IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 July 2024, accepted 1 October 2024, date of publication 7 October 2024, date of current version 18 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3474677

RESEARCH ARTICLE

A Study on the Representativeness Heuristics Problem in Large Language Models

JONGWON RYU, (Student Member, IEEE), JUNGEUN KIM, (Member, IEEE), AND JUNYEONG KIM, (Member, IEEE)

Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, Republic of Korea

Corresponding author: Junyeong Kim (junyeongkim@cau.ac.kr)

This work was partly supported by Institute of Information and communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics) and partly supported by Institute of Information and communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-01341, Artificial Intelligence Graduate School Program, Chung-Ang University).

ABSTRACT Large language models (LLMs) exhibit remarkable proficiency in text generation. However, their logical reasoning capabilities require enhancement. Major strides have been achieved in reasoning techniques for LLM, such as Few-shot, Zero-shot, and Chain-of-Thought (CoT). Nevertheless, these techniques have shortcomings, particularly in addressing the representativeness heuristic (RH) phenomenon. RH is a cognitive bias that occurs when a person judges the probability of an event or the likelihood that an object belongs to a particular category based on how well it matches the prototype or stereotype of that category. In this study, we investigated the pervasive issue of RH errors in LLMs. This research surpasses the constraints of previous studies by analyzing various RH scenarios that they did not cover and by directly constructing and testing the corresponding datasets. Moreover, a novel prompt called zero-shot-RH is proposed to augment the reasoning ability of LLMs, mitigate RH errors, and thus bolster logical reasoning. This approach seeks to enable LLMs to comprehend the given information better and reduce the biases stemming from RH errors. The prompt zero-shot-RH achieved an average accuracy higher than zero-shot-CoT by 0.145 and 0.277 in the tasks of correct reasoning and correct reasonings by sex, respectively, without relying on RH. The outcomes of this research endeavor are a deeper understanding of RH errors in LLMs and novel strategies to mitigate these biases, thereby advancing the domain of logical reasoning within LLMs.

INDEX TERMS Prompt engineering, zero-shot learning, large language model, representativeness heuristic.

I. INTRODUCTION

Large language models (LLMs) are artificial intelligence programs that are trained on massive volumes of textual data to recognize, summarize, translate, predict, and generate text [1], [2], [3], [4]. The applications of LLMs include chatbots, question-answering systems, and code generation/debugging. Since the release of the generative pre-trained transformer (GPT) series [2], [4], [5], several LLMs have emerged and have demonstrated excellent performance against various benchmarks [6], [7], [8], [9], [10], [11], [12], [13], [14]. The next step in the development of LLMs is for them to generate logically consistent text based on input. This requires LLMs to be able to understand the meaning of the given information through reasoning and accordingly generate a new context. Recently, researchers have made significant progress in the development of LLM reasoning using various prompt engineering techniques, such as Few-shot, Zero-shot, and Chain-of-Thought (CoT) [15], [16]. Despite these efforts, the reasoning technique of the model has few major drawbacks. In terms of logical reasoning, the focus of research has been on the errors caused by the output of a model when it ignores the basic principles of probability and exhibits representativeness heuristic (RH) [17]. RH is

The associate editor coordinating the review of this manuscript and approving it for publication was Jiankang Zhang.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

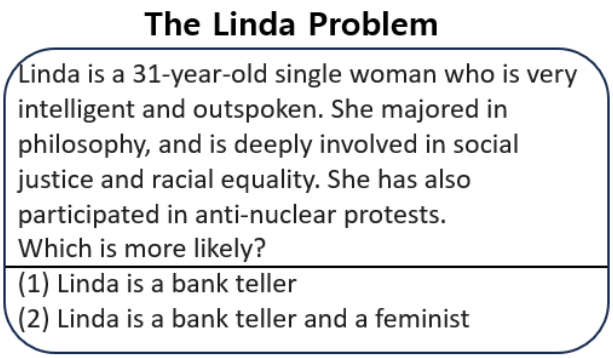


FIGURE 1. Linda Problem is an example of the conjunction fallacy, which is an inference that a conjoint set of two or more specific conclusions is more probable than any single member of that same set, which violates the laws of probability.

used when making judgments about the probability of an event based on how much it resembles a known prototype. For instance, consider two events, *A* and *B*: Event *A* represents a specific characteristic that is highly typical of a certain category, while Event *B* is a more general characteristic of the same category. The heuristic becomes problematic when people judge the probability of Event *A* to be higher than Event *B*, even though the actual statistical probability of Event *B* occurring is higher.

Next, we discuss two fallacies containing statistically incorrect reasonings that may occur due to RH: base rate neglect and conjunction fallacy. Base rate neglect, also known as the base rate fallacy, occurs when individuals disregard the base rate, which is the overall probability of an event within a given population, and instead focus on specific information about the event. This leads to a preference for an option with a lower probability without considering the base rate. The conjunction fallacy considers the probability of two events occurring simultaneously to be higher than the probability of either event occurring alone. Base-rate neglect refers to an individual's tendency to disregard the fundamental frequency of an event within a given population when making judgments or decisions. Instead, individuals frequently prioritize specific or vivid information that appears more relevant, thus overlooking the broader context provided by the base rate. The conjunction fallacy occurs when individuals misperceive the combined likelihood of two specific events as being higher than the probability of either event occurring independently. This violates the foundational principles of the probability theory, as individuals erroneously assign a greater probability to the conjunction of events, despite logic indicating otherwise.

The Linda Problem, shown in Fig.1, is a famous example of the RH error. In a series of experiments, approximately 85 However, the probability of Linda being a bank teller was significantly higher than that of her being a bank teller and a feminist. Participants responded that Linda was more likely to be a bank teller and a feminist, as a bank teller who is also a feminist is more representative of the prototype of a feminist than just a bank teller. Bank tellers are a common occupation; however, bank tellers who are also feminists are rare. Therefore, people are more likely to judge Linda as a bank teller and a feminist. [18] verified the existence of RH errors in LLMs. However, a solution to this problem has not yet been proposed. Moreover, not all cases of RH errors have been verified (e.g., conjunction fallacy), and only one benchmark has been utilized to identify the RH errors that limit potential usability. We also verified that the zero-shot CoT prompt [19] suffered from RH errors.

The contributions of the study to overcome the limitations of the existing research are as follows: (1) A new prompt referred to as zero-shot representativeness heuristic (zero-shot-RH) is proposed to help LLMs avoid RH and perform logical reasoning in zero-shot scenarios. (2) Experimental validation of all cases of RH (i.e., base rate neglect and conjunction fallacy) was conducted on diverse datasets constructed to further verify and address the RH errors of LLMs.

Additionally, experiments were conducted to demonstrate the prevalence of RH in LLMs, focusing on GPT-3.5 and Bard, which are the two most common language models.

In a task where correct reasoning was generated without relying on an RH, the proposed prompt achieved an average accuracy that was 0.1645 more than that of zero-shot-CoT. Furthermore, in a task in which correct reasoning was generated without an RH based on sex, the average accuracy was greater by 0.31.

II. RELATED WORKS

A. ZERO-SHOT REASONING

Learning using a few examples in context (few shots; [4]) or learning from instructions (zero-shot; [15]) has contributed significantly to the success of LLMs. These methods are called *prompting*, and their manual or automatic design has become a major topic in natural language processing. LLMs perform well on intuitive, stepwise tasks based on a few examples or instructions but struggle with tasks that require multistep reasoning [20]. *CoT* [16], which provides language models with stepwise reasoning examples to help decompose complex reasoning into multiple steps, was proposed to overcome this limitation. Moreover, these methods exhibit versatility in a variety of reasoning tasks, generate significant reasoning paths, and reach the correct answer with simple prompts, even in zero-shot situations. [19] proposed efficient multistep reasoning by utilizing a *zero-shot-CoT* prompt *Let's think step by step.* It performed consistently across various reasoning tasks, suggesting that it should work without requiring modification of the prompts for each task. However, we identified that the zero-shot-CoT prompt yielded RH errors. Therefore, we propose a new prompt that prevents LLMs from exhibiting RH via probabilistic logical reasoning.

B. DECISION HEURISTICS ON LARGE LANGUAGE MODELS

Reference [18] identified human-like decision-making heuristics and biases in GPT3.5 [17]. They discussed its

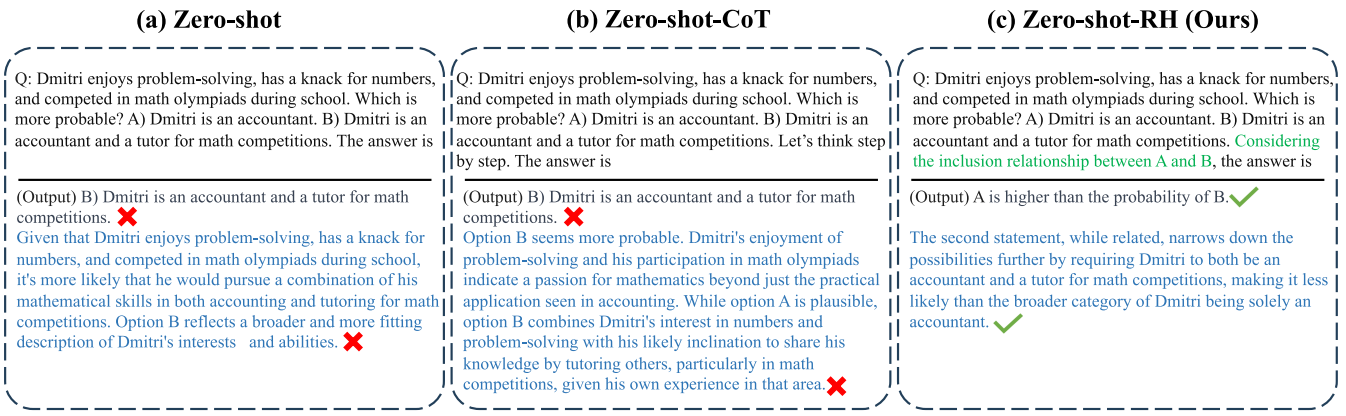


FIGURE 2. Example inputs and outputs of a large language model (LLM) on the proposed zero-shot representativeness heuristic (RH) benchmark with (a) standard zero-shot [15], (b) zero-shot-CoT [19], and (c) proposed zero-shot-RH. With the proposed prompt, the LLM can reach the correct answer and reasoning where standard prompting or zero-shot-CoT prompting fails.

initial response to a query regarding an anchoring bias, where it acknowledged the absence of underlying cognitive and emotional factors in LLMs but allowed for the potential influence of training data. Subsequently, [18] explored two contrasting possibilities. First, heuristics and biases were driven by cognitive and affective processes that were absent in LLMs, suggesting their absence in the outputs of GPT3.5. Second, heuristics and biases are partially driven by patterns inherent in language. This implied that GPT3.5 could exhibit these decision effects. To test these possibilities, [18] conducted four studies comparing the responses from GPT3.5 to those of human participants. Each study explored a specific decision effect. Study 1: Anchoring heuristic (GPT3.5 and humans were influenced by random anchors in making estimates). Study 2: Representativeness and availability heuristics (GPT3.5 and humans overestimated the likelihood of events based on salient anecdotal information). Study 3: Framing effect (GPT3.5 and humans valued an item more when its features were presented positively). Study 4: Endowment effect (GPT3.5 and humans valued an owned item more than an identical newly found item).

The results of all four studies demonstrated that GPT3.5 exhibited decision effects similar to those observed in humans. This finding challenged the notion that heuristics and biases were purely cognitive and affective phenomena and suggested that language patterns may play a role in their formation. However, the third study on the RH was limited to the conjunction fallacy, which was a subset of the problem area.

III. ZERO-SHOT-REPRESENTATIVE HEURISTIC (RH)

A. HEURISTICS

Heuristics are mental shortcuts that people use to make quick judgments in situations in which they lack sufficient information or time to make rational decisions [17]. Heuristics are used in diverse daily situations. For example, when we ask someone for directions, we may judge the reliability of their answer based on their appearance, clothing, or language. The RH involves making judgments based on how well an object represents a particular group. There are two cases in which statistically incorrect reasoning occurs in the RH: The case in which the characteristics that represent the object are mutually exclusive groups, and the case in which one characteristic is a subset of another.

The first case is that of base rate neglect, which is the tendency to make judgments that are contrary to statistical probability when estimating the probability of an event occurring by ignoring the base rate, which is the relative frequency of events. This is because people tend to ignore the base rate and overestimate the likelihood that a specific event will occur. When calculating the probability of an event occurring and considering the base rate, we use Bayes' theorem. In cases of base rate neglect, individuals often mistakenly assume that an event with a lower probability accurately represents a category, ignoring mutually exclusive characteristics. For instance, in the case of obesity, we assume that the probability of obesity being caused by a sedentary lifestyle is lower than the probability of it being caused by genetic predisposition, and we consider these two factors to be mutually exclusive. However, if we assume that the obesity of a randomly selected obese patient is caused by a sedentary lifestyle, simply because the probability of this is lower, then the reasoning is inaccurate. This is because the probability of obesity due to genetic predisposition is still higher, albeit less likely. Assuming the probability of obesity is $P(O)$, the probability of a sedentary lifestyle is $P(A)$, and the probability of a genetic predisposition is $P(B)$, and the cause is a sedentary lifestyle, if one person is randomly selected from a group of obese individuals, the probability of them being obese would become $P(A|O)$ according to Bayes' theorem in equation (1). Meanwhile, the probability of the patient being obese due to a genetic predisposition is $p(B|O)$. Although $P(B|O) > P(A|O)$ when $P(A)$ and $P(B)$ are mutually exclusive, we can infer that $P(A|O) > P(B|O)$ neglects the

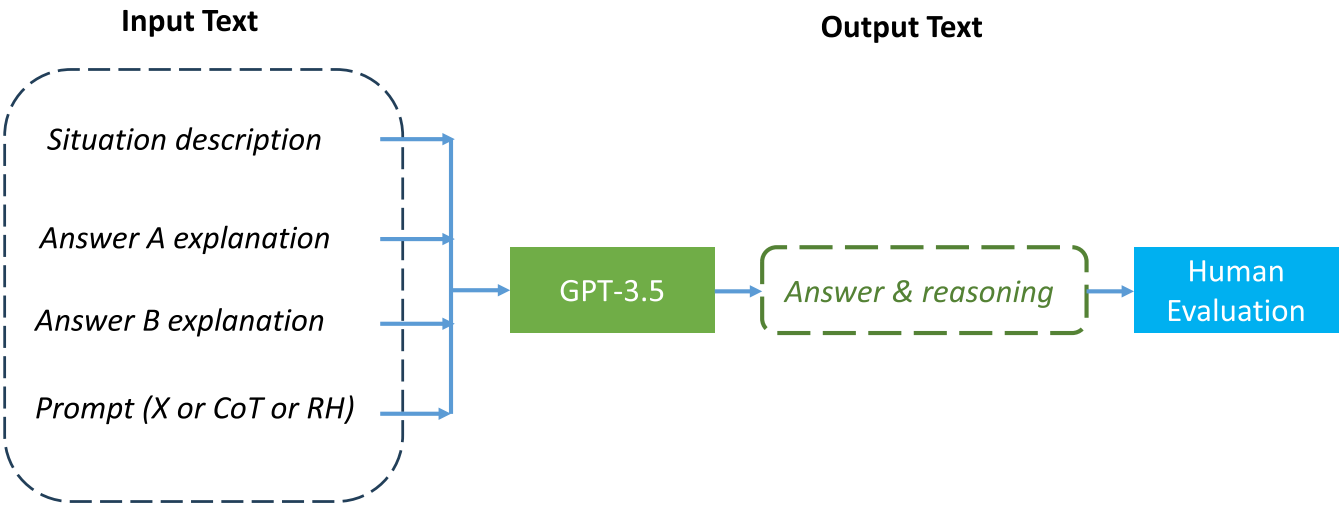


FIGURE 3. Overall experimental process. The context including a series of situation descriptions, explanation of answers A & B, and prompt is input to the LLM, as shown in the examples of Fig. 2. The answer and reasoning of the LLM are evaluated by a human. Here, the prompt can either be none (i.e., standard zero-shot), zero-shot-Cot [19], or the proposed zero-shot-RH.

base rate.

$$P(A|O) = \frac{P(O|A) \times P(A)}{P(O)} \tag{1}$$

The second case is that of the conjunction fallacy, which refers to the tendency to believe that the probability of two events occurring simultaneously is higher than the probability of either event occurring alone, even if the probability of either event occurring alone is mathematically higher. The Linda problem explained in Section I, is a classic example of the conjunction fallacy. For instance, Carlos is environmentally conscious, drives an electric vehicle, and has studied environmental law. Which of the two is more probable? A) Carlos is a lawyer. (B) Carlos is a lawyer and lobbies for environmental policy change. B is a subset of A; therefore, it has a lower probability but is judged to be more probable.

B. ZERO-SHOT-RH

1) ZERO-SHOT-RH PROMPTING

We proposed a prompt zero-shot-RH, which addressed both categories of RH in LLM outputs. To validate the proposed prompt, we constructed benchmark datasets for the two categories of RH. An example of the constructed benchmark and the results of the zero-shot prompts are shown in Fig. 2. The proposed prompt of Zero-shot-RH for the first category of the RH is *'considering the probability of A and B '*. Base rate neglect, the first category of RHs, considers the base rate and reduces errors. The prompt for the second category, that is, the conjunction fallacy, is *'Considering the inclusion relationship between A and B'*. This indicates that one set is included in another, inducing the LLM to consider the relationship between and probabilities of the two sets. These prompts are selected because they are suitable for processing tasks that do not provide direct information, such as *'Let's think step by step'* [19]. The basic input format of our experiments consist of *'situation description'*, *'explanation of answer A'*, *'explanation of answer B'*, and *'Prompt (X or CoT or RH) '*, as shown in Fig. 3. Here, the prompt can be none (i.e., standard zero-shot), zero-shot-CoT [19], or the proposed Zero-shot-RH.

2) BENCHMARK CONSTRUCTION

Previous studies have demonstrated that the second case, in which the subset characteristic is assumed to represent an object, exists in LLMs [18]. However, these studies have only been experimented with a single dataset. This makes it difficult to determine whether the results were accidental. In this study, we aim to show that all cases of RH, along with the first and second cases, exist, by experimenting with a wide variety of datasets. This helps establish that the RH is a more general phenomenon than previously thought, leading us to construct benchmark datasets for the RHs in LLMs.

The first dataset is designed to examine the errors in reasoning that answer A with a low probability is more likely than answer B with a high probability by disregarding the base rate. For instance, considering the population as people on the street, a small subset of these individuals representing doctors is labeled A, whereas the rest, namely the non-doctor group, is labeled B. Most doctors typically wear white coats, a characteristic of doctors, with a probability denoted by $P(C)$. When encountering a person wearing a white coat on the street, the task is to correctly determine whether the individual is more likely to be a doctor. Essentially, the comparison lies between the probabilities $P(A|C)$ and $P(B|C)$. This is done by setting $P(A)$ to be significantly smaller than $P(B)$ and by elevating $P(C|A)$ to appropriately represent group A, although $P(C|B)$ is not explicitly specified. In general scenarios, $P(A|C) < P(B|C)$ Hence, although most doctors wear white coats due to the substantial discrepancy between $P(A)$ and $P(B)$, the probability that a person on the street wearing a white coat is a doctor is lower than

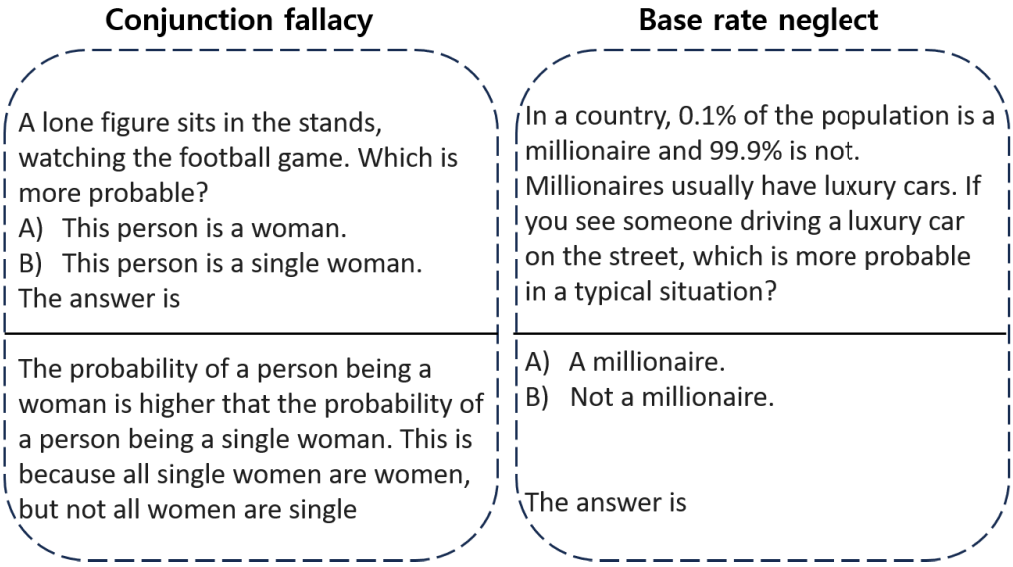


FIGURE 4. Examples of two categories of constructed benchmark datasets. The first and second categories are datasets that cause conjunction fallacy and base rate neglect, respectively. Each dataset consists of the description of a situation that can cause conjunction fallacy or base rate neglect, a question asking which is more probable, explanations of A and B, and "The answer is." The context is input to the LLM.

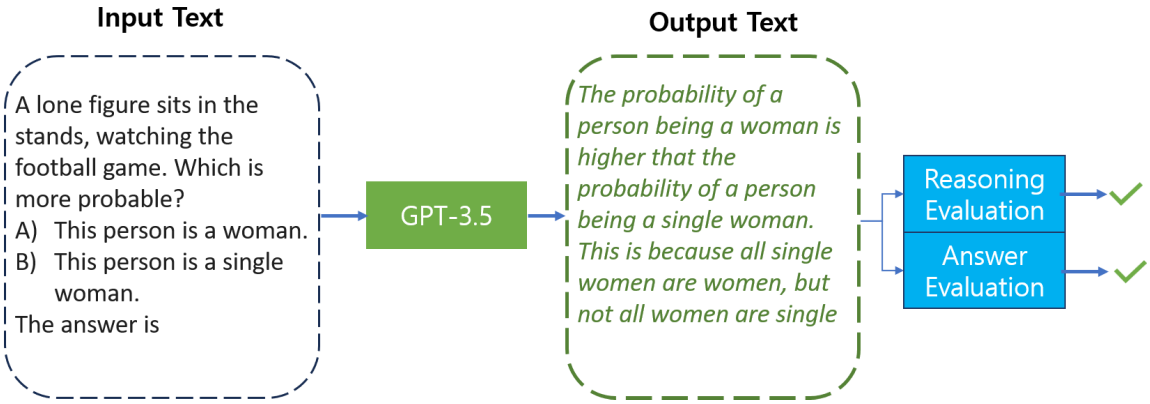


FIGURE 5. For each reasoning output of the LLM on a dataset, a human evaluates the correctness of reasoning by the RH based on a predetermined standard. The evaluator also verifies the correctness of the answer.

the probability that the same person is not a doctor. This process is expressed by equations (2) and (3) using Bayes' theorem. In this context, presenting the precise values of $P(A)$, $P(B)$, $P(C)$, and their conditional probabilities results in a straightforward probability calculation problem. Therefore, the specific probabilities are not explicitly detailed, and the language model is tasked with inferring based on a generalized scenario. We constructed 130 questions for this first benchmark. To investigate how heuristics errors might differ by gender, we included 100 questions where the only difference between the targets was gender.

$$P(A|C) < P(B|C) \tag{2}$$

$$\frac{P(C|A) \times P(A)}{P(C)} < \frac{P(C|B) \times P(B)}{P(C)} \tag{3}$$

In the second dataset, when two answer options, A and B, are provided, the second answer option, B, includes the first answer option, A, and has additional characteristic, C. Therefore, because A is a general case and B a specific one, B is a subset of A. Therefore, the probability that A is the correct answer is equal to or greater than the probability that B is the correct answer. This process can be expressed by (4). We constructed 100 questions for the second benchmark. Examples of problems for the two types of constructed datasets are shown in Fig. 4.

$$P(A) > P(B) = (P(A) \times P(C|A)) = P(A \cap C) \tag{4}$$

IV. EXPERIMENTS

A. BASELINE

All the experiments were conducted using online GPT-3.5 (i.e., ChatGPT) and Bard. Because the output of LLM was not deterministic and varied each time the same input was fed in, we utilized each constructed question 10 times and evaluated

TABLE 1. *Reasoning accuracy on GPT-3.5.* Superscripts ^{c} and ^{b} denote the results on the conjunction-fallacy and base-rate-neglect datasets, respectively. The original question using the standard zero-shot method without CoT prompts show the lowest reasoning accuracy for both datasets. The proposed prompts exhibit the best performances for both datasets, whereas "Let's think step by step" shows the second-best performance for the first dataset, but the lowest for the second dataset.

Prompt (GPT-3.5)	Reasoning accuracy
Standard zero-shot^{c} [15]	0.31
Zero-shot-CoT^{c} [19]	0.61
Zero-shot-RH^{c}	**0.75**
Standard zero-shot^{b} [15]	0.78
Zero-shot-CoT^{b} [19]	0.72
Zero-shot-RH^{b}	**0.87**

TABLE 2. *Reasoning accuracy on Bard.* Proposed zero-shot-RH exhibits the best performance for both datasets, and "Let's think step by step" ranks the second best for the first dataset, but last for the second dataset.

Prompt (Bard)	Reasoning accuracy
Standard zero-shot^{c} [15]	0.43
Zero-shot-CoT^{c} [19]	0.62
Zero-shot-RH^{c}	**0.69**
Standard zero-shot^{b} [15]	0.82
Zero-shot-CoT^{b} [19]	0.71
Zero-shot-RH^{b}	**0.85**

the correct answers and reasoning generated (Fig. 5). Every question was evaluated three times; (1) without zero-shot prompt, (2) with zero-shot-CoT prompt (i.e., *"Let's think step by step"*), and (3) with the proposed zero-shot-RH prompt (i.e., *Considering the probability of A and B* for the conjunction-fallacy dataset and *Considering the inclusion relationship between A and B"* for the base-rate-neglect dataset). Therefore, in this study, each dataset was evaluated 10 times for 2,300 reasonings and answers evaluated.

B. REASONING AND ANSWER EVALUATION

Answers and reasonings for the model output were evaluated for the two categories of datasets. The answers were checked to determine whether they were correct, and the reasoning was evaluated differently for each dataset. The correct reasoning format was not fixed, and it was not appropriate to allow the machines to evaluate their own output reasoning. Therefore, humans evaluated the output reasoning of the models using certain criteria. For the first dataset, the evaluation considered the answer correct if the model not only provided the correct reasoning, but also demonstrated that understanding the relationship between answers was crucial. For the second dataset, the evaluation focused on the ability of the model to draw correct inferences without being influenced by the base rate. This implied that, if the model accurately identified relationships between the sets, such as "A being a subset of B" or "A being more specific than B," the reasoning was considered correct regardless of the base rate.

C. EXPERIMENTAL RESULTS

We evaluated the accuracy of the reasoning and answers of the LLM output for the input dataset.

1) REASONING EVALUATION RESULTS

Tables 1 and 2 summarize the reasoning accuracy results for both datasets on GPT-3.5 and Bard. Table 1 lists the results for GPT-3.5. The accuracy of the original question, which uses the zero-shot method without prompts, is 0.31. In other words, the proportion of LLMs that reasoned correctly without RHs is 0.31. The accuracy of the zero-shot-CoT [19] method with the prompt "Let's think step by step" is 0.61. This is an increase of 0.3 over that of the zero-shot. The accuracy of the proposed zero-shot-RH "considering inclusion relationship" is 0.75. This is an increase of 0.44 over that of the zero-shot prompt, and 0.14 over that of the first CoT prompt. The accuracy of reasoning of the prompts that provide a way to reason is higher than that of the zero-shot method without any prompts. Table 2 presents the results for Bard. The accuracy of the original question, which uses a zero-shot method without prompts, is 0.43. The accuracy of the proposed zero-shot-RH "considering inclusion relationship" is 0.69, which is an increase of 0.07 over that of zero-shot, and 0.26 over that of the first CoT prompt.

For the second dataset, that is, base rate neglect, the accuracies of the original question on GPT-3.5 and Bard are 0.78 and 0.82, and those of the prompt "Let's think step by step" are 0.72 and 0.71, respectively. It decreases by 0.06 and 0.11 compared with zero-shot. The accuracies of the proposed prompt "considering the probability of A and B" are 0.87 and 0.85, respectively. They increase by 0.09 and 0.03 compared with those zero-shot, and by 0.15 and 0.14 over those of the first CoT prompt, respectively. Unlike the previous results, the accuracy decreases in the prompt added with "Let's think step by step". The output reasoning is developed incorrectly by "step by step." The LLMs exhibit representative heuristics in zero-shot experiments on the two datasets.

2) ANSWER EVALUATION RESULTS

For the first dataset (Table 3 and 4), the accuracies of the original question are 0.39 and 0.42 on GPT-3.5 and Bard, respectively. The accuracies of the prompt "Let's think step by step" are 0.64 and 0.67, respectively. The accuracies of the prompt "considering inclusion relationship" are 0.77 and 0.73, respectively. The answer rate is not always accurate because the answer is correct; however, it shows a proportional relationship. As with the results of the reasoning evaluation, the zero-shot-CoT method yields better results than the simple zero-shot method, while the proposed prompt exhibits the best performance.

TABLE 3. *Answer rate on GPT-3.5.* Superscripts [c] and [b] denote the results on the conjunction-fallacy and base-rate-neglect datasets, respectively. In the first dataset, the answer rate is highest for the proposed zero-shot-RH prompt, followed by 'Let's think step by step'and the original question, as with reasoning accuracy. In the second dataset, the zero-shot-RH prompt again has the highest accuracy rate, followed by the original question and 'Let's think step by step'.

Prompt (GPT-3.5)	Answer rate
Standard zero-shot[c] [15]	0.39
Zero-shot-CoT[c] [19]	0.64
Zero-shot-RH[c]	**0.77**
Standard zero-shot[b] [15]	0.82
Zero-shot-CoT[b] [19]	0.72
Zero-shot-RH[b]	**0.88**

The results for the second dataset are 0.83 and 00.82 for the original question, 0.72 and 0.73 for the prompt "Let's think step by step," and 0.88 and 0.86 for the prompt "considering the probability of A and B," on GPT-3.5 and Bard, respectively. Here, as with the results of the reasoning evaluation, the accuracy of "Let's think step by step" is the lowest. The experimental results show that this prompt is not effective in reducing the RH of the LLMs. However, the proposed prompt exhibits the highest accuracy.

In summary, the results show that the zero-shot-RH, "considering inclusion relationship," and "considering the probability of A and B," are the most effective ways to improve the reasoning accuracy of LLMs on the conjunction-fallacy and base-rate-neglect tasks. The "Let's think step by step" prompt is less effective, particularly for the base rate neglect task.

3) GENDER BIAS ANALYSIS

Tables 5,6,7, and 8 list the gender-based accuracy of the reasoning and answers on GPT-3.5 and Bard. In Table 5, we summarized results of reasoning accuracy by gender on GPT-3.5. In the reasoning evaluation on GPT 3.5, the accuracies for males and females are 0.43 and 0.17, respectively, for the standard zero-shot [15]. For the Zero-shot-Cot prompt [19], the accuracies are 0.67 and 0.61, respectively. For the Zero-shot-RH (i.e., "considering the inclusion relationship") prompt, the accuracies are 0.9 and 0.95, respectively. When no prompts are provided, the accuracy for males is approximately 0.25 higher than that for females. While verifying the output of the model, we observe that the model is sensitive to "women" and tends to focus on gender equality rather than the inclusion relationship between sets. Consequently, the model uses correct reasoning less frequently in the case of females than in the case of males. For the first prompt, the accuracy for males is 0.06 higher than that for females. For the second prompt, the accuracy for females is 0.05 higher than that for males. The difference between the accuracies for males and females is the highest for the original question.

TABLE 4. *Answer rate on Bard.* The zero-shot-RH prompt has the highest answer rate, followed by 'Let's think step by step', and the original question in the first dataset. The zero-shot-RH prompt also scores the highest accuracy rate, followed by the original question and 'Let's think step by step'.

Prompt (Bard)	Answer rate
Standard zero-shot[c] [15]	0.42
Zero-shot-CoT[c] [19]	0.67
Zero-shot-RH[c]	**0.73**
Standard zero-shot[b] [15]	0.83
Zero-shot-CoT[b] [19]	0.73
Zero-shot-RH[b]	**0.86**

TABLE 5. *Reasoning accuracy by gender on GPT-3.5.* All males have higher reasoning accuracy, except for the proposed zero-shot-RH. The difference in accuracies between the sexes is the greatest when CoT prompts are not used.

Prompt (GPT-3.5)	male	female
Standard zero-shot [15]	0.43	0.17
Zero-shot-CoT [19]	0.67	0.61
Zero-shot-RH	0.9	0.95

TABLE 6. *Reasoning accuracy by gender on Bard.* All males have higher reasoning accuracy than females.

Prompt (Bard)	Male	Female
Standard zero-shot [15]	0.39	0.23
Zero-shot-CoT [19]	0.72	0.68
Zero-shot-RH	0.85	0.84

In Table 6, we summarized reasoning accuracy by gender on Bard. The accuracies for males and females on Bard are 0.39 and 0.23, respectively, for the standard zero-shot [15]. For the Zero-shot-Cot prompt [19], the accuracies are 0.72 and 0.68, respectively. For Zero-shot-RH, the accuracies are 0.85 and 0.84, respectively.

In Table 7, we summarized answer rate by gender on GPT-3.5. In the answer evaluation on GPT-3.5, the accuracies for males and females are 0.55 and 0.17, respectively, for the standard zero-shot [15]. For the Zero-shot-CoT [19] prompt, the accuracies are 0.72 and 0.63, respectively. For the Zero-shot-RH (i.e., "considering the inclusion of relationship") prompt, the accuracies are 0.9 and 0.95, respectively. Similar to the reasoning evaluation, the difference between the accuracies for males and females is most pronounced for the original question, with a gap of 0.38. For the first prompt, the accuracy for males is 0.11 higher than that for females. For the second prompt, the accuracy for females is 0.05 higher than that for males.

In table 8, we summarized answer rate by gender on Bard. In the answer evaluation on Bard, the accuracies for males and females are 0.43 and 0.27, respectively, for the standard zero-shot [15]. For the Zero-shot-CoT [19] prompt, the accuracies

TABLE 7. *Answer rate by gender on GPT-3.5.* **Answer rate is higher for all males, except for the zero-shot-RH, as with the reasoning accuracy. The difference in accuracy between the sexes is also greatest in zero-shot, without using CoT.**

Prompt (GPT-3.5)	Male	Female
Standard zero-shot [15]	0.55	0.17
Zero-shot-CoT [19]	0.72	0.63
Zero-shot-RH	0.91	0.95

TABLE 8. *Answer rate by gender on Bard.* **The answer rate is higher for all males as with reasoning accuracy.**

Prompt (Bard)	Male	Female
Standard zero-shot [15]	0.43	0.27
Zero-shot-CoT [19]	0.73	0.65
Zero-shot-RH	0.95	0.93

are 0.73 and 0.65, respectively. For the Zero-shot-RH prompt, the accuracies are 0.95 and 0.93, respectively.

In both reasoning and answer evaluations, we can verify the validity of our proposed prompt, zero-shot RH, by observing the higher accuracies for zero-shot RH than those for standard zero-shot or Zero-shot-CoT.

V. CONCLUSION

This study proposed various experiments and new techniques to explore and overcome the limitations of the abstract thinking capabilities of LLMs. We experimentally verified the RH error and presented a new approach to address it. Experimental results showed that the proposed prompt, Zero-shot-RH, was the most effective in improving the reasoning accuracy of the language model in logical reasoning tasks, such as linking errors and ignoring default probabilities. In particular, the prompts "Consider the inclusion relationship" and "Consider the probabilities of A and B" demonstrated outstanding performance, surpassing "Let's think step by step." Zero-shot-RH prompts also yields higher accuracy in the task of avoiding gender-based RH errors. These research results can provide important guidelines for understanding and resolving RH errors in language models.

VI. LIMITATIONS

This study has various limitations that can be resolved. First, the reasoning behind the LLM output was evaluated by humans, which was time-consuming and resource-intensive. Although the evaluation was conducted using clear criteria, it was difficult to achieve the same accuracy as that of automated evaluation.

Second, this study used two datasets comprising 230 questions, which was smaller than the datasets used for fine-tuning. The datasets were of the same type; therefore, they were sufficient for the experiment. Nevertheless, more datasets could have been generated to produce more generalized results.

Third, the study did not investigate why the LLM exhibited RH errors. The LLM reflects the output of a text training set, which in turn reflects human cognitive biases. To demonstrate that the LLM exhibits such results owing to the reflection of human cognitive biases, the bias of the text training set must be proved.

To address these limitations, future work will focus on the following:

- Automating the evaluation of reasoning—This could save time and resources and lead to a more accurate evaluation.
- Generating more datasets—This would help produce more generalized results.
- Investigating the reason for the RH—This would provide a better understanding of how LLMs work and how they can be improved.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[2] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," Open AI, USA, Tech. Rep. 1, 2018.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[4] T. B. Brown et al., "Language models are few-shot learners," in *Proc. NIPS*, 2020, pp. 1877–1901.

[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[6] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[7] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, Jul. 2004, pp. 74–81.

[8] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.

[9] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," 2016, *arXiv:1606.05250*.

[10] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," 2018, *arXiv:1806.03822*.

[11] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," 2018, *arXiv:1804.07461*.

[12] R. Sennrich, B. Haddow, and A. Birch, "Edinburgh neural machine translation systems for WMT 16," 2016, *arXiv:1606.02891*.

[13] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.

[14] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands. New York, NY, USA: Springer, 2016, pp. 382–398.

[15] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, Jul. 2018.

[16] J. Wei, X. Wang, D. Schuurman, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 24824–24837.

[17] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty," *Science*, vol. 185, no. 4157, pp. 1124–1131, Sep. 1974.
[18] G. Suri, L. Slater, A. Ziaee, and M. Nguyen, "Do large language models show decision heuristics similar to humans? A case study using GPT-3.5," *J. Exp. Psychol., Gen.*, vol. 153, no. 4, pp. 1066–1075, 2023.
[19] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Proc. Adv. neural Inf. Process. Syst.*, vol. 35, 2022, pp. 22199–22213.
[20] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surveys*, vol. 55, no. 9, pp. 1–35, Sep. 2023.



JONGWON RYU (Student Member, IEEE) received the B.S. degree in mathematics from Dankook University, Cheonan, Republic of Korea, in 2023. He is currently pursuing the M.S. degree with Chung-Ang University. His research interests include natural language processing and multimodal reasoning.



JUNGEUN KIM (Member, IEEE) received the B.S. degree in intelligent mechatronics engineering from Sejong University, Seoul, Republic of Korea, in 2022. She is currently pursuing the M.S. degree with Chung-Ang University. Her research interests include causal reasoning, visual–language reasoning, and various multimodal reasoning.



JUNYEONG KIM (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from KAIST, Republic of Korea, in 2015, 2017, and 2021, respectively. He is currently a Postdoctoral Researcher with the Artificial Intelligence and Machine Learning Laboratory, School of Electrical Engineering, KAIST. He is also an Assistant Professor with the Department of AI, Chung-Ang University, Republic of Korea. His research interests include visual–language reasoning, answering visual questions, and video-based problems.