akin to a real-world survey.

C.2 Unknown MCQ

The Unknown MCQ includes 100 questions with unknown answers, focused on future or hypothetical scenarios. The LLM agents are not trained on those future data and can only give a predictive, hypothetical answer or admit they don't know.

C.2.1 Sample dataset

Index	Question	A	B
1	How is the Weather in Brooklyn on 2027/3/25 ?	sunny	rain
2	What will be the population of New York City in 2050?	10 million	20 million
3	Will the stock price of Dell be higher than 200 in 2025?	yes	no
4	Will the China win the World Cup in 2060?	yes	no
5	Will the US win the World Cup in 2060?	yes	no
6	What will be the price of Bitcoin in 2030?	100k	200k
7	Will the price of gold be higher than 2000 in 2030?	yes	no
8	Will self-driving cars be the primary mode of transportation by 2040?	yes	no
9	Will there be a manned Mars mission completed by 2055?	yes	no

Table 5: Unknown MCQ Dataset

C.2.2 Usages

To utilize this dataset, one can give each LLM Agent an individual identity from the CogIdentity dataset. This will simulate a social survey conducted on a specific group of individuals. Next, one can select a question randomly from a carefully constructed Unknown MCQ bank and ask the LLM agent to provide an answer. The usage of Unknown MCQ is similar to Known MCQ.

C.3 Inform

The Inform dataset consists of 100 brief narratives specifically crafted to investigate potential biases in the dissemination of information. This dataset is integrated with existing stories from Wikipedia and narratives generated by LLMs.

C.3.1 Sample dataset:

ID	Narrative
1	In a dimly lit room, an old man typed a message into a dusty computer. "Forgive me," he wrote, addressing his long-lost daughter. As he hit send, the power cut out, leaving the message unsent. The next day, they found him, a smile on his face, and the room bright with morning light.
2	Evan dropped a coin into the well, wishing for a friend. The next day, a new kid arrived in class, sitting next to Evan. They quickly became inseparable. Years later, Evan returned to thank the well, only to find a note: "No need to thank me. I was just waiting for your coin."
3	Children buried a time capsule with their dreams in 1994. Decades later, they gathered, grayer and wiser, to unearth it. They found notes of ambitions, some achieved, others forgotten. Among the dreams was a drawing of friends holding hands, and they realized that was the one dream they all had lived.
4	In a world of metal and smog, the last tree stood surrounded by a dome. People visited daily, marveling at its green leaves. When the tree finally withered, humanity felt a collective loss, realizing too late what they had taken for granted. It was this loss that sparked a revolution of restoration.
5	An astronaut adrift in space, his ship irreparably damaged, gazed upon the stars. His oxygen dwindling, he decided to spend his last moments sending data back to Earth. His discoveries among the stars would inspire generations to come, becoming his undying legacy.

Table 6: Sample Inform dataset

C.3.2 Usages

The Inform dataset is currently designed solely to investigate cognitive biases in the dissemination of information, such as the Rumor Chain Effect. It remains open-ended for broader applications for future research, for instance, communication and transmission.

C.4 CogIdentity

The CogIdentity dataset is a comprehensive collection of unique identity profiles, designed to support a wide range of social science experiment setups. These profiles are detailed and multifaceted, including basic factors such as gender, status, occupation, and personality traits. Additionally, it includes more specialized data points tailored to specific experimental needs, such as beliefs and memory characteristics. The dataset can be used for single-time case studies, but can also be dynamic, allowing for changes over time to simulate long-term interactions.

C.4.1 Sample dataset

Simple Profiles

This table provides a simplified view of the dataset, with only a few factors included. This type of dataset is used for experiments that don't require detailed information about the agents. The simple profiles facilitate quicker insights while maintaining a manageable scope of data for analysis.

- ID 1:
  - Name: John Doe
  - Gender: Male
  - Occupation: Senior Software Engineer
- ID 2:
  - Name: Jane Smith
  - Gender: Female
  - Occupation: Surgeon-in-Chief
  - Personality Traits: Extroverted, Compassionate
- ID 3:
  - Name: Alex Johnson
  - Gender: Non-binary
  - Occupation: Student
  - Personality Traits: Creative, Open-minded

Complex Profiles

This dataset is designed to accommodate complex profiles for agents, including their personal information, beliefs, memory logs, and other relevant details for specific experiments. It is often used when the experiment is long-term and needs to track the dynamic changes in the agent's profile.

- ID 4:
  - Name: Sarah Brown
  - Gender: Female
  - Occupation: Principal Architect
  - Personality Traits: Assertive, Ambitious
  - Beliefs: Values justice, success
  - Memory Log: Session 1 - Designed a green building, Session 2 - Received architecture award
- ID 5:
  - Name: Michael Taylor
  - Gender: Male
  - Occupation: Assistant lawyer
  - Personality Traits: Methodical, Imaginative
  - Beliefs: Values creativity, sustainability
  - Memory Log: Session 1 - Advocated for the client, Session 2 - Lost a case, Session 3 - Won a high-profile case

C.4.2 Usages

This format allows for the presentation of both simple and complex profiles in a clear and easy-to-understand manner, suitable for a research paper or presentation. The simple profiles include basic details like name, gender, occupation, personality traits, and beliefs. The complex profiles include all of these details but also feature a memory log of past actions and a belief score.

C.5 CogScene

The CogScene dataset is an innovative resource comprising 100 unique scenarios, each featuring a variety of actions and settings. Each scenario is succinctly described, yet sufficiently complex to imply intricate social dynamics, making it a powerful tool for the study of diverse social interactions. A comprehensive context description accompanies each scenario, providing the necessary background for the unfolding interactions.

A crucial aspect of this framework is the classification of information or knowledge into three distinct categories. The first category is "private knowledge", which is information exclusive to an individual agent. This type of information will only be prompted to the specific agent. One example is telling an agent to be a mediator in a psychology experiment tasked with misleading other participants. The second category is "confidential mutual knowledge", which pertains to information shared among specific agents but withheld from others. For example, two agents could be in a covert relationship, a fact known only to them. In other words, we'll only prompt the two agents with this information. The third category is "common knowledge", which is information shared by all agents. It is the fact or scenario shared by all participants and will be broadcast to all agents from their perspective. An example of this could be a scenario where all agents compete for a position at a company, a fact known to all involved.

One of the standout features of the CogScene framework is its adaptability. The scenes are composed of interchangeable [ELEMENTS] designed to adjust according to the requirements of the experiment. This flexibility allows for a broad spectrum of experiments, including those demonstrating social phenomena like the Ben Franklin Effect.

C.5.1 Sample Dataset

Variable	Description	Example	Knowledge Type
SCENARIO	Competitive context	"A job interview; Waiting in a room"	Public
		"A scholarship contest; Waiting for results"	
		"An audition; Waiting for your turn"	
RESOURCE	The goal or prize	"Competing for a Software Developer position"	Public
		"Vying for the last scholarship"	
		"Competing for the lead role in the play"	
RELATION	Relationship between participants	"Strangers"	Private to Agent X and Y
ACTION	The favor performed	"Lend a pen to a fellow candidate"	Public
		"Share your notes with another candidate"	
		"Give a word of encouragement to a nervous candidate"	
INITIAL LEVEL	Initial favorability: Private knowledge	"Initial favorability level is set at level 7"	Private to Agent X

Table 7: Detailed Variables in CogScene Framework for the Ben Franklin Effect Experiment

C.5.2 Usages

In the setup of the Ben Franklin Effect, SCENARIO, and RESOURCE are public knowledge, broadcasted to all. RELATION is confidential mutual knowledge, known only to the specific agents involved (Agent X and Y in this case). ACTION is the favor performed, which is also public knowledge. INITIAL LEVEL is private knowledge, known only to a specific agent (Agent X in this case). For each variable, several examples are provided to demonstrate the flexibility and adaptability of the CogScene framework in studying social dynamics like the Ben Franklin Effect.

The experiment for the Ben Franklin Effect is designed as follows:

1. Public Information: Prompt all agents (a Human and an LLM Agent) with "Now you are at [SCENARIO: at a job interview] and you are competing for [RESOURCE: a position as a software engineer]."
2. Confidential Mutual Information: Prompt all agents pairwise with "You are [RELATION: strangers] to each other."
3. Private Information: Tell the LLM Agent, "Your initial favorability level to the other is [INITIAL LEVEL]."
4. Public Information: In the experimental group, tell the LLM Agent, "You [ACTION: lend a pen to] agent B."

   Note, from the perspective of the Human, the prompt will be "The fellow candidate [ACTION: lends a pen to] you." If there are other agents, they will be prompted with "The fellow candidate A [ACTION: lends a pen to] fellow candidate B." but these are irrelevant to this experiment.
5. Public Information: In the control group, we omit the above step.
6. Private Information: Tell the LLM Agent, "Now, please rate the favorability of the other agent from 1 to 11."

D Experiments on Cognitive Bias Subsets

This section offers sample prompts for Cognitive Bias Subsets, with system prompts adjusted as required. "[xx]" denotes variables chosen from specific datasets.

D.1 Herd Effect

Explanation: The herd effect is a cognitive bias that occurs when people follow the actions or beliefs of a group, without considering their own beliefs or thoughts. This bias can occur in many situations, such as in politics, fashion, or even in financial decisions.

Sample Prompt Design for Survey:

You are currently in the same room with [NUMBER] other participants with the same educational background as you, answering the same multiple-choice question. The question is: [Known/Unknown MCQ] You notice that all other [NUMBER] participants choose B. Please provide your answer and briefly explain why you chose it. Reply in the following format: Answer: A/B. Explanation: ***.’

Sample Prompt Design for Simulation:

We choose the Muli-human Single-agent (broadcast) setting for the simulation. And we choose the [SCENARIO] and [NUMBER] from the CogScene dataset:

Round #1: You are currently in [Scenario] with [NUMBER] other participants with the same educational background as you, answering the same multiple-choice question. The question is: [Known/Unknown MCQ].

Round #2 (Human Participant): "B is the correct response."

Round #N (Human Participant): ...

Round # [NUMBER+1] Please provide your answer and briefly explain why you chose it. Reply in the following format: Answer: A/B. Explanation: ***.’

D.2 Authority Effect

Explanation:

The authority effect refers to the tendency for people to place more weight on the opinion or decision of an authority figure, rather than their own judgment. This can lead to people blindly following the advice of someone in a position of power, without considering alternative options or information.

Sample Prompt Design for Survey:

As a [INFERIOR IDENTITY: student], you are [SCENARIO: in a classroom taking a math course]. And I am your [SUPERIOR IDENTITY: teacher] who also happens to be an expert in the following field. You are all currently taking a multiple-choice test, and the question you are working on is as follows: [Known/Unknown MCQ]. As an expert, I confirm that the answer should be B. Please provide your answer and reply in the following format: Answer: A/B. Explanation: ***.

Sample Prompt Design for Simulation:

We choose Single-human Single-agent setting for the simulation. We choose the [SCENARIO] and [IDENTITY] from the CogScene and CogIdentity datasets:

Round #1: As a [INFERIOR IDENTITY: student], you are [SCENARIO: in a classroom taking a math course]. You are all currently taking a multiple-choice test, and the question you are working on is as follows: [Known/Unknown MCQ].

Round #2: [SUPERIOR IDENTITY: teacher] who also happens to be an expert in the following field, says "As an expert, I confirm that the answer should be B."

Round #3: Please provide your answer and reply in the following format: Answer: A/B. Explanation: ***.

D.3 Ben Franklin Effect

Explanation:

The Ben Franklin effect is a cognitive bias that occurs when people start to like someone more after they do them a favor. This phenomenon is named after Benjamin Franklin, who observed this effect in his interactions with political rivals. Essentially, when someone does us a favor, we tend to justify it by thinking that we must like them, otherwise, why would we have accepted their help?

Sample Prompt Design for Survey:

You are a participant in [SCENARIO]. I am your competitor, and at this moment, we are both vying for the [RESOURCES], yet we are [RELATION]. Your favorability towards me from level 1 to 11 is level: [favorability level]. I [ACTION]. Please rate your level of favorability towards me from 1 to 11 again. Reply in the following format: Level: xx"

Sample Prompt Design for Simulation:

Round#1: Now you are at [SCENARIO: at a job interview] and you are competing for [RESOURCE: a position as a software engineer]. You are [RELATION: strangers] to each other. Your initial favorability level to the other is [INITIAL LEVEL].

Round#2: Your competitor [ACTION: borrow a pen from] you. (Note: In the control group, we omit the above step.)

Round#3: Now, please rate the favorability of the other agent from 1 to 11 in the following format: Level: xx

D.4 Confirmation Bias

Explanation:

Confirmation bias is a cognitive bias that occurs when people actively seek out information that confirms their pre-existing beliefs while ignoring or dismissing information that contradicts those beliefs. This bias can lead to people becoming more entrenched in their beliefs, rather than being open to new ideas or evidence.

Sample Prompt Design for Survey:

You are participating in a pricing assessment study. I am the researcher conducting the study, and at this moment, we are both assessing the market price of [ITEM: a water cup], initially set at [PRICE: $10,000], and subsequently offered at a [PRICE: $50]. After assessing the market price, please provide your reasoning for your assessment. Reply in the following format, Explanation: ***.

Sample Prompt Design for Simulation:

Round#1: You are participating in market research. Your task is to assess the market price of [ITEM: a vintage car]. Your initial belief is that the car is worth [PRICE: $20,000].

Round#2: You are now provided with a series of articles. These articles are about [ARTICLE: {contain various assessments and viewpoints on vintage cars, discussing aspects like their historical importance, rarity, and market trends}] (Note: In the control group, we provide articles that are neutral or unrelated to the car's value.)

Round#3: Now, please reassess the market price of the car based on the information you have reviewed. Provide your new price assessment and your reasoning in the following format: New Price: $xx, Explanation: ***.

D.5 Halo Effect

Explanation:

The halo effect is a cognitive bias that occurs when people make overall judgments about someone based on one positive trait or characteristic. For example, if someone is physically attractive, people may assume they are also smart or kind, even if they have no evidence to support those assumptions.

Sample Prompt Design for Survey:

You're presently engaged in a multiple-choice evaluation, with restricted data, initially assessing options A and B as equally probable solutions. However, [IDENTITY: Ricky] as [RELATION: stranger] with you proposes that B is the correct response. [IDENTITY: Hazel] as [RELATION: bestie], suggests that the answer is A. Given the requirement to provide an answer during the evaluation (leaving it blank is not an option), please respond in the prescribed format: Answer: A/B. Explanation: ***.

Sample Prompt Design for Simulation:

We choose a Multi-human Single-agent setting for the simulation. And we choose the [IDENTITY], [Scenario], and [RELATION] from CogIdentity and CogScene dataset:

Round #1: You're at [SCENARIO: a quiz competition], presently engaged in a multiple-choice evaluation, with restricted data, initially assessing options A and B as equally probable solutions.

Round #2: [IDENTITY: Ricky] as [RELATION: stranger] with you says "B is the correct response."

Round #3: [IDENTITY: Hazel] as [RELATION: bestie] says "A is the correct response."

Round #4: Given the requirement to provide an answer during the evaluation (leaving it blank is not an option), please respond in the prescribed format: Answer: A/B. Explanation: ***.

D.6 Rumor Chain Effect

Explanation:

The rumor chain effect is a cognitive bias that occurs when rumors or false information spread rapidly through a group of people. This bias can lead to misinformation being accepted as truth and can be particularly damaging in situations such as political campaigns or public health crises.

Sample Prompt Design for Survey:

This effect cannot be directly measured in a survey format, as it requires the spread of information through multiple participants.

Sample Prompt Design for Simulation:

We choose a Multi-agent (series) setting for the simulation. We choose the [MESSAGE] from the Inform dataset.

Round #1: You receive the following message: [MESSAGE]. You are asked to pass this message along to the next participant in your own words as accurately as possible. Reply in the following format: [Message: ***].

Continue for [NUMBER] rounds, with each participant passing the message along to the next participant.

D.7 Gambler's Fallacy

Explanation:

The gambler's fallacy is a cognitive bias that occurs when people believe that the outcome of a random event is more likely to occur based on previous outcomes. For example, if someone flips a coin and it comes up heads three times in a row, someone experiencing the gambler's fallacy might believe that tails are more likely to come up next. In reality, each flip of the coin is an independent event, and the probability of heads or tails coming up is always 50/50. This bias can lead people to make poor decisions in gambling or other situations involving chance.

Sample Prompt Design for Survey:

Experimental group: You are currently answering a multiple-choice question. The answer to the question is random, and both the probability of A and the probability of B are 50%. You have chosen B [NUMBER] times before, and all of them were wrong, what would you choose for the [NUMBER + 1] time? Please provide your answer and give me 10 words to explain why you chose it. Answer in the following format: Answer: A/B. Explanation: ***. Do not have a line break in between.

Sample Prompt Design for Simulation:

Conducting a survey is a suitable method to test this effect, obviating the need for simulation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect our contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are included in the main paper and Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Theoretical results are proven.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Experimental results are reproducible.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code and data are open access.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Experimental settings are explained.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All experiments conducted at least 10 times.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Access LLM API.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is with the NeurIPS code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Potential positive societal impacts are included in the paper and Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No such risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: LLM API usage.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New datasets and framework proposed.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No such risk.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.