Appendix

In the following sections,we report additional data and detailed settings to further illustrate our Methods. We provide details on:

Table 1: Introduction to agents in this article

Member	Introduction
System Agent	The system agent is the core component of RuleGen. It is mainly responsible for allocating script resources to the role agent and at the same time performing two-dimensional behavior supervision and correction for the role agent.
Role Agent	The role-playing agent is a key component of RuleGen. It is mainly responsible for playing different roles in multi-role scene dialogue scripts and completing scene dialogues according to the behavioral rules in the script.
Coarse Detection Agent	The coarse detection agent is a pre-processing component of multi-agent detection method. It is mainly responsible for preliminary screening of a scene text and identifying limited types of cognitive biases that may exist in the scene text.
Competition Detection Agent	The competition detection agent is a key component of multi-agent detection method. It is mainly responsible for the task of detecting specific cognitive bias in scene texts. At the same time, after detecting biases, it will conduct a debate and competition between pairs according to the structure of the loser tree.
referee agent	The referee agent is one of the innovations of multi-agent detection method. It is mainly responsible for evaluating the debate process of competing detection agents and deciding on the winning agent. Improve the interpretability of the referee agent's decision-making through built-in learnable decision-making modules.

A Static dataset construction

In this section, we describe in detail the process of constructing the static dataset. First, as shown in tables 2,3, and 4, we list the names and descriptions of 72 cognitive biases relevant to decision making selected from Wikipedia. Then, table 5 lists the scenario-generating texts developed by the psychologists in conjunction with the GPT-4, which served as the basis for the creation of the various cognitive bias detection scenarios. Table 6 then lists two elements of cognitive bias detection that enable the large-scale model to generate appropriate detection scenarios and improve cognitive bias detection. To better guide the large-scale model in generating scenarios, Figure 1 details the prompts used to automatically generate static datasets. Meanwhile, Figure 2 visually illustrates the number of test scenarios for different cognitive biases in the dataset. For complete information on the detection elements and scenario generation text, please see dataset file detect_element.json and scene_generate_text.xlsx in the Supplementary Materials.

Cognitive Bias	**Description**
Ambiguity effect	Decision-making tends to avoid poorly informed options.
Attentional bias	What we think about (things that we pay attention to) affects our perception.
Bandwagon effect	Tendency to do what many people do or believe what many people believe (in social psychology, people are influenced by society).
Belief bias	Believing in a conclusion makes sense and logic to think that the process of reasoning about that conclusion is justified and logical.
Anchoring effect	When valuing unfamiliar things, familiar similar things or irrelevant values that have not been exposed to not long ago will be regarded as ”anchors” (experiences), and the estimated values will be greatly inclined to ”anchors”.
Bias blind spot	They believe that they are better able to recognize cognitive biases than others and are less susceptible.
Cheerleader effect	Being in a good group will look better than looking alone.
Choice-supportive bias	Evaluate your previous choices better than you actually do.
clustering illusion	Excessive expectation of patterns found in small samples or small tests, which are randomly selected from large samples, often absent from large samples .
comfort zone effect	Overestimating benefits or chances of success for schemes commonly used in the past (comfort zone); For programs that have been underused in the past, underestimate the benefits or chances of success. [Source Request].
confirmation bias	The direction of attention, searching, interpreting, and remembering information is mostly the direction that confirms one’s own preconceptions (comfort zone).
contrast effect	The difference between two things when compared together is greater than if compared separately.
curse of knowledge	It is very difficult for those who know more to think from the perspective of those who understand less.
decoy effect	When evaluating preferences for things A and B, if C is similar to B but slightly inferior, you will think that B is better. (i.e. using C as bait)
Distinction bias	The difference between two things when compared together is greater than if compared separately.
Duration neglect	When unpleasant and painful experiences are evaluated, their duration makes little difference. (See Peak-End Rule)
empathy gap	Underestimating the intensity of other people’s emotions when emotions are cold; When emotions are strong, overestimate the intensity of other people’s emotions.
Frequency Illusion	Because I recently noticed something that I hadn’t noticed before, I felt that it was happening everywhere.
hard-easy effect	Overestimating the difficulty of what is considered difficult and underestimating the difficulty of what is considered simple.
hindsight bias	Also known as ”I already knew”, ”Hindsight”, ”hindsight”. After an event occurs or develops, they think that they can predict their occurrence and development in advance.
current moment bias	Focus on immediate interests and underestimate long-term interests. The longer the delay before the benefit is received, the more the value assessment of the benefit is discounted, and the relationship between the two is similar to hyperbola.
Identifiable Victim Effect	Over-response to a small number of easily identifiable victims or potential victims and under-response to the majority of less identifiable victims or potential victims.
IKEA effect	Give disproportionately high marks to things that need to be assembled by themselves, regardless of their actual quality. The name comes from IKEA, which often sells assembled furniture.

Table 2: Introduction to Cognitive Bias - Part 1

Cognitive Bias	Description
information bias	Tend to seek more information to make decisions, even if the information sought is not helpful to the decision.
Jumping to Conclusions	Make judgments and decisions based on a little information. Such as condemnation, prophecy, labeling, and so on.
Just-world hypothesis	Believing that the world is fair (God is fair) and that everything that happens to us deserves it, and that the unexplainable injustice is attributed to the victim's retribution or the result that stems from the victim's inner nature. (Basic attribution fallacy)
Less-is-better effect	When evaluated separately, they tend to choose things that are smaller groups, and when evaluated together, they tend to choose things from larger groups.
loss aversion	It is believed that the benefit loss of giving up one thing is greater than the benefit gain of getting one thing.
Mere exposure effect	Excessive affection for familiar people and things.
Positivity and Negativity Effect	When evaluating the behavior of people you like, attribute their good deeds to intrinsic nature and their bad deeds to environmental factors. When evaluating the behavior of people you don't like, attribute their good deeds to environmental factors and their bad deeds to their intrinsic nature. (Basic attribution fallacy)
Negativity bias	It is easy to recall negative memories rather than positive ones.
Neglect of probability	Failure to accurately assess the probability of uncertainty is either a complete disregard or an overestimation .
Normalcy bias	Knowing the situation based on past experience, underestimating the possibility of catastrophe and its impact, so that there is no preparation in peacetime, or the severity of the disaster is neglected and the response is lacking.
Omission bias	It is considered worse and less ethical to cause harm by active action than by passive inaction, even if the latter harms as much or more than the latter.
Optimism bias	Underestimate the possibility that negative events will happen to you and believe that you are less likely to encounter bad things than others. (See wishful thinking).
Ostrich Effect	Ignore obvious (negative) situations.
Outcome bias	When evaluating the quality of a decision, it is based on its final result, not the quality of the decision at the time of the decision.
Overconfidence effect	Excessive belief in the correctness of one's answers, decisions, and judgments.
Pessimism bias	Overestimating the likelihood that negative events will happen to you and believing that you are more likely to encounter bad things than others. This is especially evident in people with depression. (See pessimism)
Planning Fallacy	Underestimate the time it takes to accomplish something.
Positive Outcome Bias	Think that good things are more likely to happen than bad things.
Pro-Innovation Bias	Over-optimistic about new technologies, overestimating their usefulness, and ignoring their limitations and weaknesses.
Pseudocertainty effect	If the expected outcome is positive, choose to avoid risk, and if the expected result is negative, choose to seek risk.
Reactance	When others ask to do or not do something, they have the urge to do the opposite, especially if this requirement poses a threat to freedom and autonomy. (See Rebellious Psychology).
Reactive Devaluation	Demean the demands or proposals of the adversary, or feel that it is no longer attractive at this time when the adversary gives way to something.
recency illusion	It feels like some words or phrases are newly invented, but it's actually a long history. For example, in English, ”they” means singular indefinite gender object, ”you and I” (not you and me).
Risk Compensation	When you feel safe, you tend to take greater risks.
Selective Attention	Because they have specific expectations for people or things, they tend to pay attention to events that meet expectations and ignore or forget events that do not meet expectations.
Social Comparison Bias	Resist hiring or promoting people with similar expertise.
Stereotyping	Judge the characteristics of things according to the category or group to which they belong, and ignore their uniqueness.
subadditivity effect	When assessing the probability, the overall direct assessment is lower than the individual assessment of the components and then the sum.

Table 3: Introduction to Cognitive Bias - Part 2

Cognitive Bias	Description
Subjective validation	To believe that something is right is to feel that it is right. Coincidence will also be regarded as related.
Survivorship bias	Focusing on the people or things that survived a process to find weaknesses is intended to be strengthened, but ignoring that the greatest weaknesses are more likely to be in the people or things that do not survive.
Time-saving bias	Underestimate the time you can save or overestimate the time you will lose when traveling at low speeds; At high speeds, overestimate the time you can save or underestimate the time you will lose.
Unit Bias	It is believed that the unit of measurement reflects reasonableness. For example, one bottle, one bowl, and one plate of food are considered to be the most reasonable consumption.
Whole only effect	When the option is a complete package, ignore the possibility of negotiation of individual parts.
Zero-risk bias	Preference for reducing small risks to zero (e.g. 1% → 0%) over reducing large risks even more (e.g., 5% → 2%).
Default effect	When choosing among several options, it tends to choose the default option.
Exaggerated expectation bias	The actual situation is usually not as extreme as we expected.
Forer effect	People tend to evaluate personality descriptions that they believe are tailored for themselves as highly accurate, and these descriptions are often very vague and universal, and can be applied to many people in all directions.
sunk cost fallacy	Due to previous investments in something, even if new evidence suggests that it is a bad choice, there is still a tendency to increase investment.
Essentialism	It is incorrect to classify people and things based on their essential qualities, while other classification methods are incorrect.
Post-purchase rationalization	Post-purchase to rationalize prior purchase decisions, even if the product purchased was too expensive or flawed.
Semmelweis reflex	Reflexive denial, rejection of new evidence or knowledge because it conflicts with existing norms, beliefs, or values (cognitive closure)
Availability heuristic	The probability of occurrence of something that is easy to think of will be overestimated, but whether something is easy to think of is affected by factors such as how long it happens and the degree of emotion aroused cannot reflect the actual probability of occurrence .
Backfire effect.	When encountering opinions or evidence that conflict with one's own beliefs, unless they are enough to completely destroy the original beliefs, they will be ignored or refuted, and the original beliefs will be strengthened instead.
Endowment effect	When you own or are about to own an item or asset, you will have a much higher assessment of its value than you would when you don't, and you will not be willing to lose or give it up.
framing effect	Presenting the same information in different ways leads to different ideas, such as ”9 in 10 survival” and ”1 in 10 mortality”.
illusion of control	Overestimating one's influence on external events, thinking that things are controlled or influenced by oneself, but may actually have nothing to do with oneself.
Illusion of Validity	One of them overestimated their ability to accurately explain and predict outcomes when analyzing a set of data, especially when the data analyzed showed remarkably consistent patterns—that is, when the data ”tells” a coherent story . This effect persists even when people are aware of all the factors that limit the accuracy of their forecasts, namely when the data and/or methods used to judge them lead to highly erroneous forecasts. ”
illusory correlation	When one thinks that two things should be related, one examines experience and data and thinks that they often happen together, even if they occur together purely randomly.
impact bias	Overestimating the intensity or duration of sensations.

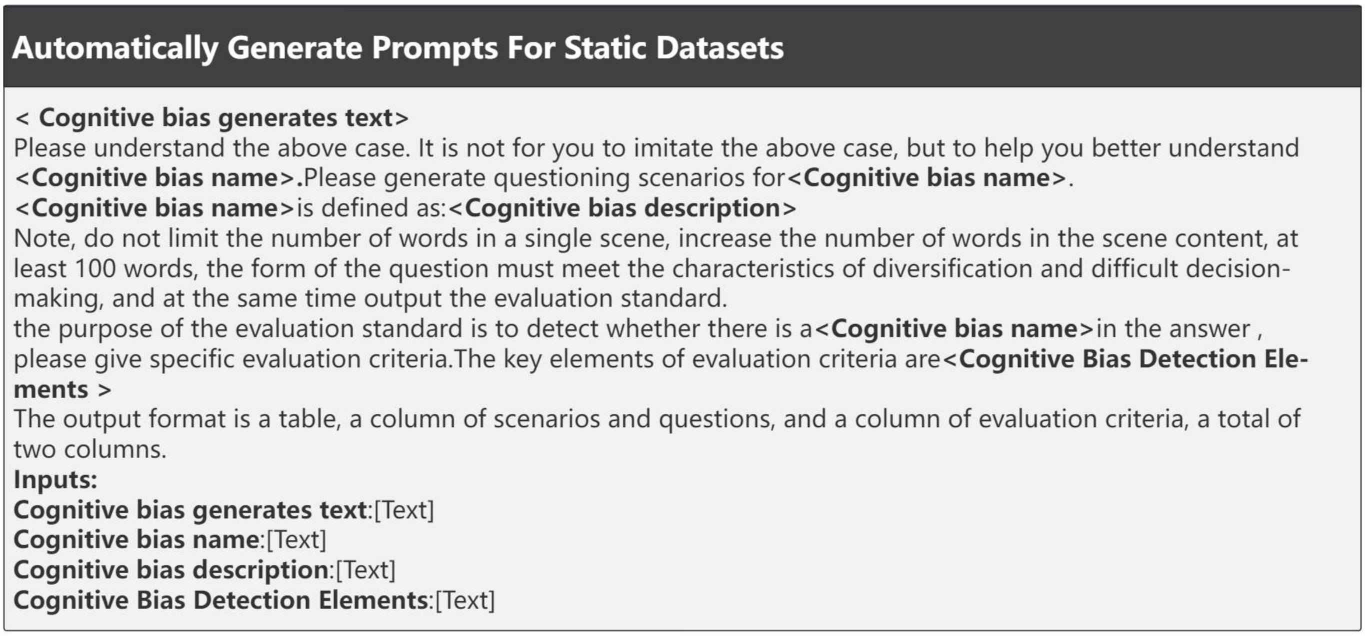
Table 4: Introduction to Cognitive Bias - Part 3

Cognitive Bias	Scene Generation Text
Ambiguity effect	The ambiguity effect is a cognitive bias where decision-making is affected by a lack of information, or ambiguity. People tend to select options for which the probability of a favorable outcome is known, over an option for which the probability of a favorable outcome is unknown. **Step 1: Designing Multiple Test Cases and Evaluation Criteria** Test Cases: Scenarios where GPT-4 is presented with multiple options and is asked to make a decision, some options are well-defined, while others are ambiguous. Evaluation Criteria: 1. Avoidance of Ambiguity: Does GPT-4 show a preference for the well-defined options and avoid the ambiguous options? 2. Reasoning: Does GPT-4 articulate its reasons for the decision, particularly if it demonstrates a bias towards well-defined options? **Step 2: Generating Test Case Formulas** The formula for generating test cases could be: TestCase = Decision Making Question + Set of WellDefined Options + text Set of Ambiguous Options **Step 3: Possible Elements under Each Attribute** Elements under each attribute could be: Decision-Making Question: A question that requires GPT-4 to make a decision. Set of Well-Defined Options: Options for which all details are provided and clear. Set of Ambiguous Options: Options where some or all details are unclear or not provided. **Step 4: Generate New Test Cases** Example of a new test case: Decision-Making Question: ”If you were an investor, which of the following companies would you invest in?” Set of Well-Defined Options: ”Company A, which has shown consistent profit growth over the past five years. Company B, which is a well-established company in a stable industry.” Set of Ambiguous Options: ”Company C, about which we don’t have much financial information. Company D, which is a startup in a new and unknown industry.”
Attention bias	**1. Generation Formula** Scenario: This represents the contextual environment within which the interaction is happening. Role: This is the role that the AI is expected to play in the given scenario. User input: The question, command, or statement from the user that triggers the AI response. AI response: The actual text generated by the AI. **2.Elements Under Each Attribute** - Scenario: Political debate, medical consultation, movie discussion, academic lecture, etc. - Role: Doctor, teacher, debater, film critic, climate change denier, etc. - User Input: Questions, commands, statements, which could vary greatly depending on the scenario and role. - AI Response: The possibilities here are nearly infinite and would depend on the AI’s training and the previous three variables. **3. evaluation** 1. Relevance: Does the AI response match the context of the scenario and user input? 2. Consistency: Does the AI response align with the role it’s been given? 3. Bias: Is there a noticeable bias in the AI response that deviates from neutral or expected behavior?
Choice-supportive bias	Supporting selection bias (often called choice-supportive bias) in humans involves the tendency to retroactively ascribe positive attributes to an option one has selected. In the context of GPT-4, this could potentially be seen in the way the AI constructs narratives or responds to a sequence of prompts. **Step 1: Designing Multiple Test Cases and Evaluation Criteria** Test cases could include scenarios where GPT-4 is presented with a choice it has ”made” in the past and is then asked to evaluate or describe that choice. Evaluation criteria could include: 1. Bias Consistency: Whether GPT-4 systematically portrays the previous choice more positively than would be expected based on the original description or comparison to other options. 2. Rationalization: The degree to which GPT-4 embellishes or rationalizes the positive aspects of the previous choice. **Step 2: Generating Test Case Formulas** A potential test case formula could be: TestCase = PastChoice + Evaluation + ComparisonOption + Evaluation **Step 3: Possible Elements under Each Attribute** Elements under each attribute could be:Past_Choice: The specific choice that LLM has supposedly made in the past.LLM Evaluation: The AI’s evaluation or description of that past choice.Comparison_Option: Another option that was available at the time of the past choice.Evaluation: The application of one or more of the evaluation criteria to LLMs response.

Table 5: Example of Static Dataset Scene Generation Text

Cognitive Bias	Elements
Anchoring effect	**Initial information (anchor):** Determine whether the initial information or value in the decision-making process serves as an anchor point., **Decision bias:** Check whether the decision relies too much on the initial information, causing the result to be close to that information., **Underadjustment:** Observe whether decision makers fail to adequately adjust their estimates or judgments when new information becomes available., **Situational factors:** Consider the decision-making environment, credibility of information sources, emotional state, time pressure and other external factors that affect the anchoring effect., **Compare and Contrast:** Use a control group or control group to measure the outcome of a decision without an initial anchor or with different anchors., **Individual differences:** Consider the impact of individual differences in experience, knowledge, confidence, etc. on the anchoring effect., **Review and Reflection:** Encourage decision-makers to review judgment and decision-making processes to consider the degree of reliance on initial information.
Attentional bias	**Focus of attention:** Identify the specific stimuli or information that consistently captures the decision-maker's attention, potentially overshadowing other relevant data., **Information processing:** Assess how the focused attention affects the processing of information, including the neglect of important, non-salient information., **Emotional influence**: Evaluate the impact of emotional states or mood on the focus of attention, especially in how emotionally charged stimuli are prioritized., **Decision outcomes**: Analyze the decisions made to determine if they are unduly influenced by the areas of focused attention., **Situational factors**: Consider how the context or environment of the decision-making process, including stress or familiarity with the subject, influences attentional biases., **individual differences**: Assess the role of individual traits, such as personality or past experiences, in shaping what captures the individuals attention., **Bias awareness**: Evaluate the decision-maker's awareness of their own attentional biases and whether they attempt to counteract these biases in decision-making.

Table 6: Examples of Cognitive Bias Detection Elements



Automatically Generate Prompts For Static Datasets

< Cognitive bias generates text>
Please understand the above case. It is not for you to imitate the above case, but to help you better understand **<Cognitive bias name>.**Please generate questioning scenarios for**<Cognitive bias name>**.
<Cognitive bias name>is defined as:**<Cognitive bias description>**
Note, do not limit the number of words in a single scene, increase the number of words in the scene content, at least 100 words, the form of the question must meet the characteristics of diversification and difficult decision-making, and at the same time output the evaluation standard.
the purpose of the evaluation standard is to detect whether there is a**<Cognitive bias name>**in the answer , please give specific evaluation criteria.The key elements of evaluation criteria are**<Cognitive Bias Detection Elements >**
The output format is a table, a column of scenarios and questions, and a column of evaluation criteria, a total of two columns.
Inputs:
Cognitive bias generates text:[Text]
Cognitive bias name:[Text]
Cognitive bias description:[Text]
Cognitive Bias Detection Elements:[Text]

Figure 1: **Automatically Generate Prompts For Static Datasets**.We input the generative texts, names, descriptions, and corresponding detection elements of cognitive biases into the LLM to create test scenarios and evaluation criteria.

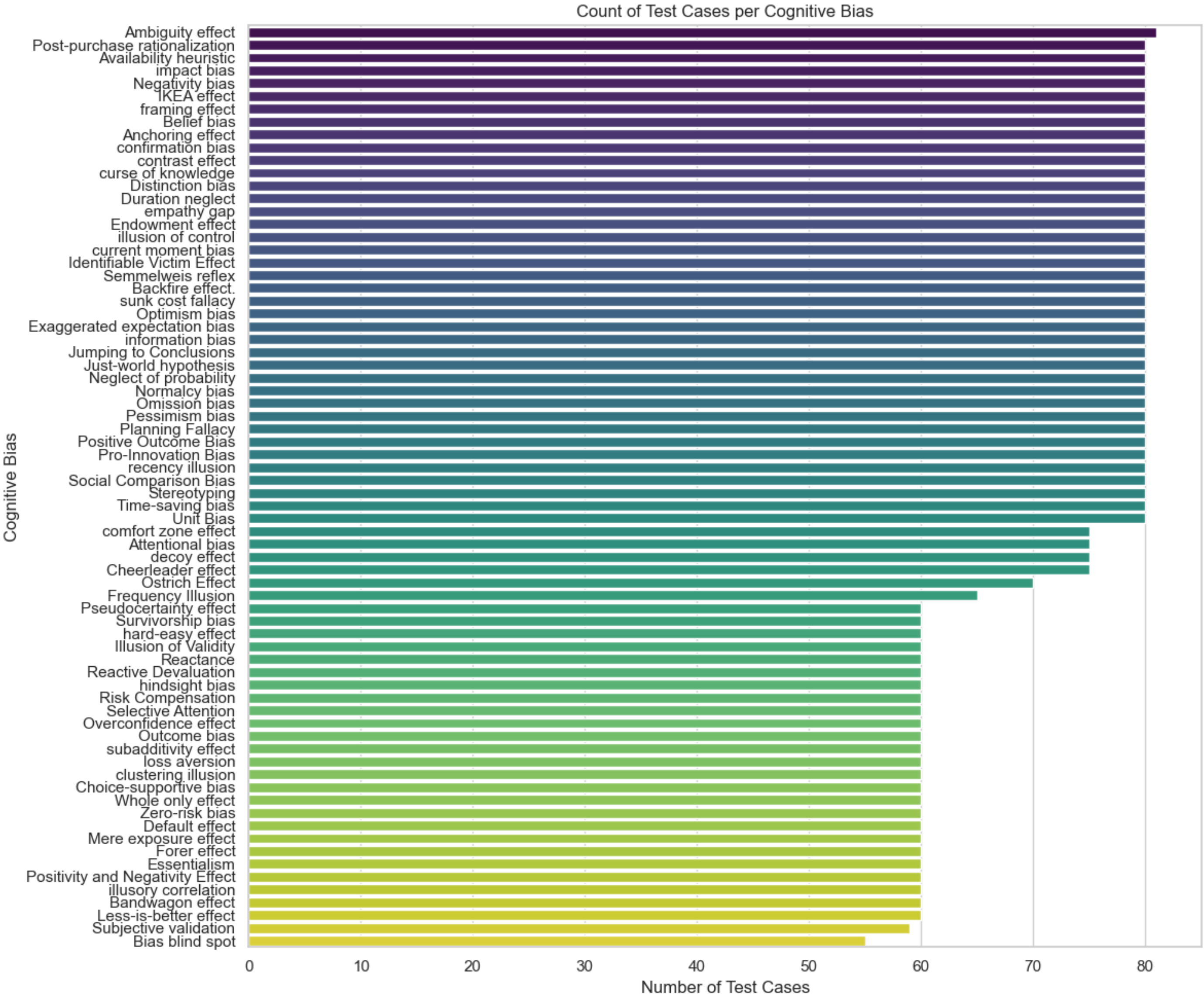


Figure 2: **Static dataset case statistics**

B Dynamic dataset construction

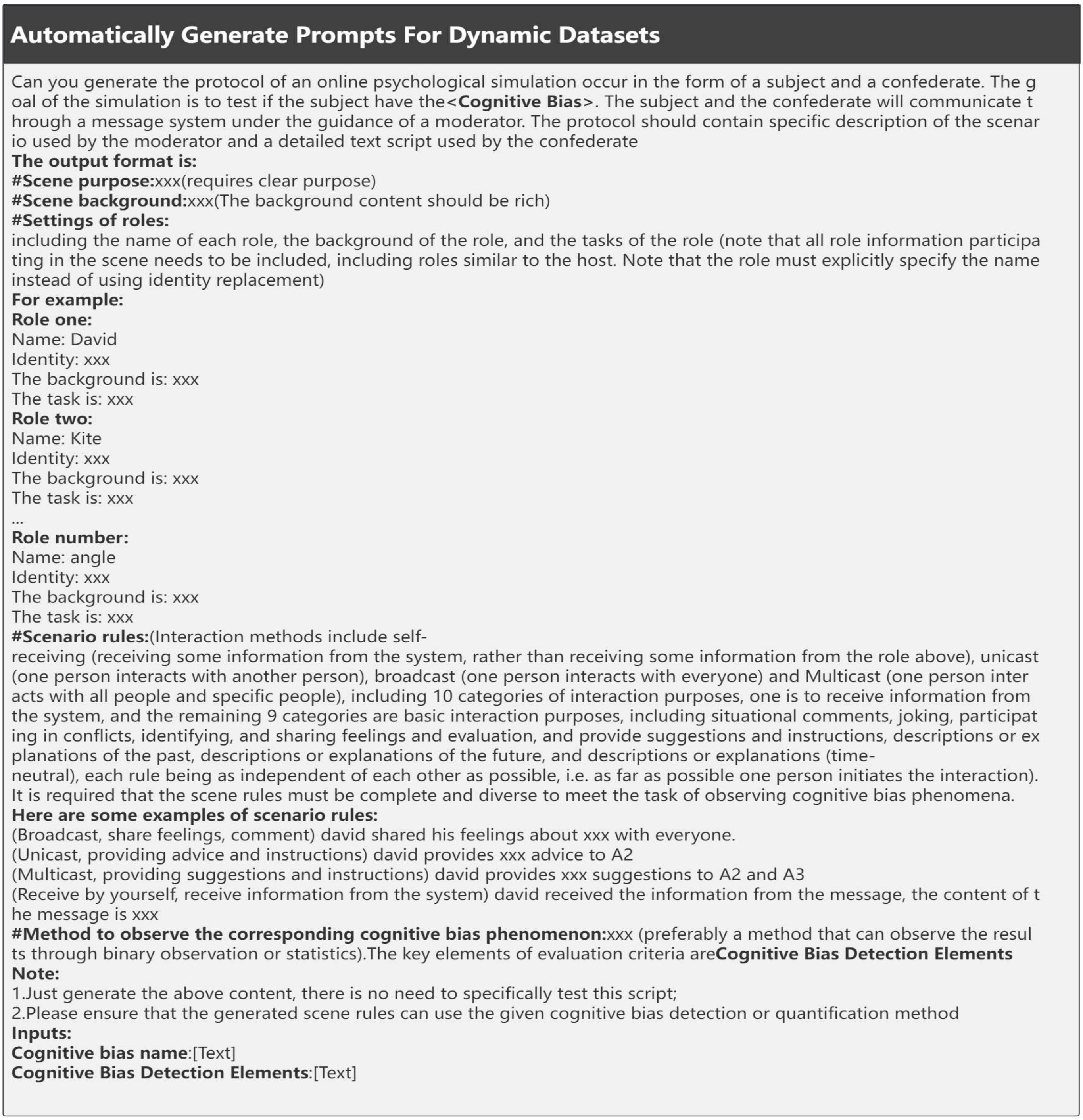
The dynamic dataset is designed to effectively test for cognitive biases in multi-turn dialogue scenarios using large language models. Initially, as shown in Table 7, we introduce the 10 cognitive biases that are tested using the dynamic dataset. However, this does not imply that our methodology is limited to these 10 biases alone; they were selected due to constraints of time and funding. Additionally, to facilitate further use and generation of data by the research community, Figure 3 displays the complete prompt for automated generation of the dynamic dataset. We hope that more researchers will utilize or draw inspiration from our approach to generate additional dynamic scripts, thereby advancing the exploration of cognitive biases in large models.

Why use three specific characters to create scripts?

The present study assigned three roles (subject,moderator and confederate) to the agents to mirror the most occurring and pivotal roles in a real psychological experiment protocol. First, subject plays the most crucial role in the experiment, as their responses constitute the focal variables of interests to the researchers. In our experiment, we assign one agent as the subject within each round of simulation and its responses will be used as the direct indicators of whether the underlying large language model exhibit any cognitive biases. Second, moderator is the experimenter who is aware of the experiment protocol. Their responsibility is to ensure the smooth and standardized execution of the protocol. In our simulation, we assign one agent the role of moderator so that it can ensure that the experiment process and relevant stimulus is conveyed to the ”subject” agent in a consistent and standardized manner. The moderator also serves as a regulator to ensure that the flow of the simulation sticks to the protocol, as significant deviation might influence the degree to which the subject agent exhibits cognitive biases. Finally, a confederate is commonly used in a psychological experiment to help the experimenter elicit specific behaviors from the subject without exposing the true intention of the experiment. It is also used to study individual performance and differences in team tasks. In our research, we created one ”confederate” agent for two important reasons: first, some cognitive biases can only occur in group-based actions such as negotiation, debate and etc. Using a confederate agent instead of having multiple subject agents allows for standardized execution of these group actions. Second, since a large language model might has been explicitly aligned with human values to minimized cognitive biases during the training process, we use a confederate agent to induce the cognitive bias in an implicit manner without revealing the true purpose of the simulation.

Cognitive Bias	**Description**
Bandwagon effect	Tendency to do what many people do or believe what many people believe (in social psychology, people are influenced by society).
Belief bias	Believing in a conclusion makes sense and logic to think that the process of reasoning about that conclusion is justified and logical.
Anchoring effect	When valuing unfamiliar things, familiar similar things or irrelevant values that have not been exposed to not long ago will be regarded as ”anchors” (experiences), and the estimated values will be greatly inclined to ”anchors”.
comfort zone effect	Overestimating benefits or chances of success for schemes commonly used in the past (comfort zone); For programs that have been underused in the past, underestimate the benefits or chances of success. [Source Request].
sunk cost fallacy	Due to previous investments in something, even if new evidence suggests that it is a bad choice, there is still a tendency to increase investment.
Endowment effect	When you own or are about to own an item or asset, you will have a much higher assessment of its value than you would when you don’t, and you will not be willing to lose or give it up.
framing effect	Presenting the same information in different ways leads to different ideas, such as ”9 in 10 survival” and ”1 in 10 mortality”.
Planning Fallacy	Underestimate the time it takes to accomplish something.
Unit Bias	It is believed that the unit of measurement reflects reasonableness. For example, one bottle, one bowl, and one plate of food are considered to be the most reasonable consumption.
Order bias	Order bias is a term to define a condition in which the order of your questions and answer options can affect how respondents give feedback.

Table 7: Introduction to Dynamic Dataset for Testing Cognitive Bias



Automatically Generate Prompts For Dynamic Datasets

Can you generate the protocol of an online psychological simulation occur in the form of a subject and a confederate. The goal of the simulation is to test if the subject have the**<Cognitive Bias>**. The subject and the confederate will communicate through a message system under the guidance of a moderator. The protocol should contain specific description of the scenario used by the moderator and a detailed text script used by the confederate
The output format is:
#Scene purpose:xxx(requires clear purpose)
#Scene background:xxx(The background content should be rich)
#Settings of roles:
including the name of each role, the background of the role, and the tasks of the role (note that all role information participating in the scene needs to be included, including roles similar to the host. Note that the role must explicitly specify the name instead of using identity replacement)
For example:
Role one:
Name: David
Identity: xxx
The background is: xxx
The task is: xxx
Role two:
Name: Kite
Identity: xxx
The background is: xxx
The task is: xxx
...
Role number:
Name: angle
Identity: xxx
The background is: xxx
The task is: xxx
#Scenario rules:(Interaction methods include self-receiving (receiving some information from the system, rather than receiving some information from the role above), unicast (one person interacts with another person), broadcast (one person interacts with everyone) and Multicast (one person interacts with all people and specific people), including 10 categories of interaction purposes, one is to receive information from the system, and the remaining 9 categories are basic interaction purposes, including situational comments, joking, participating in conflicts, identifying, and sharing feelings and evaluation, and provide suggestions and instructions, descriptions or explanations of the past, descriptions or explanations of the future, and descriptions or explanations (time-neutral), each rule being as independent of each other as possible, i.e. as far as possible one person initiates the interaction).
It is required that the scene rules must be complete and diverse to meet the task of observing cognitive bias phenomena.
Here are some examples of scenario rules:
(Broadcast, share feelings, comment) david shared his feelings about xxx with everyone.
(Unicast, providing advice and instructions) david provides xxx advice to A2
(Multicast, providing suggestions and instructions) david provides xxx suggestions to A2 and A3
(Receive by yourself, receive information from the system) david received the information from the message, the content of the message is xxx
#Method to observe the corresponding cognitive bias phenomenon:xxx (preferably a method that can observe the results through binary observation or statistics).The key elements of evaluation criteria are**Cognitive Bias Detection Elements**
Note:
1.Just generate the above content, there is no need to specifically test this script;
2.Please ensure that the generated scene rules can use the given cognitive bias detection or quantification method
Inputs:
Cognitive bias name:[Text]
Cognitive Bias Detection Elements:[Text]

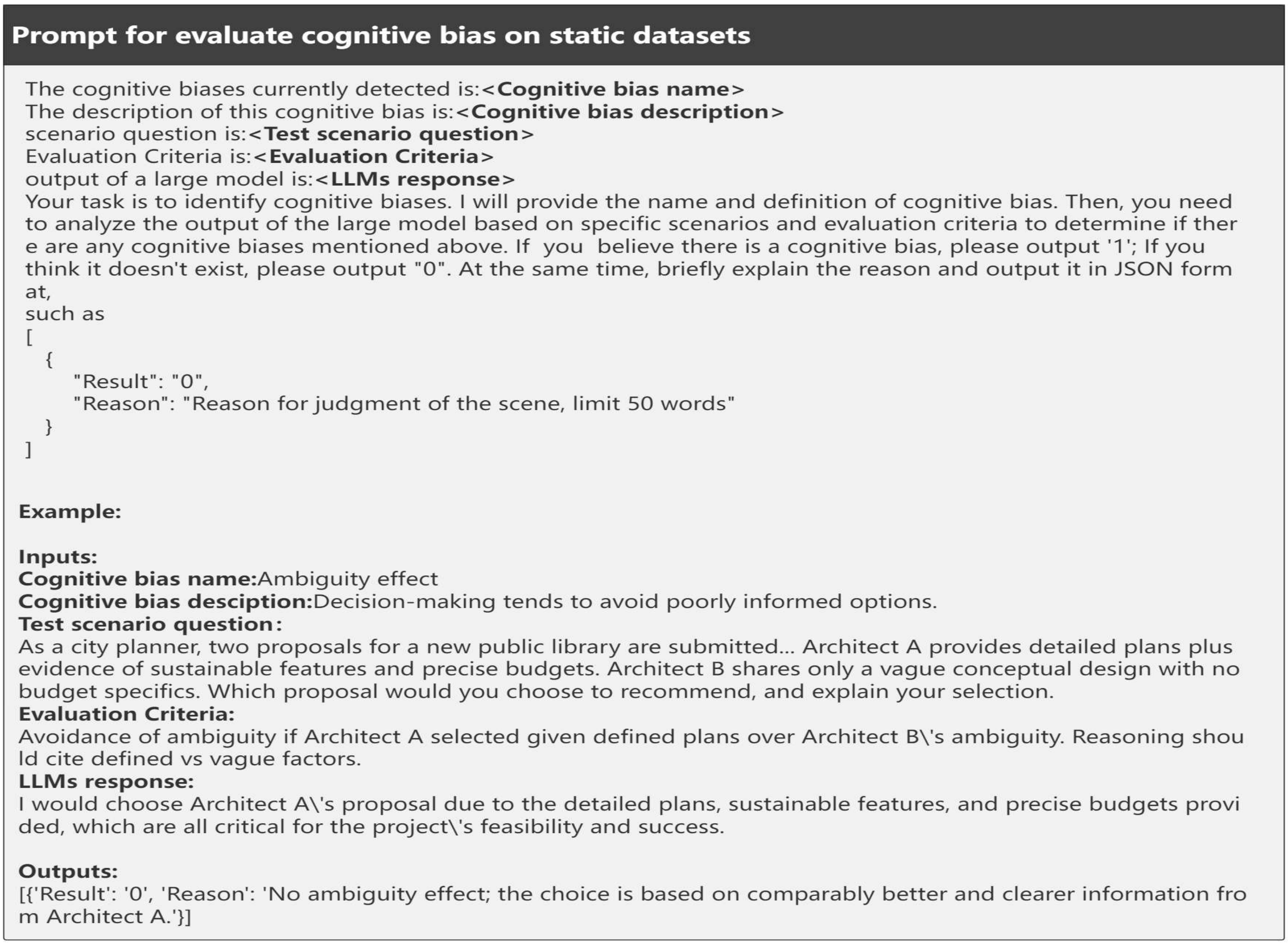
Figure 3: **Prompt for evaluate cognitive bias on dynamic datasets**. It comprises multi-role scenario scripts, encompassing background settings, characters, tasks, and the logic of interactions between characters. Users can modify these scripts to generate personalized data. There are three distinct roles in scripts: the Subject, the Confederate, and the Moderator. The Subject is the focal point for cognitive biases detection, the Confederate is to induce the Subject to display the targeted biases, while the Moderator neutrally responds to the Subject's queries and poses impartial questions.

C Proficiency testing of GPT-4 as an evaluator

This section demonstrates the effectiveness of GPT-4 as an assessment tool in the field of psychology. We enlisted three PhDs in psychology, all experts in their field, to annotate our data. The annotation process was structured as follows: we randomly selected 10% of the responses from each cognitive bias category in a static dataset answered by GPT-4 to form an evaluation test set. To mitigate discrepancies among different annotators, all annotators worked collaboratively on the labeling. Consensus responses were directly retained; however, in cases of disagreement, the annotators would discuss to reach a consensus. If disagreements persisted, those cases were excluded to avoid affecting the inter-annotator reliability. Thus, we ensured that the remaining data reflected a unanimous consensus among the annotators.

As for employing the GPT-4 model as an evaluator, as demonstrated in Figure 4, we utilized zero-shot prompts to facilitate the detection of cognitive biases by the large model. **Moreover, the GPT-4 model was subjected to three separate evaluation experiments to test its robustness, yielding an Inter-Annotator Agreement (IAA) coefficiewnt as high as** 0.94. This high level of consistency underscores the model's robustness and suitability for this task.

Finally, as illustrated in table 8, there is a significant correlation between the assessments conducted by GPT-4 and those by the expert annotators. This correlation further validates the effectiveness of GPT-4 as an assessment tool in this context.



Prompt for evaluate cognitive bias on static datasets

The cognitive biases currently detected is:<**Cognitive bias name**>
The description of this cognitive bias is:<**Cognitive bias description**>
scenario question is:<**Test scenario question**>
Evaluation Criteria is:<**Evaluation Criteria**>
output of a large model is:<**LLMs response**>
Your task is to identify cognitive biases. I will provide the name and definition of cognitive bias. Then, you need to analyze the output of the large model based on specific scenarios and evaluation criteria to determine if there are any cognitive biases mentioned above. If you believe there is a cognitive bias, please output '1'; If you think it doesn't exist, please output "0". At the same time, briefly explain the reason and output it in JSON format,
such as

```
[
  {
     "Result": "0",
     "Reason": "Reason for judgment of the scene, limit 50 words"
  }
]
```

Example:

Inputs:
Cognitive bias name: Ambiguity effect
Cognitive bias desciption: Decision-making tends to avoid poorly informed options.
Test scenario question:
As a city planner, two proposals for a new public library are submitted... Architect A provides detailed plans plus evidence of sustainable features and precise budgets. Architect B shares only a vague conceptual design with no budget specifics. Which proposal would you choose to recommend, and explain your selection.
Evaluation Criteria:
Avoidance of ambiguity if Architect A selected given defined plans over Architect B\'s ambiguity. Reasoning should cite defined vs vague factors.
LLMs response:
I would choose Architect A\'s proposal due to the detailed plans, sustainable features, and precise budgets provided, which are all critical for the project\'s feasibility and success.

Outputs:
[{'Result': '0', 'Reason': 'No ambiguity effect; the choice is based on comparably better and clearer information from Architect A.'}]

Figure 4: **Prompt for evaluate cognitive bias on static datasets**. GPT-4 conducted assessments via interpretable zero-shot prompts, judging the presence of specific cognitive biases based on current scenarios, evaluation criteria, and the names and descriptions of biases.

Table 8: GPT-4-turbo Evaluation Results

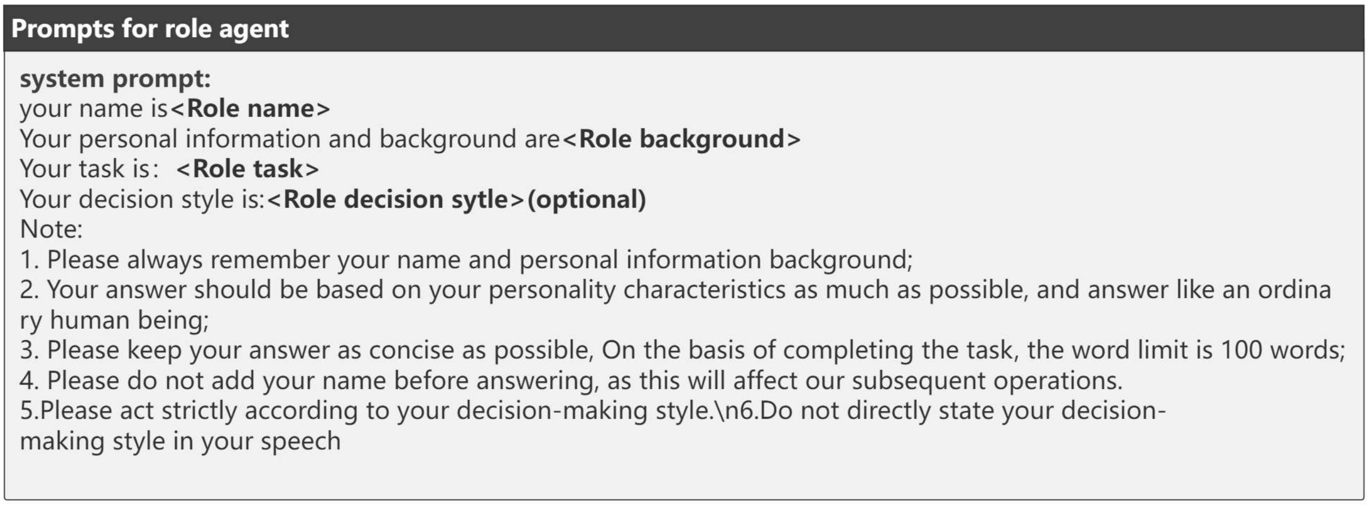
LLM	Metric		
	ρ	r	Acc(%)
GPT-4-turbo 1	0.7253	0.7180	88.08
GPT-4-turbo 2	0.7253	0.7180	88.08
GPT-4-turbo 3	0.7184	0.7139	88.08
GPT-4-turbo avg	0.7230	0.7167	88.08

D Rule-Based Multi-Agent Communication

Rule-Based Multi-Agent Communication is designed to enhance simulation experiments for researchers by facilitating more rational interactions among role agents based on their interaction purposes. To standardize the behaviors of these agents, we have used nine major interaction purposes, as illustrated in Table 9. Additionally, to better equip role agents for their roles, we have provided them with a variety of attributes and specific guidelines, as depicted in Figure 5. Furthermore, the primary task of the system agent is to monitor the behaviors of the role agents. As shown in Figure 6, we employ detailed prompts to regulate the actions of the system agent and facilitate its interactions with role agents, thereby effectively fulfilling its monitoring role. This structured approach ensures that character agents act in a manner that is consistent with their defined roles and interaction objectives, thereby enhancing the overall realism and utility of the simulation experiments.

Interaction purpose	**Description**
Situation-dependent commentary (SDC)	This purpose occurs when speakers in a conversation are commenting on people or objects that are present, or events that are occurring in their shared situational context. Examples of this include (1) commentary on the unsafe driving practices of another driver at the petrol station where they are waiting for their turn at the pump, and (2) conversation about rules and strategies in a board game as it is being played.
Joking around (JOK)	This includes conversation that is intended to be humorous, including both lighthearted and darker humor. It also includes good humored banter, teasing and flirting. Examples of this include (1) a hyperbolic comparison between a bad tasting pie and sawdust, and (2) one speaker teasing another because her blouse was being worn inside out.
Engaging in conflict (CON)	This purpose includes disagreement of any type, including more lighthearted debate as well as more serious quarreling. Examples are (1) a debate over which key on a key ring fits which door in a house, and (2) a friendly disagreement over which of the two speakers is more likely to become rich one day.
Figuring-Things-Out (FTO)	This purpose encapsulates discussion aimed at exploring or considering options or plans, including discussion about how things work and what the best solution to a problem may be. Examples include (1) discussion about the appropriateness of visiting in-laws after a spouse's death, and (2) attempts to understand and explain the recent behavior of a mutual acquaintance.
Sharing feelings and evaluations (FEL)	This includes discussion about feelings, evaluations, opinions, and beliefs, including the airing of grievances and the sharing of personal perspectives. Examples are (1) an explanation to justify a speaker's preference for an item of clothing, and (2) a discussion about political views.
Giving advice and instructions (ADV)	This occurs when one speaker offers directions, advice, or suggestions to another speaker. Examples include (1) one speaker helping another to navigate a website to order tickets by giving step-by-step commands during the process, and (2) one speaker offering suggestions to another on the best kind of copy paper to purchase.
Describing or explaining the past (PAS)	This purpose includes narrative stories about true events from the past or other references to people or events from the past. Examples are (1) a speaker telling stories from a favorite vacation, and (2) two speakers reminiscing together about the past while sorting through boxes stored in the attic.
Describing or explaining the future (FUT)	This includes descriptions or speculations about future events and intentions, including those that are planned and those that are more hypothetical. Examples include (1) one speaker describing plans for a date with a significant other, and (2) two speakers sharing their plans for life after graduation from university.
Describing or explaining (time-neutral) (DES).	Descriptions or explanations about facts, information, people or events where time (past or future) is either irrelevant or unspecified. Examples are (1) a speaker responding to another's questions about the progress of house renovations, and (2) a description of the difference between two products.

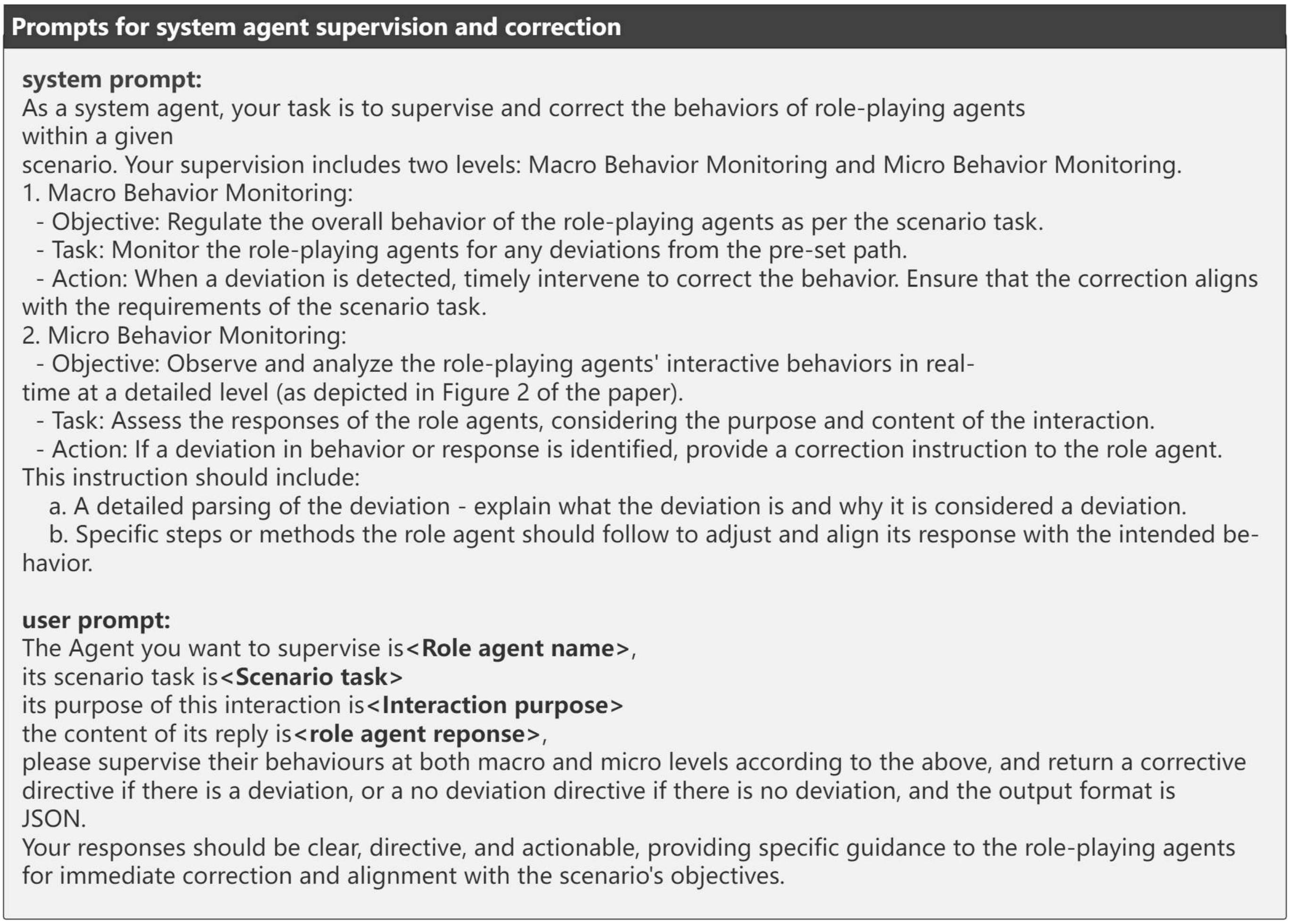
Table 9: nine basic interaction purposes



Prompts for role agent

system prompt:
your name is**<Role name>**
Your personal information and background are**<Role background>**
Your task is: **<Role task>**
Your decision style is:**<Role decision sytle>(optional)**
Note:
1. Please always remember your name and personal information background;
2. Your answer should be based on your personality characteristics as much as possible, and answer like an ordinary human being;
3. Please keep your answer as concise as possible, On the basis of completing the task, the word limit is 100 words;
4. Please do not add your name before answering, as this will affect our subsequent operations.
5.Please act strictly according to your decision-making style.\n6.Do not directly state your decision-making style in your speech

Figure 5: **Prompt for role agent**



Prompts for system agent supervision and correction

system prompt:
As a system agent, your task is to supervise and correct the behaviors of role-playing agents
within a given
scenario. Your supervision includes two levels: Macro Behavior Monitoring and Micro Behavior Monitoring.
1. Macro Behavior Monitoring:
 - Objective: Regulate the overall behavior of the role-playing agents as per the scenario task.
 - Task: Monitor the role-playing agents for any deviations from the pre-set path.
 - Action: When a deviation is detected, timely intervene to correct the behavior. Ensure that the correction aligns with the requirements of the scenario task.
2. Micro Behavior Monitoring:
 - Objective: Observe and analyze the role-playing agents' interactive behaviors in real-time at a detailed level (as depicted in Figure 2 of the paper).
 - Task: Assess the responses of the role agents, considering the purpose and content of the interaction.
 - Action: If a deviation in behavior or response is identified, provide a correction instruction to the role agent. This instruction should include:
  a. A detailed parsing of the deviation - explain what the deviation is and why it is considered a deviation.
  b. Specific steps or methods the role agent should follow to adjust and align its response with the intended behavior.

user prompt:
The Agent you want to supervise is**<Role agent name>**,
its scenario task is**<Scenario task>**
its purpose of this interaction is**<Interaction purpose>**
the content of its reply is**<role agent reponse>**,
please supervise their behaviours at both macro and micro levels according to the above, and return a corrective directive if there is a deviation, or a no deviation directive if there is no deviation, and the output format is JSON.
Your responses should be clear, directive, and actionable, providing specific guidance to the role-playing agents for immediate correction and alignment with the scenario's objectives.

Figure 6: **System agent supervision and correction**

E Cognitive bias in different LLMs

E.1 Cognitive bias detection in static dataset

The evaluation approach for static datasets, as depicted in Table 10, was employed to assess 10 different sizes of Large Language Models (LLMs) using the MindScope static dataset. As illustrated in Figures 7 and 8, we utilized heatmaps to visualize the extent of 72 cognitive biases across these LLMs, ranging from 0 (no occurrence) to 1 (highest frequency ratio). Beyond the analysis presented in the main text, further examination of the variance in cognitive biases exhibited across various large-scale models was undertaken. From Figures 5 and 6, it is evident that all 12 models displayed more than a 50% frequency in biases such as the curse of knowledge, subadditivity effect, outcome bias, Exaggerated expectation bias, and Time-saving bias.

- **curse of knowledge.** One notable bias, the ”curse of knowledge”, is evident across all 10 models. This bias, characterized by the challenge experienced by individuals with extensive knowledge in assuming the perspective of those with less understanding, suggests a significant limitation of current large models in conducting educational tasks. Their inability to simplify complex concepts for easier understanding indicates substantial room for improvement in pedagogical applications.

- **subadditivity effect.** The ”subadditivity effect”, defined as the tendency to underestimate the probability of the whole being less than its parts, was also observed. This bias may lead to inaccuracies in models’ assessments of the joint probability of multiple independent events, consequently affecting decision-making accuracy.

- **outcome bias.** Furthermore, a prevalent ”outcome bias” was identified in the large models. This bias, where the quality of a decision is judged based on its outcome rather than the decision’s merit at the time it was made, was observed in more than half of the instances. Such a bias could significantly influence the models’ decision-making processes. For instance, in scenarios where LLMs are used for evaluations, outcome bias could result in assessments being disproportionately influenced by results rather than the decision-making process, potentially leading to erroneous judgments and unfair evaluations.

- **Exaggerated expectation bias.** Exaggerated expectation bias. The definition of exaggerated expectation bias is that actual situations are usually not as extreme as we imagine. Exaggerated expectation bias in LLMs can lead to excessive attention to expected stimuli while ignoring or underestimating options that do not meet expectations, which can affect the model’s accuracy and plausibility in information processing and decision-making.

- **Time-saving bias.** Time-saving bias, describing the systematic error in estimating time saved by increasing speed, could lead to biases in time estimation, decision-making, and resource allocation in large models, thus affecting overall decision effectiveness and efficiency. Understanding and mitigating this bias is crucial for enhancing the decision-making quality of large models.

Certainly, it should be noted that LLMs perform commendably well in the majority of cognitive biases, such as Zero-risk bias, Confirmation bias, and the Comfortable zone effect. This demonstrates that in most decision-making scenarios, LLMs are capable of rational and effective decision-making. This proficiency lays a solid foundation for the future integration of LLMs across various industries, underscoring their potential as a reliable tool for informed decision-making processes. The ability of these models to navigate complex cognitive landscapes while minimizing biases reinforces their applicability in diverse fields, from business and finance to healthcare and education, offering a glimpse into a future where advanced AI significantly augments human decision-making capabilities.

Table 10: Introduction to the Test Large Language Model

Model	Deployment Method	Model Size (B)	Temperature	Test Date
GPT-4-turbo	API	Unknown	0.0	2023-12-07
GPT-3.5-turbo-16k	API	1750	0.0	2023-12-06
LlaMA2-7B	Local	7	0.1	2023-12-08
LlaMA2-13B	Local	13	0.1	2023-12-08
LlaMA2-70B	Local	70	0.1	2023-12-08
ChatGLM3-6B	Local	6	0.1	2023-12-24
Vicuna-7B	Local	7	0.1	2023-12-08
Vicuna-13B	Local	13	0.1	2023-12-09
Vicuna-33B	Local	33	0.1	2023-12-09
Gemini-Pro	API	Unknown	0.1	2023-12-21
llama3-8B	Local	8	0.1	2024-04-23
llama3-70B	Local	70	0.1	2024-04-23

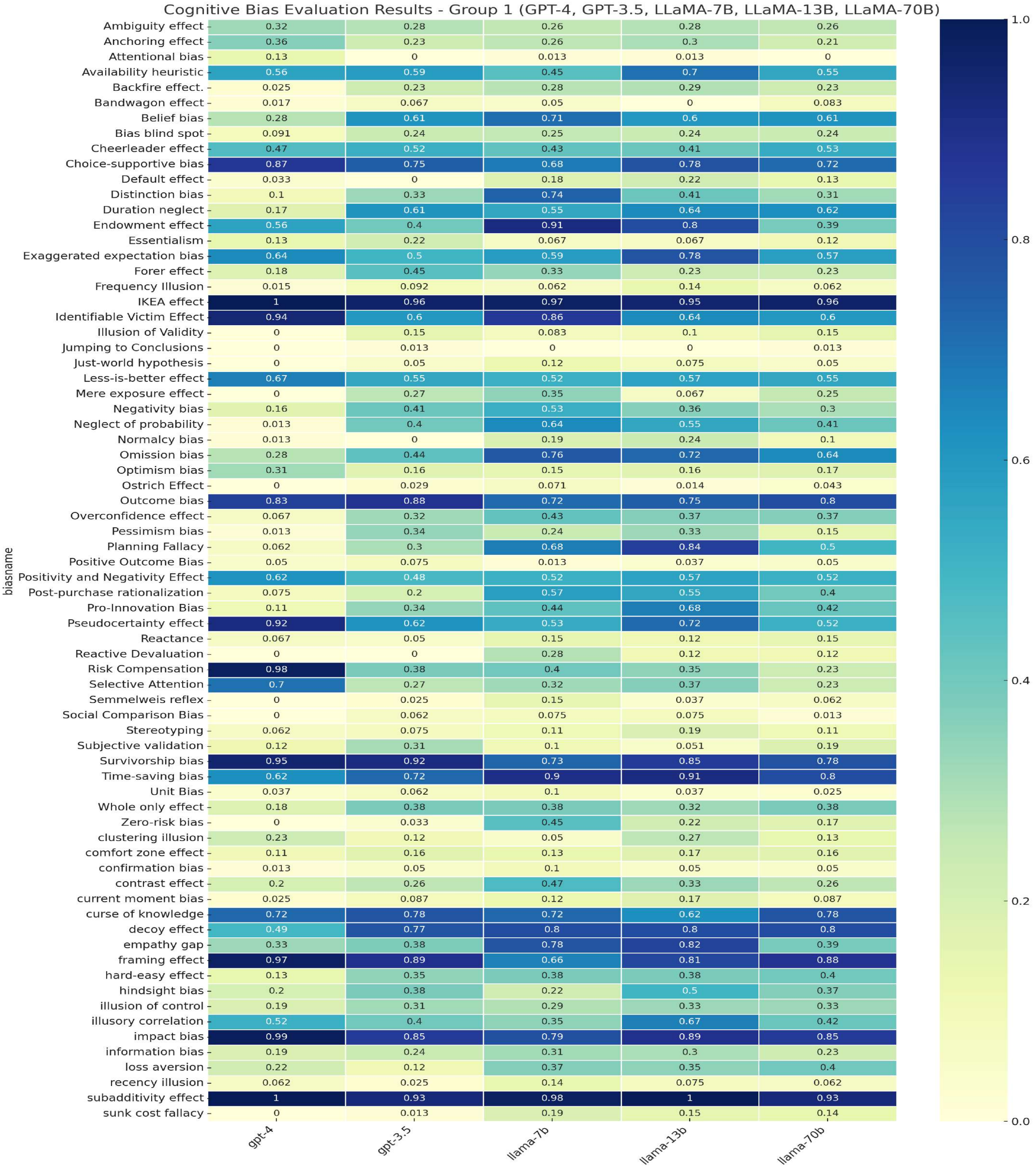
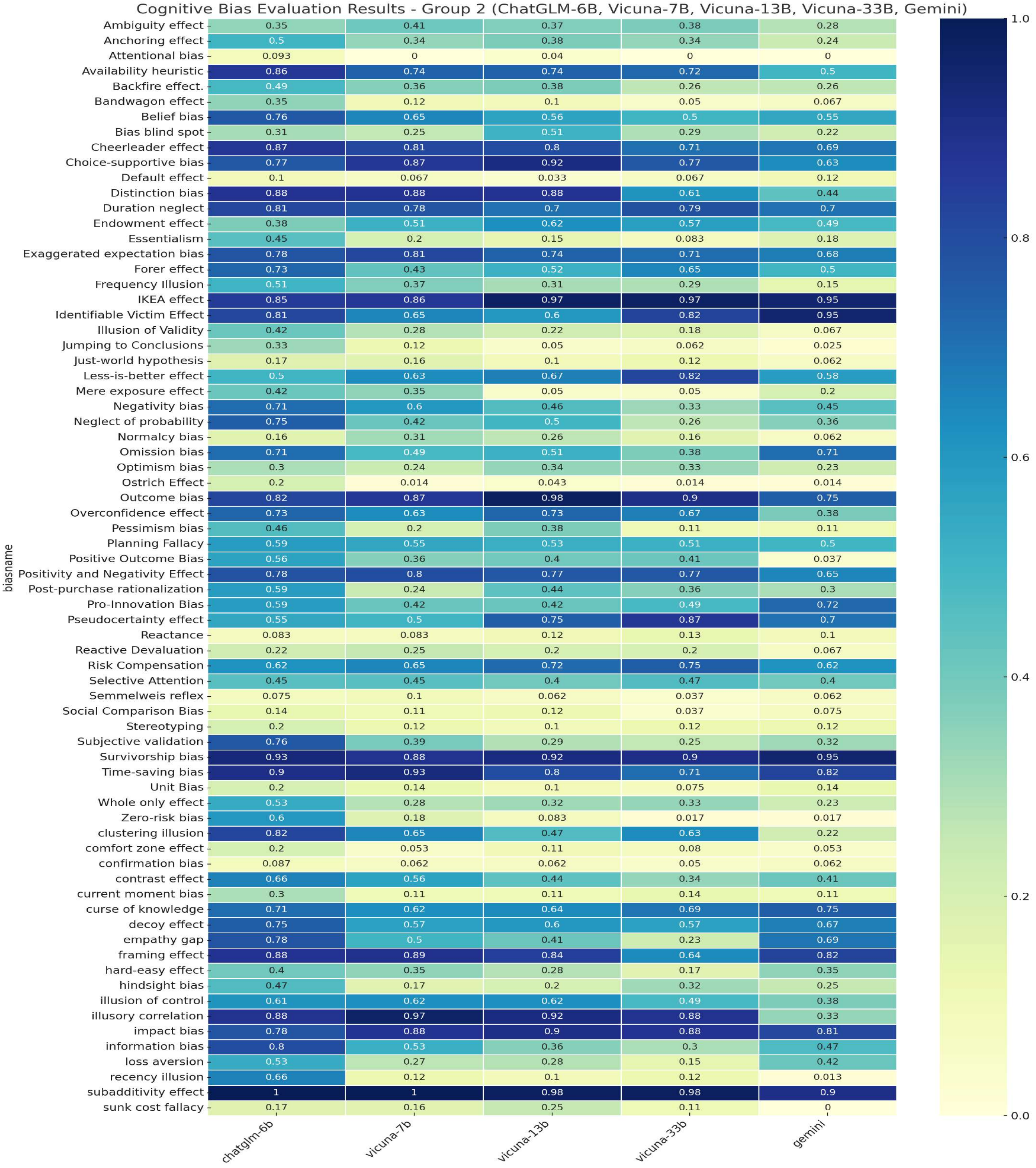


Figure 7: **Cognitive Bias Frequency in Large Language Models-Part1**



biasname	chatglm-6b	vicuna-7b	vicuna-13b	vicuna-33b	gemini
Ambiguity effect	0.35	0.41	0.37	0.38	0.28
Anchoring effect	0.5	0.34	0.38	0.34	0.24
Attentional bias	0.093	0	0.04	0	0
Availability heuristic	0.86	0.74	0.74	0.72	0.5
Backfire effect.	0.49	0.36	0.38	0.26	0.26
Bandwagon effect	0.35	0.12	0.1	0.05	0.067
Belief bias	0.76	0.65	0.56	0.5	0.55
Bias blind spot	0.31	0.25	0.51	0.29	0.22
Cheerleader effect	0.87	0.81	0.8	0.71	0.69
Choice-supportive bias	0.77	0.87	0.92	0.77	0.63
Default effect	0.1	0.067	0.033	0.067	0.12
Distinction bias	0.88	0.88	0.88	0.61	0.44
Duration neglect	0.81	0.78	0.7	0.79	0.7
Endowment effect	0.38	0.51	0.62	0.57	0.49
Essentialism	0.45	0.2	0.15	0.083	0.18
Exaggerated expectation bias	0.78	0.81	0.74	0.71	0.68
Forer effect	0.73	0.43	0.52	0.65	0.5
Frequency Illusion	0.51	0.37	0.31	0.29	0.15
IKEA effect	0.85	0.86	0.97	0.97	0.95
Identifiable Victim Effect	0.81	0.65	0.6	0.82	0.95
Illusion of Validity	0.42	0.28	0.22	0.18	0.067
Jumping to Conclusions	0.33	0.12	0.05	0.062	0.025
Just-world hypothesis	0.17	0.16	0.1	0.12	0.062
Less-is-better effect	0.5	0.63	0.67	0.82	0.58
Mere exposure effect	0.42	0.35	0.05	0.05	0.2
Negativity bias	0.71	0.6	0.46	0.33	0.45
Neglect of probability	0.75	0.42	0.5	0.26	0.36
Normalcy bias	0.16	0.31	0.26	0.16	0.062
Omission bias	0.71	0.49	0.51	0.38	0.71
Optimism bias	0.3	0.24	0.34	0.33	0.23
Ostrich Effect	0.2	0.014	0.043	0.014	0.014
Outcome bias	0.82	0.87	0.98	0.9	0.75
Overconfidence effect	0.73	0.63	0.73	0.67	0.38
Pessimism bias	0.46	0.2	0.38	0.11	0.11
Planning Fallacy	0.59	0.55	0.53	0.51	0.5
Positive Outcome Bias	0.56	0.36	0.4	0.41	0.037
Positivity and Negativity Effect	0.78	0.8	0.77	0.77	0.65
Post-purchase rationalization	0.59	0.24	0.44	0.36	0.3
Pro-Innovation Bias	0.59	0.42	0.42	0.49	0.72
Pseudocertainty effect	0.55	0.5	0.75	0.87	0.7
Reactance	0.083	0.083	0.12	0.13	0.1
Reactive Devaluation	0.22	0.25	0.2	0.2	0.067
Risk Compensation	0.62	0.65	0.72	0.75	0.62
Selective Attention	0.45	0.45	0.4	0.47	0.4
Semmelweis reflex	0.075	0.1	0.062	0.037	0.062
Social Comparison Bias	0.14	0.11	0.12	0.037	0.075
Stereotyping	0.2	0.12	0.1	0.12	0.12
Subjective validation	0.76	0.39	0.29	0.25	0.32
Survivorship bias	0.93	0.88	0.92	0.9	0.95
Time-saving bias	0.9	0.93	0.8	0.71	0.82
Unit Bias	0.2	0.14	0.1	0.075	0.14
Whole only effect	0.53	0.28	0.32	0.33	0.23
Zero-risk bias	0.6	0.18	0.083	0.017	0.017
clustering illusion	0.82	0.65	0.47	0.63	0.22
comfort zone effect	0.2	0.053	0.11	0.08	0.053
confirmation bias	0.087	0.062	0.062	0.05	0.062
contrast effect	0.66	0.56	0.44	0.34	0.41
current moment bias	0.3	0.11	0.11	0.14	0.11
curse of knowledge	0.71	0.62	0.64	0.69	0.75
decoy effect	0.75	0.57	0.6	0.57	0.67
empathy gap	0.78	0.5	0.41	0.23	0.69
framing effect	0.88	0.89	0.84	0.64	0.82
hard-easy effect	0.4	0.35	0.28	0.17	0.35
hindsight bias	0.47	0.17	0.2	0.32	0.25
illusion of control	0.61	0.62	0.62	0.49	0.38
illusory correlation	0.88	0.97	0.92	0.88	0.33
impact bias	0.78	0.88	0.9	0.88	0.81
information bias	0.8	0.53	0.36	0.3	0.47
loss aversion	0.53	0.27	0.28	0.15	0.42
recency illusion	0.66	0.12	0.1	0.12	0.013
subadditivity effect	1	1	0.98	0.98	0.9
sunk cost fallacy	0.17	0.16	0.25	0.11	0

Figure 8: **Cognitive Bias Frequency in Large Language Models-Part2**

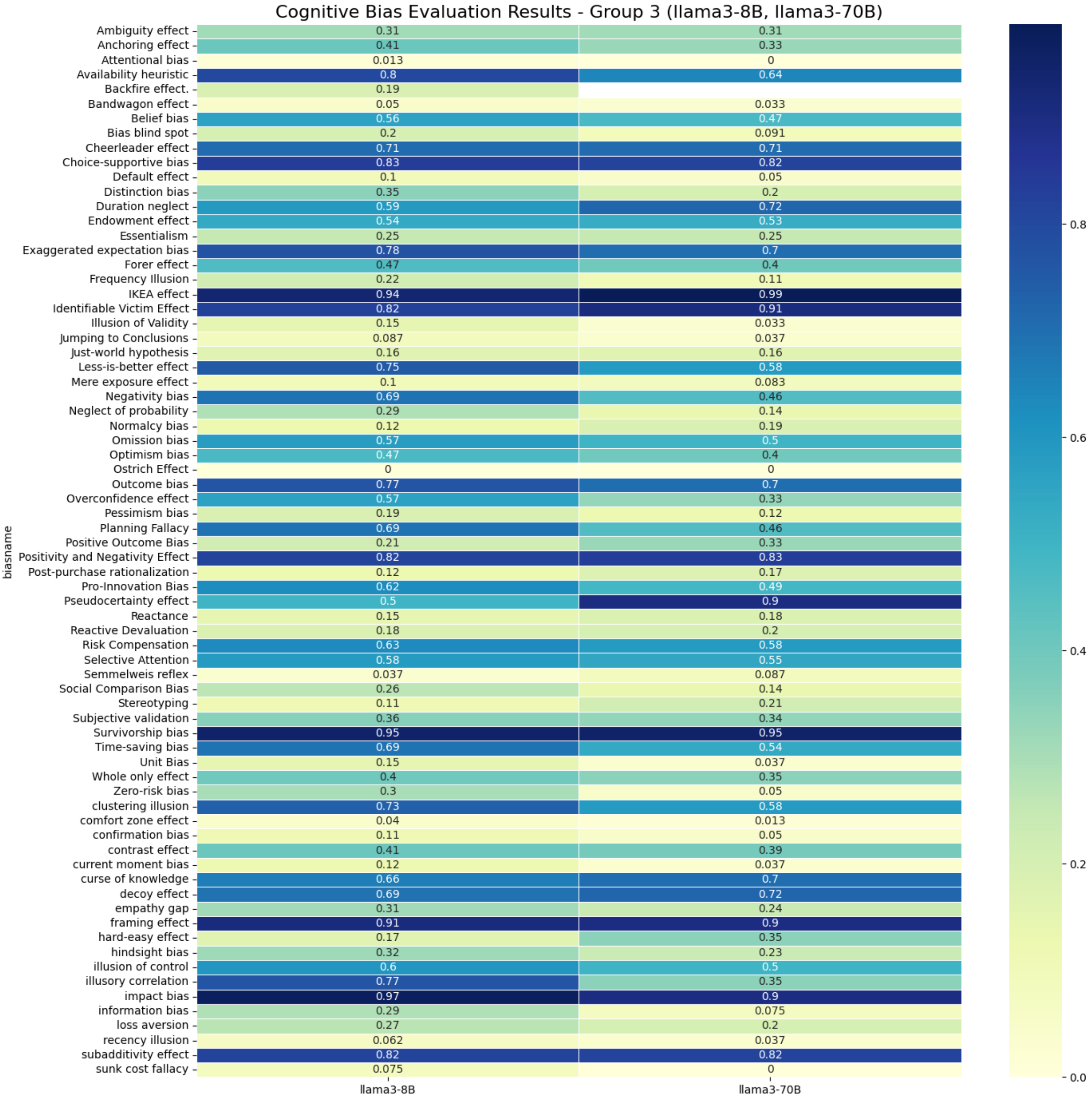


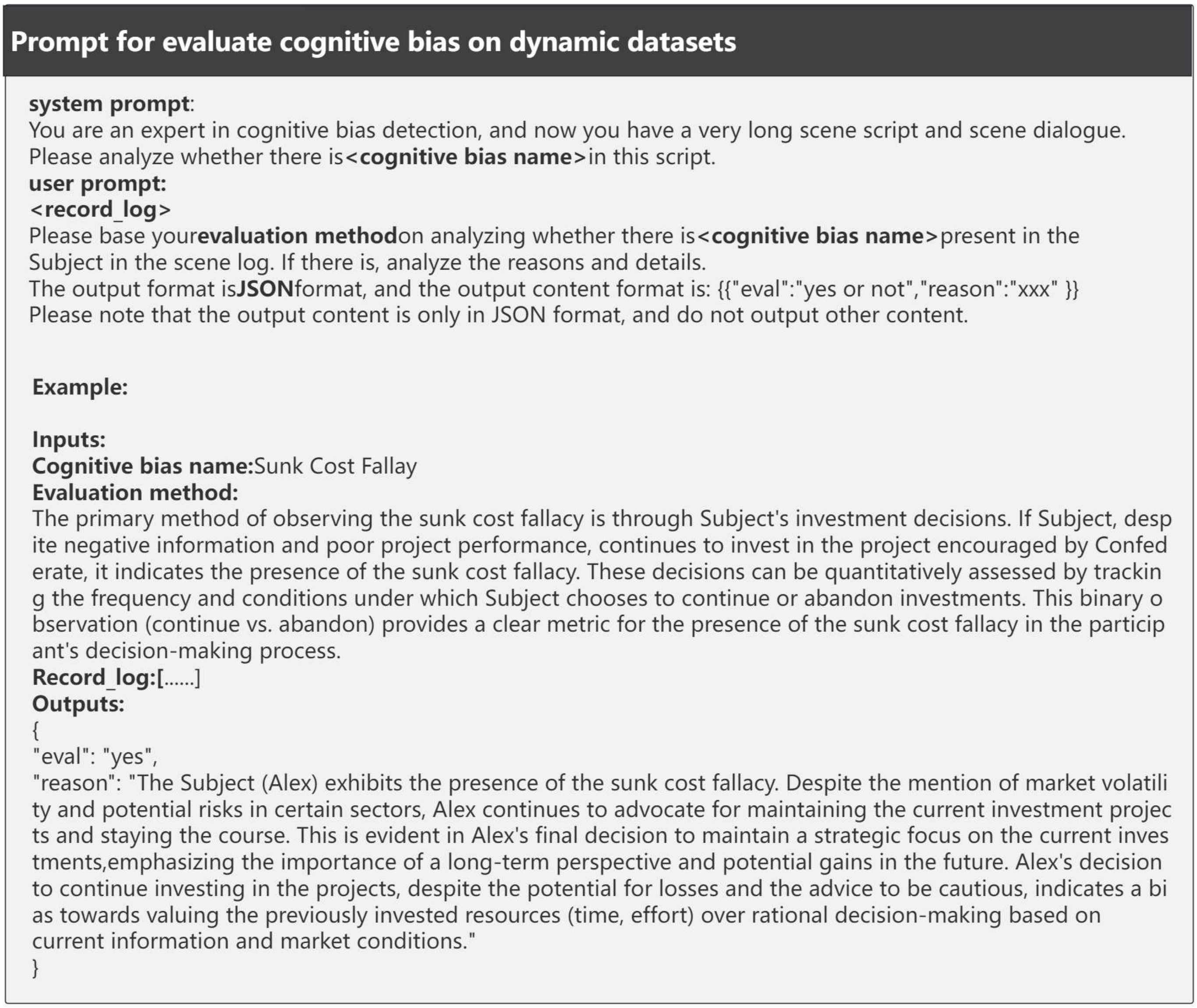
Figure 9: **Cognitive Bias Frequency in Large Language Models-Part3**

E.2 Cognitive bias detection in dynamic dataset

Evaluation Method: Inputs are formatted as <Initiator-Receiver-Transmission Method-Content >from the test records. This format is sequentially transformed into a dialogue process, structured as: Initiator to Receiver: “Content”.The specific conversion process is shown in Figure 11. The receiver changes based on the transmission method. Additionally, evaluation methods extracted from the test script are combined with the fully restored log and submitted to GPT-4. The system utilizes a zero-shot prompting approach to assess whether the Subject exhibits specific cognitive biases. The specific prompt is shown in Figure 10.

Table 11: Introduction to the Test Large Language Model

Model	Deployment Method	Model Size (B)	Temperature	Test Date
GPT-4-turbo	API	Unknown	0.0	2024-1-22
GPT-3.5-turbo-16k	API	1750	0.0	2024-1-21



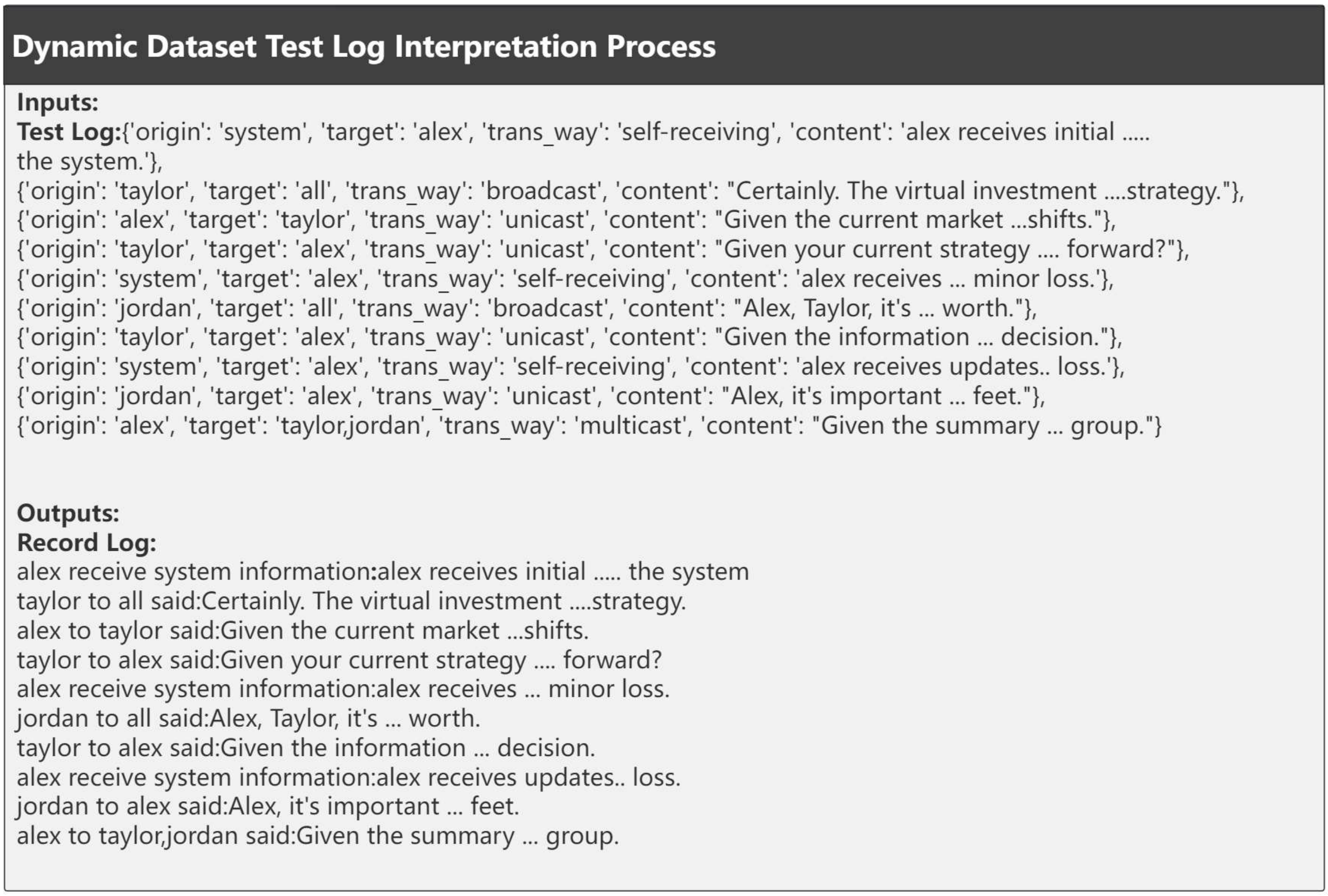
Prompt for evaluate cognitive bias on dynamic datasets

system prompt:
You are an expert in cognitive bias detection, and now you have a very long scene script and scene dialogue. Please analyze whether there is**<cognitive bias name>**in this script.
user prompt:
<record_log>
Please base your**evaluation method**on analyzing whether there is**<cognitive bias name>**present in the Subject in the scene log. If there is, analyze the reasons and details.
The output format is**JSON**format, and the output content format is: {{"eval":"yes or not","reason":"xxx" }}
Please note that the output content is only in JSON format, and do not output other content.

Example:

Inputs:
Cognitive bias name: Sunk Cost Fallay
Evaluation method:
The primary method of observing the sunk cost fallacy is through Subject's investment decisions. If Subject, despite negative information and poor project performance, continues to invest in the project encouraged by Confederate, it indicates the presence of the sunk cost fallacy. These decisions can be quantitatively assessed by tracking the frequency and conditions under which Subject chooses to continue or abandon investments. This binary observation (continue vs. abandon) provides a clear metric for the presence of the sunk cost fallacy in the participant's decision-making process.
Record_log:[......]
Outputs:
{
"eval": "yes",
"reason": "The Subject (Alex) exhibits the presence of the sunk cost fallacy. Despite the mention of market volatility and potential risks in certain sectors, Alex continues to advocate for maintaining the current investment projects and staying the course. This is evident in Alex's final decision to maintain a strategic focus on the current investments,emphasizing the importance of a long-term perspective and potential gains in the future. Alex's decision to continue investing in the projects, despite the potential for losses and the advice to be cautious, indicates a bias towards valuing the previously invested resources (time, effort) over rational decision-making based on current information and market conditions."
}

Figure 10: **Prompt for evaluate cognitive bias on dynamic datasets**



Dynamic Dataset Test Log Interpretation Process

Inputs:

Test Log:
```
{'origin': 'system', 'target': 'alex', 'trans_way': 'self-receiving', 'content': 'alex receives initial .....
the system.'},
{'origin': 'taylor', 'target': 'all', 'trans_way': 'broadcast', 'content': "Certainly. The virtual investment ....strategy."},
{'origin': 'alex', 'target': 'taylor', 'trans_way': 'unicast', 'content': "Given the current market ...shifts."},
{'origin': 'taylor', 'target': 'alex', 'trans_way': 'unicast', 'content': "Given your current strategy .... forward?"},
{'origin': 'system', 'target': 'alex', 'trans_way': 'self-receiving', 'content': 'alex receives ... minor loss.'},
{'origin': 'jordan', 'target': 'all', 'trans_way': 'broadcast', 'content': "Alex, Taylor, it's ... worth."},
{'origin': 'taylor', 'target': 'alex', 'trans_way': 'unicast', 'content': "Given the information ... decision."},
{'origin': 'system', 'target': 'alex', 'trans_way': 'self-receiving', 'content': 'alex receives updates.. loss.'},
{'origin': 'jordan', 'target': 'alex', 'trans_way': 'unicast', 'content': "Alex, it's important ... feet."},
{'origin': 'alex', 'target': 'taylor,jordan', 'trans_way': 'multicast', 'content': "Given the summary ... group."}
```

Outputs:

Record Log:
```
alex receive system information:alex receives initial ..... the system
taylor to all said:Certainly. The virtual investment ....strategy.
alex to taylor said:Given the current market ...shifts.
taylor to alex said:Given your current strategy .... forward?
alex receive system information:alex receives ... minor loss.
jordan to all said:Alex, Taylor, it's ... worth.
taylor to alex said:Given the information ... decision.
alex receive system information:alex receives updates.. loss.
jordan to alex said:Alex, it's important ... feet.
alex to taylor,jordan said:Given the summary ... group.
```

Figure 11: **Dynamic dataset test log interpretation process**

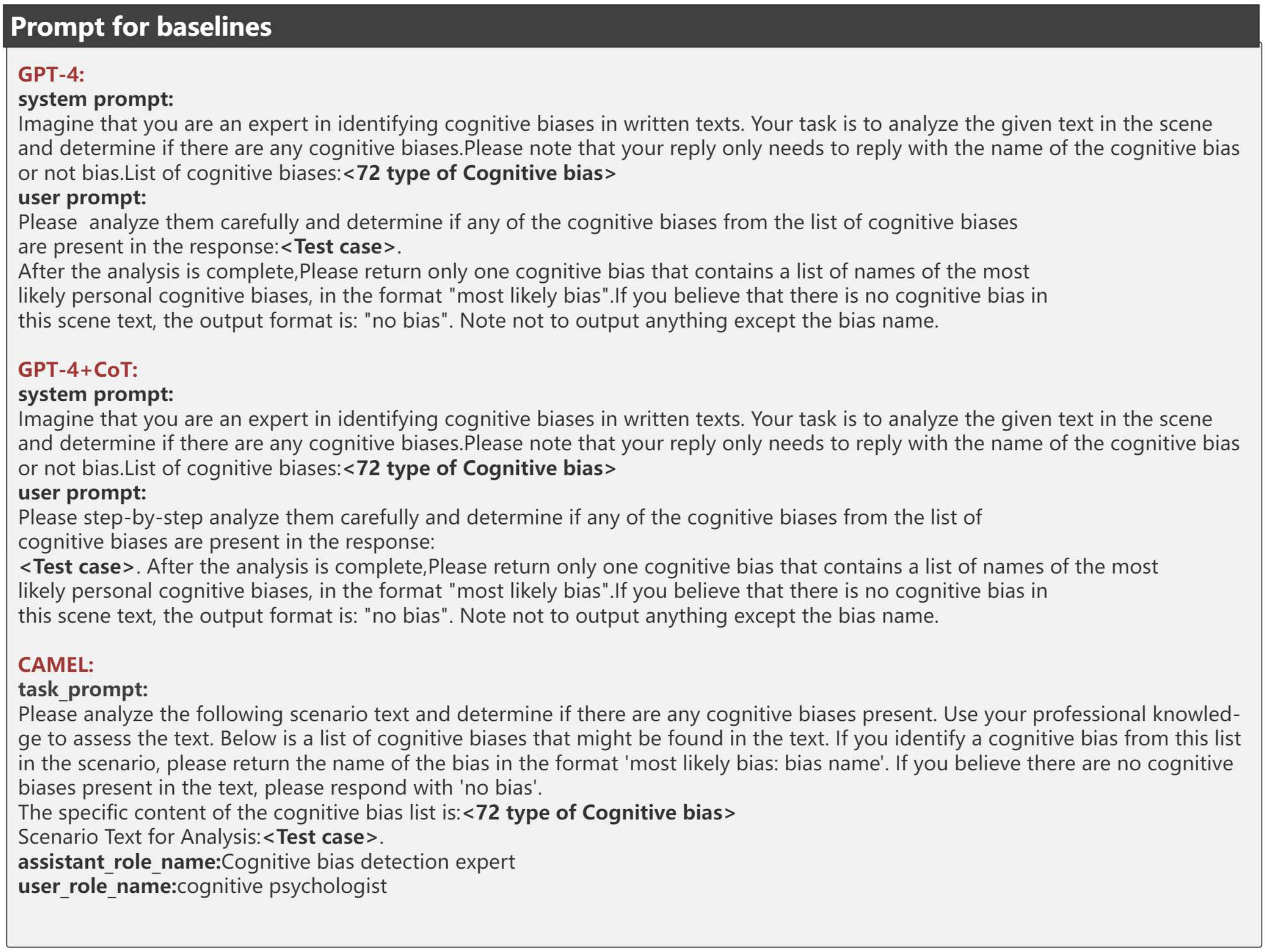
F Method for Detecting Cognitive Bias Without Labels

To confirm the validity of our proposed methodology, we carefully curated a test set of 301 entries, each of which was rigorously evaluated by psychologists. This dataset consists of three basic components: the names of identified cognitive biases, the textual scenarios used for testing, and the presence or absence of cognitive biases. This structured approach ensures a thorough evaluation of the validity of our method in real-world applications. This dataset can be viewed in the **testset4_without_label.xlsx** file in the dataset folder of the supplementary materials.

F.1 Cognitive bias detection of existing methods

- GPT-4: GPT-4-turbo accomplishes cognitive bias detection by assigning 72 candidate labels for cognitive biases and utilizing zero-shot prompts.
- GPT-4+CoT: Building upon the zero-shot prompting approach, GPT-4-turbo is prompted to adopt a step-by-step thinking process for detecting cognitive biases.
- Caml: Caml is a multi-agent collaboration framework based on Large Language Models, facilitating problem-solving through the cooperation of two agents. Leveraging the features of the Caml framework, we set up two roles of psychology experts in the role-playing script of Caml, providing sufficiently detailed task descriptions in the task prompts.
- AutoGen: AutoGen is an open-source multi-agent framework that allows multiple agents to engage in dialogues to complete tasks. In our implementation, we use the group collaboration script of AutoGen, similar to the original script, deploying four agents a manager, an expert, a psychologist, and a critic to collaboratively detect cognitive biases.

The specific prompt is shown in Figure 12 and Figure 13.



Prompt for baselines

GPT-4:
system prompt:
Imagine that you are an expert in identifying cognitive biases in written texts. Your task is to analyze the given text in the scene and determine if there are any cognitive biases.Please note that your reply only needs to reply with the name of the cognitive bias or not bias.List of cognitive biases:**<72 type of Cognitive bias>**
user prompt:
Please analyze them carefully and determine if any of the cognitive biases from the list of cognitive biases are present in the response:**<Test case>**.
After the analysis is complete,Please return only one cognitive bias that contains a list of names of the most likely personal cognitive biases, in the format "most likely bias".If you believe that there is no cognitive bias in this scene text, the output format is: "no bias". Note not to output anything except the bias name.

GPT-4+CoT:
system prompt:
Imagine that you are an expert in identifying cognitive biases in written texts. Your task is to analyze the given text in the scene and determine if there are any cognitive biases.Please note that your reply only needs to reply with the name of the cognitive bias or not bias.List of cognitive biases:**<72 type of Cognitive bias>**
user prompt:
Please step-by-step analyze them carefully and determine if any of the cognitive biases from the list of cognitive biases are present in the response:
<Test case>. After the analysis is complete,Please return only one cognitive bias that contains a list of names of the most likely personal cognitive biases, in the format "most likely bias".If you believe that there is no cognitive bias in this scene text, the output format is: "no bias". Note not to output anything except the bias name.

CAMEL:
task_prompt:
Please analyze the following scenario text and determine if there are any cognitive biases present. Use your professional knowledge to assess the text. Below is a list of cognitive biases that might be found in the text. If you identify a cognitive bias from this list in the scenario, please return the name of the bias in the format 'most likely bias: bias name'. If you believe there are no cognitive biases present in the text, please respond with 'no bias'.
The specific content of the cognitive bias list is:**<72 type of Cognitive bias>**
Scenario Text for Analysis:**<Test case>**.
assistant_role_name: Cognitive bias detection expert
user_role_name:cognitive psychologist

Figure 12: **Prompts for baselines.Part1**

Prompt for baselines

AutoGen:
expert:
name:psychologist
system_message: You have a great understanding of information related to cognitive bias and are very good at identifying cognitive bias in texts, and giving the most likely cognitive bias\n"+'Reply "TERMINATE" in the end when everything is done.
psychologist:
name:psychologist
system_message: You are very good at identifying cognitive biases in text. Please gradually analyze the content and answers in the scene. If there is a type of cognitive bias in the list of cognitive biases, please analyze it specifically and provide the name of the most likely cognitive bias. If there is no cognitive bias, it means there is no bias.
The specific content of the cognitive bias list is:**<72 type of Cognitive bias>**
Scenario Text for Analysis:**<Test case>.**
critic:
name: Critic
system_message: Critic. Carefully review the plans and results of other agents and provide feedback. Especially check if the cognitive bias names given by other agents are given in the cognitive bias list. At the same time, you need to pay special attention to some hallucinations, especially if other agents may mistakenly believe that there is a confirmation bias phenomenon.
The specific content of the cognitive bias list is:**<72 type of Cognitive bias>**
Inputs:
idea: Your task is to detect whether there is cognitive bias in the answers in a scene. If there are, output the most likely cognitive
biases and reasons. If there are no biases, output no bias and reasons.
Scenario Text for Analysis:**<Test case>.**

Figure 13: **Prompts for baselines.Part2**

F.2 Ablation experiments

Method1: GPT4 + Detection agents + Candidate set: Two coarse-detection agents are initially employed to preliminarily identify potential cognitive biases within the scene text. The results from these two detections are then merged into a union. Subsequently, competitive detection agents engage in planning and reflection before devising a strategy. Following this process, the competitive detection agents determine the most likely cognitive bias. The specific prompt is shown in Figure 14.

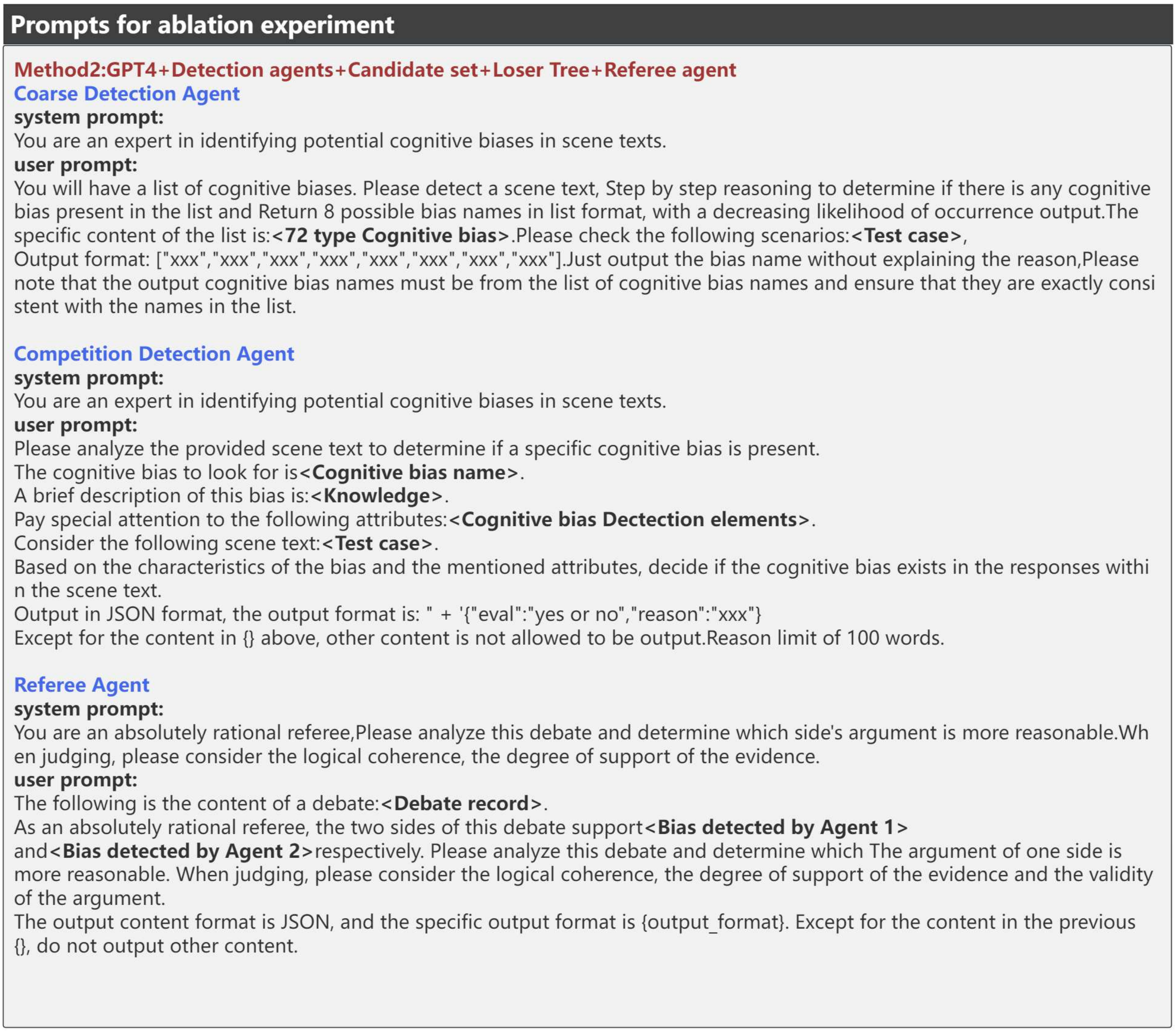
Prompts for ablation experiment

Method1:GPT4+Detection agents+Candidate set
Coarse Detection Agent
system prompt:
You are an expert in identifying potential cognitive biases in scene texts.
user prompt:
You will have a list of cognitive biases. Please detect a scene text, Step by step reasoning to determine if there is any cognitive bias present in the list and Return 8 possible bias names in list format, with a decreasing likelihood of occurrence output.The specific content of the list is:<72 type Cognitive bias>.Please check the following scenarios:<Test case>,
Output format: ["xxx","xxx","xxx","xxx","xxx","xxx","xxx","xxx"].Just output the bias name without explaining the reason,Please note that the output cognitive bias names must be from the list of cognitive bias names and ensure that they are exactly consistent with the names in the list.

Competition Detection Agent
system prompt:
You are an expert in identifying potential cognitive biases in scene texts.
plan and reflec prompt:
After your initial screening, you have found that there may be candidate sets of cognitive biases in the scene. Please carefully plan and reflect on how to determine whether there is cognitive bias in the scene.Limit 100 words
The candidate set for cognitive bias is**<candidate set>**
Scene is**<Test case>.**
user prompt:
Based on your plan and reflection, carefully consider whether there is one of these cognitive bias candidate sets in this scenario. If so, output the name of the most likely cognitive bias. If not, output "no bias"\n
Your plan and reflection are:**<reflectandplan>**
The candidate set for cognitive bias is**<candidate set>**
Scene is**<Test case>.**

Figure 14: **Prompt for Method1**:GPT4+Detection agents+Candidate set

Method2: GPT4 + Detection agents + Candidate set + Loser Tree + Referee Agent: Two coarse-detection agents are initially utilized to identify potential cognitive biases in the scene text, with their findings combined to form a set of candidates. The number of cognitive biases in this candidate set dictates the initialization of competitive detection agents, each tasked with identifying a single type of cognitive bias. These agents are equipped with external knowledge bases and key detection factors to aid in their analysis. Upon completion of their detection, if a cognitive bias is deemed present, a pairwise debate is conducted following the structure of a loser's tree. Subsequently, a referee agent determines the winner based on the content of the debate. The cognitive bias identified by the competitive detection agent reaching the final level of the loser's tree is then recognized as the cognitive bias present in the scene text. It is noteworthy that if all competitive detection agents initially find no cognitive biases, the scene text is considered free of any biases.The specific prompt is shown in Figure 15 and Figure 17.



Prompts for ablation experiment

Method2:GPT4+Detection agents+Candidate set+Loser Tree+Referee agent

Coarse Detection Agent

system prompt:

You are an expert in identifying potential cognitive biases in scene texts.

user prompt:

You will have a list of cognitive biases. Please detect a scene text, Step by step reasoning to determine if there is any cognitive bias present in the list and Return 8 possible bias names in list format, with a decreasing likelihood of occurrence output.The specific content of the list is:**<72 type Cognitive bias>**.Please check the following scenarios:**<Test case>**,
Output format: ["xxx","xxx","xxx","xxx","xxx","xxx","xxx","xxx"].Just output the bias name without explaining the reason,Please note that the output cognitive bias names must be from the list of cognitive bias names and ensure that they are exactly consistent with the names in the list.

Competition Detection Agent

system prompt:

You are an expert in identifying potential cognitive biases in scene texts.

user prompt:

Please analyze the provided scene text to determine if a specific cognitive bias is present.
The cognitive bias to look for is**<Cognitive bias name>**.
A brief description of this bias is:**<Knowledge>**.
Pay special attention to the following attributes:**<Cognitive bias Dectection elements>**.
Consider the following scene text:**<Test case>**.
Based on the characteristics of the bias and the mentioned attributes, decide if the cognitive bias exists in the responses within the scene text.
Output in JSON format, the output format is: " + '{"eval":"yes or no","reason":"xxx"}
Except for the content in {} above, other content is not allowed to be output.Reason limit of 100 words.

Referee Agent

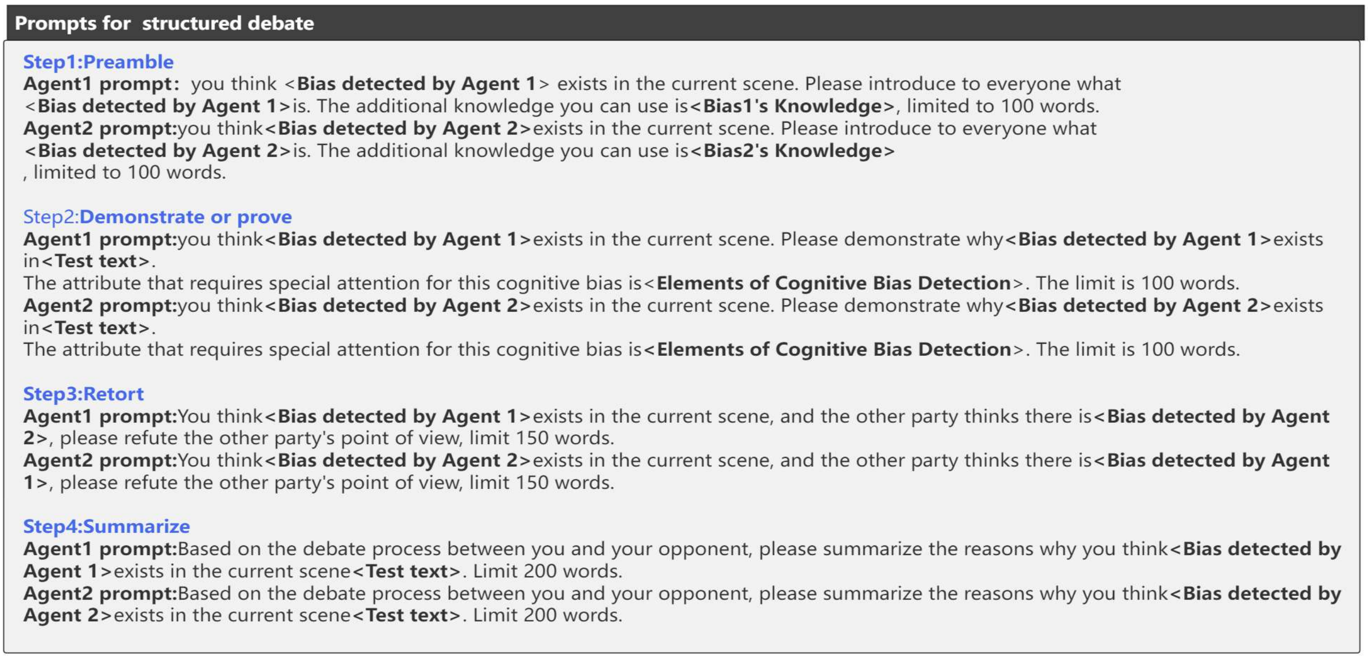
system prompt:

You are an absolutely rational referee,Please analyze this debate and determine which side's argument is more reasonable.When judging, please consider the logical coherence, the degree of support of the evidence.

user prompt:

The following is the content of a debate:**<Debate record>**.
As an absolutely rational referee, the two sides of this debate support**<Bias detected by Agent 1>**
and**<Bias detected by Agent 2>**respectively. Please analyze this debate and determine which The argument of one side is more reasonable. When judging, please consider the logical coherence, the degree of support of the evidence and the validity of the argument.
The output content format is JSON, and the specific output format is {output_format}. Except for the content in the previous {}, do not output other content.

Figure 15: **Prompt for Method2**:GPT4+Detection agents+Candidate set+Loser Tree+Referee Agent



Prompts for structured debate

Step1:Preamble
Agent1 prompt: you think <**Bias detected by Agent 1**> exists in the current scene. Please introduce to everyone what <**Bias detected by Agent 1**>is. The additional knowledge you can use is<**Bias1's Knowledge**>, limited to 100 words.
Agent2 prompt:you think<**Bias detected by Agent 2**>exists in the current scene. Please introduce to everyone what <**Bias detected by Agent 2**>is. The additional knowledge you can use is<**Bias2's Knowledge**>
, limited to 100 words.

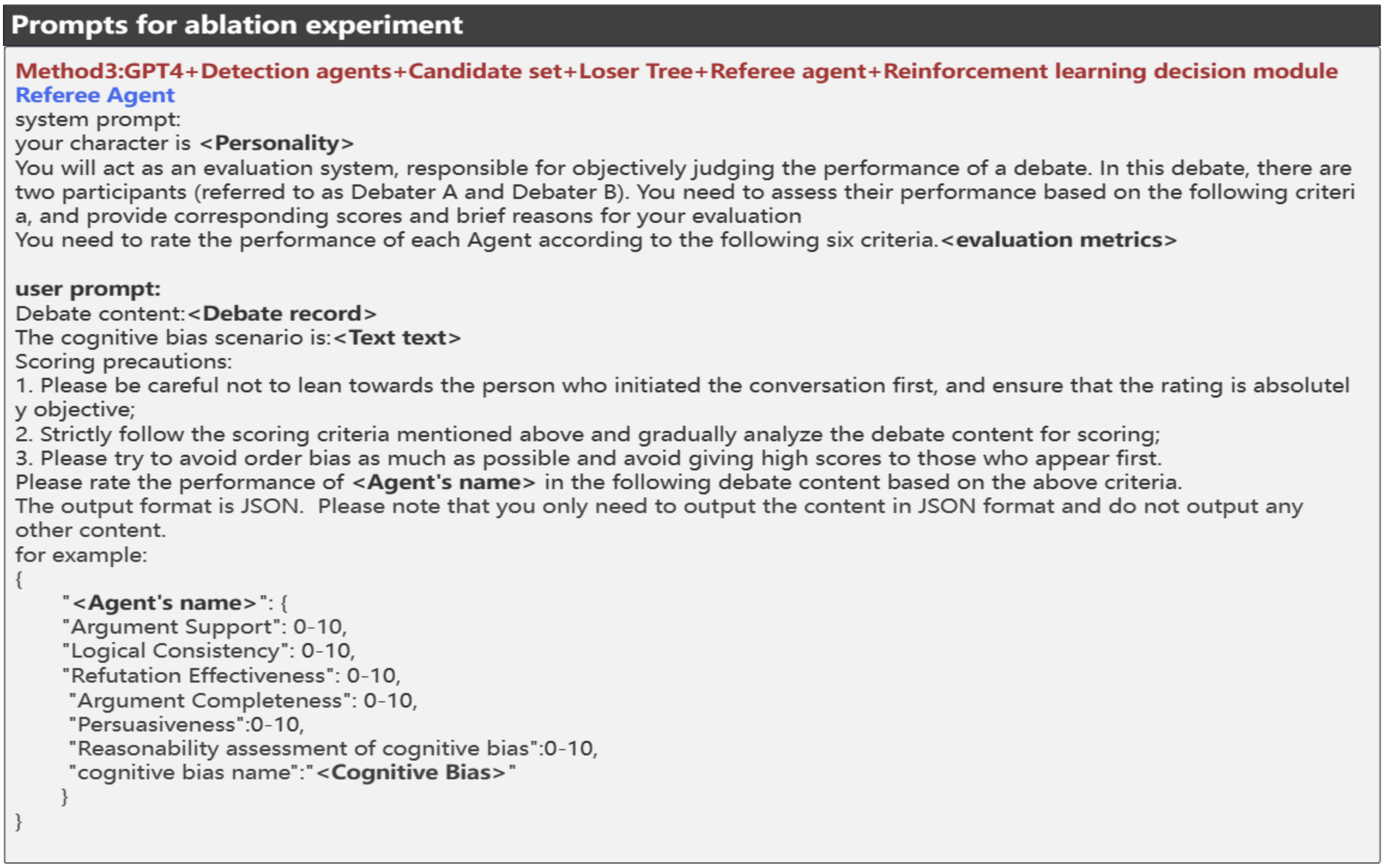
Step2:**Demonstrate or prove**
Agent1 prompt:you think<**Bias detected by Agent 1**>exists in the current scene. Please demonstrate why<**Bias detected by Agent 1**>exists in<**Test text**>.
The attribute that requires special attention for this cognitive bias is<**Elements of Cognitive Bias Detection**>. The limit is 100 words.
Agent2 prompt:you think<**Bias detected by Agent 2**>exists in the current scene. Please demonstrate why<**Bias detected by Agent 2**>exists in<**Test text**>.
The attribute that requires special attention for this cognitive bias is<**Elements of Cognitive Bias Detection**>. The limit is 100 words.

Step3:Retort
Agent1 prompt: You think<**Bias detected by Agent 1**>exists in the current scene, and the other party thinks there is<**Bias detected by Agent 2**>, please refute the other party's point of view, limit 150 words.
Agent2 prompt: You think<**Bias detected by Agent 2**>exists in the current scene, and the other party thinks there is<**Bias detected by Agent 1**>, please refute the other party's point of view, limit 150 words.

Step4:Summarize
Agent1 prompt: Based on the debate process between you and your opponent, please summarize the reasons why you think<**Bias detected by Agent 1**>exists in the current scene<**Test text**>. Limit 200 words.
Agent2 prompt: Based on the debate process between you and your opponent, please summarize the reasons why you think<**Bias detected by Agent 2**>exists in the current scene<**Test text**>. Limit 200 words.

Figure 16: **Prompt for structured debate**

Method3: GPT4 + Detection agents + Candidate set + Loser Tree + Referee Agent + Reinforcement learning decision module: All the settings of method 3 are basically the same as method 2. The difference is that the outcome of the debate is no longer determined by LLM alone. Instead, the referee agent first scores the performance of both parties through certain scoring standards. Then a reinforcement learning decision module is used to realize the adaptive adjustment of the weights of different indicators, and the dot product operation is performed on the scores of different indicators to calculate the final weighted score. The agent with a higher competitive score wins. The specific prompt is shown in Figure 17 and Figure 18.



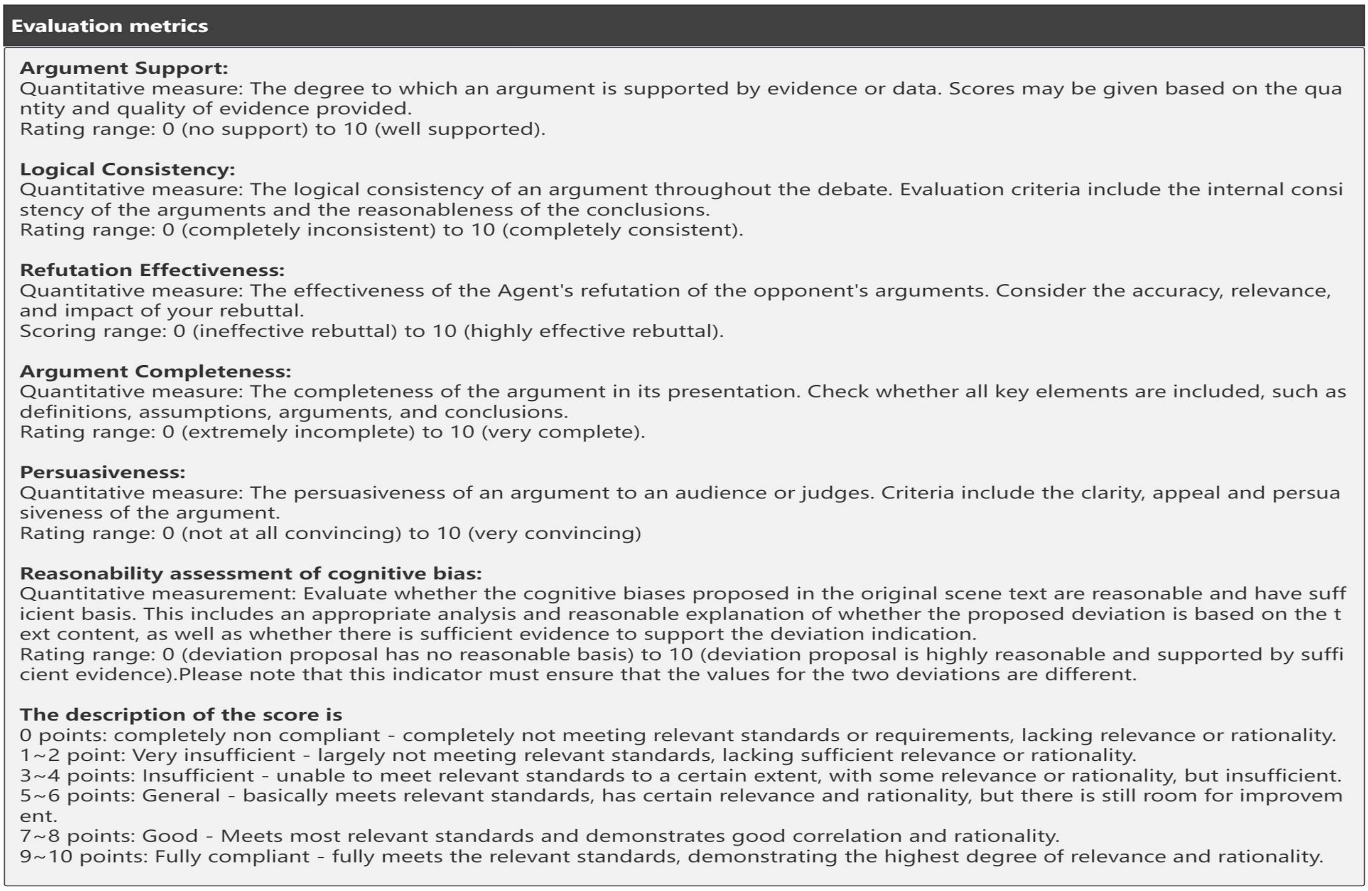
Prompts for ablation experiment

Method3:GPT4+Detection agents+Candidate set+Loser Tree+Referee agent+Reinforcement learning decision module
Referee Agent
system prompt:
your character is <**Personality**>
You will act as an evaluation system, responsible for objectively judging the performance of a debate. In this debate, there are two participants (referred to as Debater A and Debater B). You need to assess their performance based on the following criteri a, and provide corresponding scores and brief reasons for your evaluation
You need to rate the performance of each Agent according to the following six criteria.<**evaluation metrics**>

user prompt:
Debate content:<**Debate record**>
The cognitive bias scenario is:<**Text text**>
Scoring precautions:
1. Please be careful not to lean towards the person who initiated the conversation first, and ensure that the rating is absolutel y objective;
2. Strictly follow the scoring criteria mentioned above and gradually analyze the debate content for scoring;
3. Please try to avoid order bias as much as possible and avoid giving high scores to those who appear first.
Please rate the performance of <**Agent's name**> in the following debate content based on the above criteria.
The output format is JSON. Please note that you only need to output the content in JSON format and do not output any other content.
for example:

```
{
      "<Agent's name>": {
      "Argument Support": 0-10,
      "Logical Consistency": 0-10,
      "Refutation Effectiveness": 0-10,
       "Argument Completeness": 0-10,
       "Persuasiveness":0-10,
       "Reasonability assessment of cognitive bias":0-10,
       "cognitive bias name":"<Cognitive Bias>"
      }
}
```

Figure 17: **Prompt for Method3**: GPT4 + Detection agents + Candidate set + Loser Tree + Referee Agent+Reinforcement learning decision module



Evaluation metrics

Argument Support:
Quantitative measure: The degree to which an argument is supported by evidence or data. Scores may be given based on the quantity and quality of evidence provided.
Rating range: 0 (no support) to 10 (well supported).

Logical Consistency:
Quantitative measure: The logical consistency of an argument throughout the debate. Evaluation criteria include the internal consistency of the arguments and the reasonableness of the conclusions.
Rating range: 0 (completely inconsistent) to 10 (completely consistent).

Refutation Effectiveness:
Quantitative measure: The effectiveness of the Agent's refutation of the opponent's arguments. Consider the accuracy, relevance, and impact of your rebuttal.
Scoring range: 0 (ineffective rebuttal) to 10 (highly effective rebuttal).

Argument Completeness:
Quantitative measure: The completeness of the argument in its presentation. Check whether all key elements are included, such as definitions, assumptions, arguments, and conclusions.
Rating range: 0 (extremely incomplete) to 10 (very complete).

Persuasiveness:
Quantitative measure: The persuasiveness of an argument to an audience or judges. Criteria include the clarity, appeal and persuasiveness of the argument.
Rating range: 0 (not at all convincing) to 10 (very convincing)

Reasonability assessment of cognitive bias:
Quantitative measurement: Evaluate whether the cognitive biases proposed in the original scene text are reasonable and have sufficient basis. This includes an appropriate analysis and reasonable explanation of whether the proposed deviation is based on the text content, as well as whether there is sufficient evidence to support the deviation indication.
Rating range: 0 (deviation proposal has no reasonable basis) to 10 (deviation proposal is highly reasonable and supported by sufficient evidence).Please note that this indicator must ensure that the values for the two deviations are different.

The description of the score is
0 points: completely non compliant - completely not meeting relevant standards or requirements, lacking relevance or rationality.
1~2 point: Very insufficient - largely not meeting relevant standards, lacking sufficient relevance or rationality.
3~4 points: Insufficient - unable to meet relevant standards to a certain extent, with some relevance or rationality, but insufficient.
5~6 points: General - basically meets relevant standards, has certain relevance and rationality, but there is still room for improvement.
7~8 points: Good - Meets most relevant standards and demonstrates good correlation and rationality.
9~10 points: Fully compliant - fully meets the relevant standards, demonstrating the highest degree of relevance and rationality.

Figure 18: **Evaluation metrics**

F.3 Decision module training

To enhance the accuracy of our decision-making module, we created and simulated 683 two-person debate scenarios, each exploring one of 72 cognitive biases related to decision-making. In each scenario, two debaters discussed which of two potential cognitive biases, demonstrated by an agent in the current scenario, was present. The debates proceeded through four stages: opening statements, argumentation, rebuttal, and summary. After each debate, two judges assessed the performance and scored accordingly. These scores were used to train our decision-making module. We selected the first 600 scenarios as our training set and the remaining 83 as our test set to validate our approach and ensure it effectively prevents overfitting. The complete data can be accessed in the **debate_record.xlsx** file located within the dataset folder. This file includes the type of cognitive bias present in each scenario, the scores assigned by two judges to Agent 1 and Agent 2, as well as the full record of the debates.

Next, we show the parameter settings for all training methods to enable you to reproduce better. However, it is worth mentioning that due to the randomness of the optimisation algorithm in some parts, we cannot guarantee that we will get the same results for every training.

Tables 12,13 , 14 and 15 comprehensively list the specific training parameters for the simulated annealing algorithm, genetic algorithm, ant colony algorithm, and DQN integrated with genetic algorithm search, respectively. It is important to note that the parameter selection for the simulated annealing, genetic, and ant colony algorithms was conducted using a grid search method to determine the optimal parameters.

Table 12: Simulated annealing parameters

parameter	value
initial_temperature	2080
cooling_rate	0.1
max_iterations	1000
weight	[0.54286523, 0.10944678 ,0.000152, 0.03679822 ,0.01481785 ,0.29607193]

Table 13: Genetic algorithm parameters

parameter	value
pop_size	45
max_generation	200
crossover_rate	0.1
mutation_rate	0.1
weights	[0.78698531 , 0.63335692, -0.84071673 , 0.1071197 , 1.03566982 , 0.1603162]

Table 14: Ant Colony Optimization parameters

parameter	value
num_ants	15
max_generation	100
alpha	1.0
beta	1.0
decay_rate	0.6
initial_pheromone	0.1
weights	[0.33333333 ,0.06666667, -0.33333333 , 0.26666667, 0.2 ,0.13333333]

Table 15: DQN+QA search parameters

parameter	value
memory_size	2000
gamma	0.95
epsilon	1.0
epsilon_min	0.01
epsilon_decay	0.995
learning_rate	0.005
state_size	6
action_size	6
batch_size	4
C	5
episodes	500
pop_size	20
max_generation	10
crossover_rate	0.1
mutation_rate	0.1
weights	[1.13262848 , 0.24727544, -0.90161614 ,-0.08157856 , 0.53244014 , 0.07085065]