Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models

Yuyan Chen*
chenyuyan21@m.fudan.edu.cn
Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
Shanghai, China

Qiang Fu†
qifu@microsoft.com
Microsoft
Beijing, China

Yichen Yuan
axclbkj@gmail.com
Shanghai Key Laboratory of Data Science
Shanghai, China

Zhihao Wen
zhwen.2019@phdcs.smu.edu.sg
Singapore Management University
Singapore, Singapore

Ge Fan
ge.fan@outlook.com
Tencent
Shenzhen, China

Dayiheng Liu
liudayiheng.ldyh@alibaba-inc.com
DAMO Academy
Hangzhou, China

Dongmei Zhang
dongmeiz@microsoft.com
Microsoft
Beijing, China

Zhixu Li†
zhixuli@fudan.edu.cn
Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
Shanghai, China

Yanghua Xiao†
shawyh@fudan.edu.cn
Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Fudan-Aishu Cognitive Intelligence Joint Research Center
Shanghai, China

ABSTRACT

Large Language Models (LLMs) have gained widespread adoption in various natural language processing tasks, including question answering and dialogue systems. However, a major drawback of LLMs is the issue of hallucination, where they generate unfaithful or inconsistent content that deviates from the input source, leading to severe consequences. In this paper, we propose a robust discriminator named RelD to effectively detect hallucination in LLMs' generated answers. RelD is trained on the constructed RelQA, a bilingual question-answering dialogue dataset along with answers generated by LLMs and a comprehensive set of metrics. Our experimental results demonstrate that the proposed RelD successfully detects hallucination in the answers generated by diverse LLMs. Moreover, it performs well in distinguishing hallucination in LLMs' generated answers from both in-distribution and out-of-distribution datasets. Additionally, we also conduct a thorough analysis of the types of hallucinations that occur and present valuable insights. This research significantly contributes to the detection of reliable answers generated by LLMs and holds noteworthy implications for mitigating hallucination in the future work.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**.

KEYWORDS

Hallucination Detection, Large Language Models, Reliable Answers

ACM Reference Format:
Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023. Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3583780.3614905

1 INTRODUCTION

Large language models (LLMs) have revolutionized various fields [78], including logical reasoning [3, 40], question answering [48], code generation [30], and vertical domains [42]. However, LLMs encounter numerous challenges that hinder their optimal performance. These challenges include the inability to update knowledge in real-time [9], the lack of genuine emotion and thought [7], and the generation of long-winded and verbose answers [28], among others. Notably, one of the most critical failures is the presence of factual errors in the generated text [5], which gives rise to "Hallucinations" as depicted in Fig 1. The existence of such "Hallucinations" poses a severe hindrance to the widespread adoption of LLMs in non-chatbot scenarios, particularly in domains like medicine and finance

*Work done while this author was an intern at Microsoft Research.
†The corresponding authors.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3614905

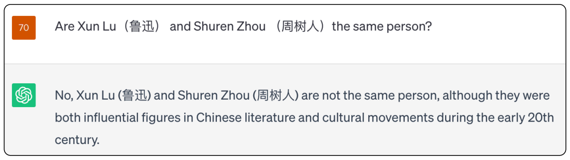

Figure 1: The given answer, produced by ChatGPT, exhibits "Hallucinations" by incorrectly treating "Shuren Zhou" and "Xun Lu" as separate individuals, despite they referring to the same person.

where factual accuracy is crucial. The potential risks associated with erroneous information can lead to significant economic losses or even jeopardize human safety [1]. Consequently, the elimination of factual errors in LLMs has become an essential requirement in both industry and academia.

The issue of hallucinations in natural text generation has long been acknowledged by researchers [27, 35, 37], and the causes of these hallucinations are complex and multifaceted. On one hand, the large-scale data corpus employed for training LLMs unavoidably contains some erroneous information, which gets learned and stored in the model parameters [45, 54, 59]. Consequently, when generating text, LLMs tend to prioritize their parameterized knowledge, thereby resulting in the production of hallucinatory content [44]. On the other hand, the decoder component of LLMs is typically trained using maximum likelihood estimation [4, 56]. During training, ground-truth serves as the input prefix for predicting subsequent tokens. However, during inference, the next token is predicted based on the generated history sequence [23]. This discrepancy in the prediction process makes it easier for hallucinations to occur.

Existing research on detecting hallucinations of LLMs' generated answers primarily encompasses statistical, model-based, and human-based evaluations [27, 34]. Statistical evaluation involves direct calculation of vocabulary matching between the generated text and reference target text, employing metrics such as ROUGE [38] and BLEU [51]. Some studies also utilize the Knowledge F1 (KF1) [65] metric to reduce knowledge hallucination in state-of-the-art chatbots. This KF1 metric is particularly suitable for detecting hallucinations in knowledge dialogue scenarios. Additionally, Shen et al. [64] conduct a large-scale assessment, including correctness and unanswerable question identification, to evaluate ChatGPT's reliability in generic question-answering scenarios. Ye et al. [76] undertake a preliminary study to assess the robustness, consistency, and credibility of LLM systems. However, these metrics rely on vocabulary matching and surface-level metrics, which may not capture semantic coherence or accurately detect hallucinations. Model-based evaluation defines the hallucination score based on the entailment probability between the source text and the generated text. This involves judging whether a hypothesis (i.e., generated text) is entailed by the premise (i.e., reference text). Model-based evaluation incorporates various metrics, including Information Extraction (IE)-based metrics, QA-based metrics [16, 63, 70], Natural Language Inference (NLI) metrics [17, 18, 25], Faithfulness Classification metrics [25, 41, 79], and LM-based metrics [19, 67]. For example, Honovich et al.[25] employ the Q^2 method of QA systems to assess the consistency between the response and external knowledge. Azaria et al.[2] utilize the internal state and hidden layer activations of LLMs to detect the truthfulness of generated statements. However, these methods lack a comprehensive set of metrics to effectively balance the advantages and disadvantages of different evaluation criteria. As a result, models often rely heavily on single labels without considering a broader range of factors. Human-based evaluation involves scoring hallucinatory text or directly comparing it with the ground truth [61, 65], which inevitably increases research costs.

To address these limitations and achieve a more balanced approach, we combine automatic metrics with model-based evaluation, which aims to align with trends observed in human evaluation scores [33]. Therefore, in this work, we focus on building a robust discriminator, RelD, which is trained on the constructed RelQA, a bilingual question-answering dialogue dataset along with answers generated by LLMs and a comprehensive set of metrics, in order to effectively detect hallucinations in the generated answers of LLMs. Specifically, the RelQA dataset comprises 274,426 samples, encompassing diverse sources such as Wikipedia, Baidu Zhidao, Bing user queries, and Chinese high school reading comprehension, etc. These datasets cover a range of domains including Wikipedia, news, education, and stories, utilizing various formats such as extractive reading comprehension and multiple-choice questions. To comprehensively evaluate LLMs' generated answers in the RelQA dataset, we adopt a set of comprehensive metrics, including LLM-assessment metrics, human metrics, machine metrics, and composite metrics. Additionally, we introduce a novel and robust discriminator, RelD, which is trained on RelQA, to detect hallucinations and analyze the types of them present in the generated answers of LLMs. Our experimental results demonstrate that RelD performs admirably in detecting hallucinations across diverse LLMs and for both in-distribution and out-of-distribution datasets. Our contributions in this paper can be outlined as follows:

- We design a novel and robust discriminator RelD, which aims to detect hallucinations in the generated answers of various LLMs.
- In order to train RelD, we construct RelQA, a bilingual question-answering dialogue dataset along with answers generated by LLMs and a comprehensive set of metrics, including LLM-assessment metrics, human metrics, machine metrics, and composite metrics.
- Our experimental results demonstrate that the discriminator RelD effectively detects hallucinations in the answers generated by different LLMs, exhibiting proficiency in both in-distribution and out-of-distribution datasets. Additionally, we make detailed analysis for types of hallucinations and provide valuable insights into the underlying causes of hallucination.

2 DATA CONSTRUCTION

In this section, we present the process of constructing RelQA. We begin by using questions from various existing nine datasets as inputs to different LLMs to generate corresponding answers. Next, we design a comprehensive set of metrics to evaluate the reliability of these generated answers. The combined collection of the original nine datasets, the generated answers by LLMs, and the evaluation metrics is referred to as RelQA. RelQA is used to train a discriminator RelD.

2.1 DATA COLLECTION

RelQA consists of nine sub-datasets: SQuAD [55], DuReader [24], HotpotQA [75], MSMARCO [46], NewsQA [69], QuAC [11], CoQA [58], TriviaQA-Web [29], and TriviaQA-Wikipedia [29]. The detailed collecting steps are as follows:

Step 1 (Dataset Selection): These datasets are selected due to their unique characteristics, diverse sources, and the enrichment they bring to the overall collection. They cover extractive reading comprehension (ERC), multiple-choice (MC), and multi-turn dialogues (MTD) categories. They originate from sources such as Wikipedia, Baidu Zhidao, Bing search, and other platforms, while encompassing domains such as student education, news, web articles, and general knowledge.

Step 2 (Formatting and Integration): To ensure compatibility and remove dataset boundaries, we perform formatting and integration for all selected datasets based on the aforementioned categories. Each dataset follows a specific standardized format, as illustrated in Table 1 (the second column). We represent the datasets of all categories as $\{L_i, D_i\}$, where L_i denotes a specific dataset and D_i denotes its standardized format.

Step 3 (Preprocessing): To facilitate effective processing and generation of answers, we employ preprocessing techniques on the dataset. This involves two primary aspects: personalized prompt instruction design and addressing the limitations associated with long texts. For personalized prompt instruction design, we create question-adaptive prompt instructions for each question based on the question type, as shown in Table 1 (the third column). These prompt instructions guide LLMs in generating better answers that align with different types of questions. To address the challenge of long texts, we implement a sliding window approach [31], segmenting the texts into smaller windows, each containing 4,000 tokens. This ensures that LLMs receive clear prompt instructions and can effectively handle texts of varying lengths, resulting in more accurate and contextually appropriate answers.

Step 4 (Answer Generation): We employ several powerful LLMs, including LLaMA [68], BLOOM [62], GPT-J [71], GPT-3 [6], and GPT-3.5 [1], to generate answers for evaluation. In the case of longer texts, we slide the window over the text and generate outputs for each window. The generated outputs for each window are stored to facilitate subsequent filtering and selection of the optimal answers. To maintain answer stability, we ask an LLM to generate the answer three times for each question and select the majority answer as the final answer. Furthermore, to ensure the overall quality and reliability of the generated answers, we conduct quality assurance procedures, including automated checks to identify and re-generate incomplete sentences by detecting missing sentence-ending punctuation, among others.

2.2 METRIC SELECTION

To evaluate the reliability of LLMs' generated answers, it is crucial to select appropriate metrics that capture different aspects of answer quality. We employ four types of metrics, including LLM-assessment metric, human metric, machine metric, and composite metric, to comprehensively evaluate the generated answers.

LLM-assessment metric is inspired by the concept of LLMs' self-evaluation, where LLMs occasionally demonstrate the ability to assess their own output correctly without human intervention [10, 74]. This metric comprises two specific indicators: the goodness of a generated answer and the similarity between the generated answer and the ground-truth answer. By obtaining the goodness score and similarity score of a generated answer, we can evaluate its quality and how closely it aligns with the ground-truth answer. Higher scores indicate better quality and semantic alignment. The LLM-assessment metric provides valuable insights into the LLMs' ability to evaluate the quality of generated answers.

Human metric plays a significant role in evaluating the LLM's performance from a human perspective. It includes a human score, which is a binary label assigned to each answer based on the degree of match between the LLM's generated answer and the ground-truth answer, along with the assigned goodness score. The human metric labeling is as follows: i) When the LLM's generated answer is the same as the ground-truth answer and receives a goodness score of 4 or 5, the human metric is labeled as 1. This indicates that the LLM has successfully generated a correct and high-quality answer that aligns with the expected answer. ii) When the LLM's generated answer is different from the ground-truth answer and receives a goodness score of 1, 2, or 3, the human metric is labeled as 2. This suggests that the LLM's generated answer is incorrect or of lower quality compared to the ground-truth answer. iii) For cases where the LLM's generated answer neither matches the ground-truth answer nor falls within the aforementioned goodness score ranges, the human metric is labeled as 0. This label represents a neutral or ambiguous classification, indicating that the answer may require further examination or subjective judgment. The human metric captures the human perception of the LLM's performance.

Machine metric draws inspiration from question-answering and dialogue systems, which rely on objective metrics to assess the quality of generated answers. It encompasses various categories, including accuracy metrics, overlap metrics, similarity metrics, and diversity metrics. Examples of machine metrics include F1 score, Recall, BLEU [51], BERT score [77], ROUGE (ROUGE-1, ROUGE-2, ROUGE-L) [38], Distinct-N (Distinct-1, Distinct-2) [36], Greedy matching, and Embedding scores (average, extreme) [39]. Specifically, accuracy metrics assess the correctness of generated answers compared to the ground truth, including F1 score. Overlap metrics measure the overlap between generated answers and the ground truth, including BLEU, Recall, ROUGE. Similarity metrics capture the semantic similarity between generated answers and the ground truth, including BERT score, Greedy matching and Embedding scores (average, extreme). Diversity metrics measure the diversity of the generated answers, including Distinct-N. These metrics objectively evaluate the semantic alignment, relevance, diversity, and quality of generated answers, enabling a comprehensive assessment of LLMs' answers.

Composite metric is designed to provide a comprehensive evaluation of a model's performance by combining multiple aspects. It includes a final score and a final tag to summarize the evaluation. Each of the metrics mentioned above contributes to the final score, with specific emphasis given to certain metrics. For instance, Recall

[1]https://chat.openai.com/

Table 1: The format and prompt instuctions of three types of datasets. a_i: the answer in ERC or MTD, or the correct answer in MC. a'_i: the wrong answers in MC.

Type	Format	Prompt instruction
ERC	$D_i = \{c_i, q_i, a_i\}$	Given the following context c_i and the question q_i. Please provide the answer.
MC	$D_i = \{c_i, q_i, a_i, a'_i\}$	Given the following context c_i and the question q_i. Please select the best answer from the candidate answers $\{a_i, a'_i\}$.
MTD	$D_i = \{h_i, q_i, a_i\}$	Given the history conversation h_i and the current question q_i. Please provide the answer.

Table 2: The distribution of each dataset in RelQA on LLM-assessment metric.

Dataset	Goodness Low	Goodness Medium	Goodness High	Similarity Low	Similarity Medium	Similarity High
SQuAD	0.11%	0.42%	99.47%	33.71%	2.50%	63.8%
DuReader	2.77%	5.60%	91.63%	15.73%	34.01%	50.26%
HotpotQA	1.47%	1.35%	97.18%	37.57%	5.52%	56.9%
MSMARCO	1.62%	2.43%	95.95%	13.58%	11.53%	74.89
NewsQA	0.66%	0.91%	98.43%	21.67%	25.44%	52.89
QUAC	8.87%	8.41%	82.72%	60.28%	18.3%	21.41
CoQA	1.37%	3.08%	95.55%	18.45%	7.43%	74.13
TriviaQA-web	1.25%	0.63%	98.12%	31.18%	6.16%	62.66
TriviaQA-wiki	1.36%	0.66%	97.99%	31.36%	6.54%	62.11

Table 3: The distribution of each dataset in RelQA on Human metric.

Dataset	Human score Reliable	Human score Unreliable	Human score Ambiguous
SQuAD	32.79%	0.49%	66.71%
DuReader	0.42%	8.31%	91.27%
HotpotQA	19.75%	2.73%	77.52%
MSMARCO	6.95%	3.99%	89.06%
NewsQA	2.09%	1.53%	96.38%
QUAC	0.81%	17.16%	82.03%
CoQA	8.08%	4.22%	87.71%
TriviaQA-web	25.77	1.75	72.49%
TriviaQA-wiki	24.29	1.87	73.84%

and ROUGE (ROUGE-1, ROUGE-2, ROUGE-L) may be assigned higher weights (e.g., twice the weight) to highlight the importance of maintaining information. The weights of different metrics can be dynamically optimized to better assess their importance in real-world scenarios as demonstrated in Experiment 4.3. The final tag is a binary label assigned based on the average score. If the average score is greater than 0.5, it is labeled as 1; otherwise, it is labeled as 0. The final tag simplifies the evaluation outcome, indicating whether the LLMs' generated answer is considered reliable or not. In summary, these metrics collectively evaluate the quality of answers generated by LLMs compared to the ground-truth answers.

2.3 DATA EXPLORATORY ANALYSIS

In this section, we conduct a data exploratory analysis of the constructed RelQA dataset, which comprises a total of 1,372,130 samples, including generated answers by five selected LLMs. Among these, 743,910 samples are assigned as reliable and 628,220 samples as unreliable based on the final tag metric. We divide the possible ranges of all metrics into three equal parts, representing low, medium, and high levels. Fig 2 illustrates the distribution of each dataset at the high level for each metric. We also present the distributions of different datasets among various metrics as shown in Table 2, Table 3, Table 4 and Table 5.

First, we analyze the differences in the LLM-assessment metric across different datasets. Regarding the "goodness" metric, the QUAC dataset performs poorly in terms of answer quality, with a high score percentage of 82.72%, while the SQuAD dataset excels in generating high-quality answers, with a high score percentage

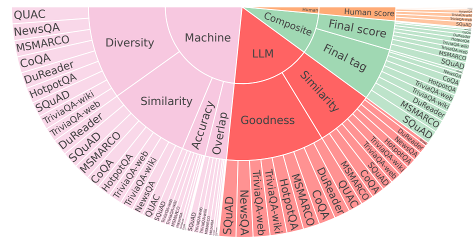

Figure 2: A data exploratory analysis of the constructed RelQA based on different metrics.

of 99.47%. Other datasets generally achieve high score percentages above 90%. Regarding the "similarity" metric, the MSMARCO dataset demonstrates the highest similarity to the reference answers, with a high similarity percentage of 74.89%. Conversely, the QUAC dataset also performs poorly in terms of similarity, with a low similarity percentage of 60.28%.

Next, we analyze the differences in the human metric across different datasets. The proportions of reliable evaluations vary significantly in the "human score" metric. The lowest proportion is 0.42% for DuReader-master, while the highest is 32.79% for SQuAD. Similarly, the proportions of unreliable evaluations differ, with the lowest being 0.49% for SQuAD and the highest being 17.16% for QUAC. Additionally, the proportion of ambiguous evaluations is highest for newsQA at 96.38% and lowest for QUAC at 66.71%.

Afterwards, we analyze the differences in the machine metric across different datasets. In terms of "accuracy metrics", the QUAC dataset performs the worst, with a high score percentage of only 4.54%. The high score percentages for other datasets range between 4.54% and 30.8%, with a median around 20%. In terms of "overlap metrics", the QUAC dataset also performs poorly in terms of low overlap, with a low score percentage of 87.52%. The low score percentages for other datasets range from 32.47% to 75.28%, with no significant high scores observed overall. Regarding "similarity metrics", DuReader, SQuAD, and MSMARCO perform well in terms of high similarity scores, with the highest scores being 95.89%, 94.71%, and 93.41% respectively. In contrast, newsQA and QUAC exhibit lower similarity scores, with the highest scores being 66.6% and 64.13% respectively. Notably, there are consistencies between the similarity scores in machine metrics and the similarity scores in LLM-assessment metrics. In "diversity metrics", QUAC, newsQA, and MSMARCO perform well in terms of high diversity scores, with the highest scores being 97.77%, 96.83%, and 94.97% respectively. This is likely due to the higher question diversity in these datasets, allowing models to exhibit more creativity and diversity in generating answers. Other datasets also maintain high diversity scores, all above 80%.

Finally, we analyze the differences in composite evaluation metrics across different datasets. In terms of the "final score" metric,

Table 4: The distribution of each dataset in RelQA on Machine metric.

Dataset	Accuracy			Overlap			Similarity			Diversity		
	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High
SQuAD	25.27%	30.03%	44.69%	32.47%	25.95%	41.58%	0.19%	5.10%	94.71%	0.00%	13.45%	86.55%
DuReader	49.35%	34.53%	16.12%	56.51%	30.81%	12.67%	0.13%	3.98%	95.89%	0.03%	5.86%	94.10%
HotpotQA	53.79%	21.49%	24.73%	60.26%	15.91%	23.83%	0.38%	20.57%	79.06%	0.00%	9.61%	90.39%
MSMARCO	33.99%	35.91%	30.09%	37.69%	35.92%	26.38%	0.19%	6.40%	93.41%	0.00%	5.03%	94.97%
NewsQA	70.53%	22.92%	6.56%	75.28%	19.04%	5.68%	1.52%	31.88%	66.60%	0.00%	3.17%	96.83%
QUAC	85.63%	9.83%	4.54%	87.52%	8.59%	3.89%	0.51%	35.36%	64.13%	0.01%	2.22%	97.77%
CoQA	56.09%	28.10%	15.81%	64.49%	22.06%	13.46%	0.54%	18.22%	81.24%	0.00%	5.77%	94.23%
TriviaQA-web	48.26%	20.93%	30.8%	54.17%	15.49%	30.34%	1.00%	20.74%	78.26%	0.00%	17.88%	82.12%
TriviaQA-wiki	47.83%	21.71%	30.46%	53.56%	16.40%	30.05%	1.06%	21.75%	77.19%	0.01%	17.73%	82.26%

Table 5: The distribution of each dataset in RelQA on Composite metric.

Dataset	Final score			Final tag	
	Low	Medium	High	Reliable	Unreliable
SQuAD	3.56%	44.01%	52.43%	78.57%	21.43%
DuReader	6.32%	67.85%	25.83%	58.57%	41.43%
HotpotQA	15.89%	56.02%	28.10%	47.75%	52.25%
MSMARCO	5.52%	51.88%	42.60%	72.29%	27.71%
NewsQA	27.65%	60.55%	11.80%	33.15%	66.85%
QUAC	50.08%	43.04%	6.88%	16.44%	83.56%
CoQA	15.43%	62.79%	21.78%	45.75%	54.25%
TriviaQA-web	17.83%	49.50%	32.67%	53.34%	46.66%
TriviaQA-wiki	19.32%	48.36%	32.33%	53.41%	46.59%

the QUAC dataset performs the worst, with a high composite score percentage of 6.88%. Conversely, the SQuAD dataset achieves the highest composite score, with a high percentage of 52.43%. It is evident that none of the datasets achieve particularly high composite scores. In terms of the "final tag" metric, the SQuAD dataset exhibits the highest proportion indicating answer reliablity, at 78.57%, while the QUAC dataset has the lowest proportion at 16.44%. This aligns with the human metric, as the SQuAD dataset primarily consists of simple extractive reading comprehension, making it easier for models to generate reliable answers. On the other hand, QUAC involves open-domain dialogue with more complex semantic understanding, posing challenges for models to generate reliable answers.

3 DISCRIMINATOR

In this section, we introduce a novel and robust discriminator called RelD, which is designed to assess the reliability of answers generated by LLMs. To ensure that RelD closely aligns with human evaluation, we employ an appropriate method to train RelD and make it fit the final score based on human evaluation. The process of constructing RelD is illustrated in Fig. 3.

3.1 REGRESSION TO MULTI-CLASS CLASSIFICATION

Initially, we employ a regression approach to train the discriminator RelD in order to fit the final score and align with human evaluation. However, our experiments reveal that the regression approach performs poorly, possibly due to the use of the mean square error loss function. Consequently, we convert the regression task into a classification task to improve the fitting. Specifically, In this process, we normalize the final score into different numbers of classes, such as four, six, eight, and ten, for multi-class classification. For instance, we assign the first category in a four-category classification to final scores ranging from 0 to 0.25. After experiments as

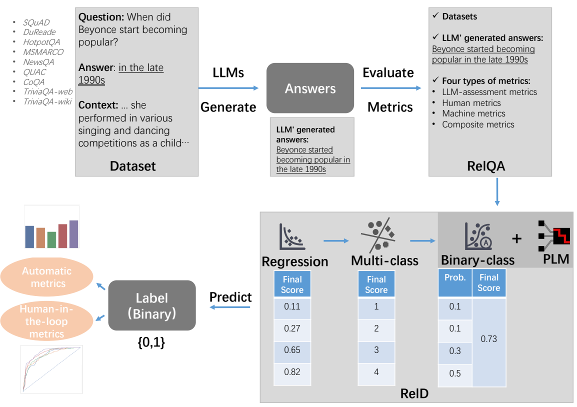

Figure 3: The process of building the discriminator RelD, which is trained on the constructed dataset RelQA and used to detect hallucination of LLMs' generated answers.

shown in Sec. 4.3, we ultimately choose a ten-class classification approach. The theoretical foundation of this method mainly lies in information theory and the cross-entropy loss function. Cross-entropy is a common information theory measure used to quantify the distance between two probability distributions. In the case of multi-classification problems, the cross-entropy loss function is defined as follows:

$$L = -\sum(y_i \cdot \log(p_i)), \tag{1}$$

where y_i represents the true label of the i-th category, and p_i represents the predicted probability of the i-th category by the discriminator RelD. Our objective is to minimize this loss function during the training of RelD. In practice, we employ the softmax function to convert the original output of RelD into a probability distribution.

One potential advantage of this method is that the classification task, which focuses on distinguishing different categories, may facilitate capturing subtle differences among the final scores. Furthermore, the cross-entropy loss function exhibits greater stability compared to the mean square error loss function when dealing with imbalanced datasets. However, it is important to note that in certain situations, multi-class tasks may introduce overly complex information, leading to a notable disparity between the concepts learned by the discriminator and human intuitive perception. For example, dividing a problem into five categories, such as "not reliable", "weakly reliable", "moderately reliable", "strongly reliable" and "highly reliable", may surpass most people's intuitive understanding of the fundamental categories of "reliable" and "unreliable".

3.2 MULTI-CLASS TO BINARY-CLASS CLASSIFICATION

Based on the aforementioned analysis, we further convert the multi-class task into a binary classification task, which may better align with human intuitive perception. Here, we present three possible approaches for this conversion, each with its theoretical support and definition:

Normalization. This method is based on threshold decision theory. It involves converting all class information into binary labels by directly normalizing the final score to 0 and 1, which serves as the final probability value for classification. However, this approach may result in some information loss as continuous scores are transformed into discrete classes.

Discrete Values. This method is grounded in maximum likelihood estimation, a commonly used parameter estimation technique in statistics. Here, we consider the highest predicted probability from the discriminator as the final probability value for classification. For example, in a four-class classification scenario, if the probabilities corresponding to the classes are 0.1, 0.1, 0.1, and 0.7, respectively, we would use 0.7 as the final probability value. The advantage of this method lies in its simplicity, although the drawback is that we do not know which class the maximum probability value corresponds to.

Weighted Average Probability. The theoretical basis for this method stems from decision theory, particularly the concept of expected utility, which involves taking a weighted average of all possible outcomes and their corresponding utilities (in this case, predicted probabilities). The goal of this approach is to determine a weighted average value that best represents the predicted probabilities for each class from the discriminator. In this method, we multiply the probability of each class predicted by the discriminator with its corresponding weight, summing them up to obtain a final probability value. This value can then be used for binary classification tasks. The formula for this method is as follows:

$$p_i' = \frac{(\sum w_i \cdot p_i) - w_{\min}}{w_{\max} - w_{\min}}, \tag{2}$$

where p_i represents the probability output of the discriminator for class i, w_i denotes the weight for class i, and $w_{\min}$ and $w_{\max}$ are the minimum and maximum weights, respectively. We set the threshold to 0.5 and use the cross-entropy loss function for approximation. It allows for a more refined fitting of regression tasks and has demonstrated better performance compared to the previous two methods, as indicated by Sec. 4.3.

3.3 Backbone of the Discriminator

We utilize a Pre-trained Language Model (PLM), such as ELECTRA [12], as the backbone of the discriminator RelD. Through our experiments, we have demonstrated that ELECTRA outperforms other PLMs, including BERT [14], RoBERTa [43], and DeBERTa [22], as indicated in Section 4.3. RelD takes questions along with contexts and LLMs' generated answers as input, generating a classification label to determine the reliability of a generated answer. It uses the weighted average probability approach to fit the ground-truth answers.

Table 6: Performance of RelD among the selected LLMs on the validation dataset.

LLM	LLaMA	BLOOM	GPT-J	GPT-3	GPT-3.5
Automatic	0.855	0.846	0.827	0.863	0.881
Human	0.826	0.830	0.835	0.869	0.894
Average score	0.841	0.838	0.831	0.866	0.888

4 EXPERIMENTS

In this section, we conduct experiments to evaluate the effectiveness of RelD in detecting the reliability of LLMs' generated answers using both automatic metrics and human-in-the-loop metrics.

4.1 EXPERIMENTAL SETUP

The experiments are conducted using TESLA A100 GPUs for answer generation and GTX 3090 GPUs for training RelD with PyTorch in Python. During the training of RelD, we set the batch size to 32 and the sequence length to 128. Hyperparameters such as weight decay (0.01), β_1 (0.9), and β_2 (0.999) are maintained. The learning rate is set to 2e-05. We train RelD for 20 epochs.

Baselines and metrics. We validate the effectiveness of the proposed RelD on well-known LLMs, including LLaMA (LLaMA-7B)[68], BLOOM (BLOOM-7B)[62], GPT-J (GPT-J-6B)[71], GPT-3[6], and GPT-3.5 [1]. To evaluate the performance of RelD, we use accuracy (ACC) as the automatic metrics and ROC curve analysis with the area under the ROC curve (AUC) as the human-in-the-loop metrics. The automatic evaluation process utilizes the final tag as the ground-truth label, while the human-in-the-loop evaluation involves human ratings as the ground-truth labels. Specifically, we randomly select 9,000 QA pairs, with 1,000 from each dataset in RelQA, for human ratings. We enroll nine volunteers and divide them into three groups to ensure evaluation stability. Each group provides scores of 0 or 1 for the randomly selected 3,000 QA pairs. Inter-rater agreement is calculated using Krippendorff's Alpha (IRA) to ensure the confidence of the human ratings. For controversial ratings with low agreement (<0.7), we discard the corresponding QA pair and replace it with another.

4.2 MAIN RESULTS

We conduct experiments to evaluate the effectiveness of the proposed RelD as follows:

Experiment 1: RelD's Performance across Different LLMs. We conduct ten-fold cross-validation and report the average performance on the validation dataset. Based on the results presented in Table 6, it's observed that both the automatic and human-in-the-loop evaluations consistently exceed 0.8 for all LLMs, with minimal variation between different models ($p<0.01$). The strong correlation between the automatic and human-in-the-loop evaluations ($p<0.01$) suggests that the automatic scoring of the RelQA dataset could largely replace human scoring. It also indicates the robustness of RelD in detecting the reliability of different LLMs.

Experiment 2: RelD's Performance on IID and OOD Datasets We evaluate the performance of RelD on both In-distribution (IID) and Out-of-distribution (OOD) datasets. We randomly assign nine datasets from RelQA to the IID and OOD sets in various ratios, such as 1:8, 2:7, 3:6, and 4:5, and vice versa. For example, we train on 8

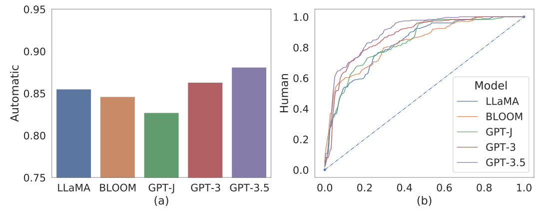

Figure 4: The visualization of RelD’s performance among the selected LLMs on the validation dataset, including both automatic and human-in-the-loop metrics.

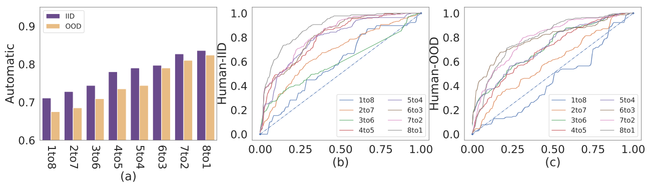

Figure 5: Performance of RelD on automatic metrics (a) and human-in-the-loop metrics (b)(c), including results on IID validation dataset (b) and OOD dataset (c) among the selected LLMs.

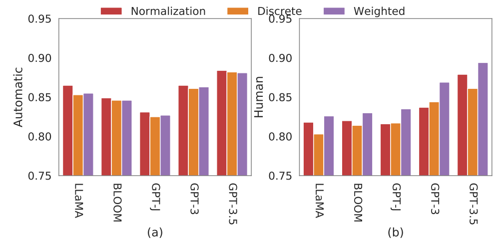

Figure 6: The performance of using weighted average probability is compared with using normalization and discrete values in automatic metrics (a) and human-in-the-loop metrics (b) on the validation dataset among the selected LLMs.

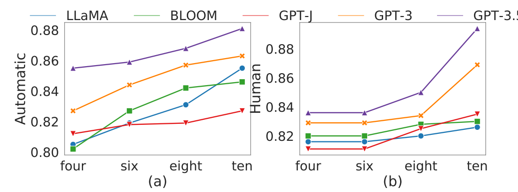

Figure 7: The performance of different numbers of categories in automatic (a) and human-in-the-loop metrics (b) on the validation dataset among the selected LLMs.

datasets and validate on 1 dataset. To ensure a balanced quantity of data in both the IID and OOD sets, we perform downsampling by randomly selecting 3,000 samples from each dataset. Considering that different datasets acting as IID or OOD may yield different results, we conduct five experiments for each ratio group and provide average values along with the range of error. This approach allows us to accurately assess the generalization ability of RelD. To evaluate the performance on the IID dataset, we use 30% of the IID data as a validation dataset. For the OOD evaluation, we directly test RelD on the entire OOD dataset. The results are presented in Table 7 and Fig. 5. We observe that when the IID ratio is set to 5 or higher, RelD consistently achieves automatic and human-in-the-loop evaluations above 0.7 on both the IID and OOD datasets. This indicates that RelD exhibits a strong generalization capability in handling OOD data as well as alignment with human evaluation predictions.

4.3 ABLATION STUDY

After that, we conduct several experiments to evaluate the effectiveness of different modules in the proposed RelD. All results are performed on the validation dataset using ten-fold cross-validation.

Experiment 3: Effectiveness of Weighted Average Probability. We compare the performance of using normalization, discrete values, and weighted average probability in the conversion from multi-class to binary-class classification in both automatic and human-in-the-loop metrics. The results are presented in Fig. 6. We observe that while using weighted average probability slightly underperforms normalization in terms of automatic metrics, it significantly outperforms normalization and discrete values in human-in-the-loop metrics across all LLMs. Therefore, we adopt weighted average probability as it offers a more intuitive and aligned approach from a human perspective.

Experiment 4: Optimal Number of Categories. We investigate the impact of the number of categories when converting regression into multi-class classification. We test four categories, six categories, eight categories, and ten categories. The results are shown in Fig. 7. It is evident that a higher number of categories leads to improved performance in human-in-the-loop metrics. This suggests that a larger number of categories brings the classification task closer to regression and enhances alignment with human cognition. Consequently, we ultimately convert the regression task into a ten-category classification task and then discern it as a binary classification using weighted average probability.

Experiment 5: Optimizing Weights of Each Metric Relying solely on prior knowledge to determine the weights of each metric may not achieve the best performance. Therefore, we explore the optimal weights for each metric. To achieve this, we calculate the optimal weight for each metric as the weighted average of two values: the AUC when each metric is treated as the ground-truth compared to human evaluation, and the Pearson coefficient between each metric and human evaluation. In our experiment, we set the ratio for the former as 0.9 and for the latter as 0.1, as it yields the best performance. The optimal weights of each metric are depicted in Fig. 8(a). Subsequently, we evaluate whether the optimal weights can enhance the performance of RelD in detecting hallucination of LLMs’ generated answers as shown in Fig. 8(b)(c). Remarkably, we observe improvements in both automatic (b) and human-in-the-loop metrics (c) after optimizing the weights of each metric.

Experiment 6: Backbone Selection for RelD. We experiment with different PLMs, including BERT [14], RoBERTa [43], DeBERTa [22], and ELECTRA [12], for RelD in order to choose the most effective backbone, as shown in Table 8. Through this comparison, we observe that ELECTRA achieves the best performance in both automatic and human-in-the-loop metrics. Consequently, we select ELECTRA as the preferred backbone for RelD.

4.4 EXPLORATORY ANALYSIS

We classify the predictions generated by RelD into four categories, as presented in Table 9. To gain insights into the characteristics

Table 7: Performance of RelD on IID and OOD datasets. IID results are based on a 30% validation dataset from the IID dataset, while OOD results are obtained from the entire OOD dataset.

LLM	Metrics	Distribution	1 to 8	2 to 7	3 to 6	4 to 5	5 to 4	6 to 3	7 to 2	8 to 1	Average
LLaMA	**Automatic**	IID	$0.698_{\pm.021}$	$0.723_{\pm.018}$	$0.762_{\pm.029}$	$0.785_{\pm.016}$	$0.776_{\pm.012}$	$0.806_{\pm.022}$	$0.821_{\pm.014}$	$0.832_{\pm.010}$	$0.775_{\pm.018}$
		OOD	$0.672_{\pm.023}$	$0.675_{\pm.020}$	$0.701_{\pm.017}$	$0.747_{\pm.011}$	$0.735_{\pm.028}$	$0.798_{\pm.026}$	$0.815_{\pm.024}$	$0.820_{\pm.013}$	$0.745_{\pm.020}$
	Human	IID	$0.550_{\pm.019}$	$0.693_{\pm.027}$	$0.721_{\pm.015}$	$0.758_{\pm.010}$	$0.763_{\pm.023}$	$0.791_{\pm.018}$	$0.839_{\pm.011}$	$0.862_{\pm.025}$	$0.747_{\pm.019}$
		OOD	$0.487_{\pm.021}$	$0.547_{\pm.017}$	$0.585_{\pm.029}$	$0.634_{\pm.022}$	$0.732_{\pm.015}$	$0.748_{\pm.014}$	$0.73_{\pm.027}$	$0.744_{\pm.012}$	$0.651_{\pm.020}$
BLOOM	**Automatic**	IID	$0.701_{\pm.024}$	$0.729_{\pm.026}$	$0.755_{\pm.013}$	$0.790_{\pm.017}$	$0.777_{\pm.020}$	$0.801_{\pm.028}$	$0.817_{\pm.016}$	$0.827_{\pm.011}$	$0.775_{\pm.019}$
		OOD	$0.674_{\pm.018}$	$0.678_{\pm.021}$	$0.705_{\pm.012}$	$0.750_{\pm.010}$	$0.739_{\pm.019}$	$0.799_{\pm.016}$	$0.817_{\pm.023}$	$0.822_{\pm.015}$	$0.748_{\pm.017}$
	Human	IID	$0.539_{\pm.013}$	$0.680_{\pm.014}$	$0.695_{\pm.028}$	$0.747_{\pm.019}$	$0.759_{\pm.022}$	$0.778_{\pm.024}$	$0.834_{\pm.012}$	$0.854_{\pm.011}$	$0.736_{\pm.018}$
		OOD	$0.462_{\pm.0120}$	$0.521_{\pm.011}$	$0.546_{\pm.025}$	$0.628_{\pm.016}$	$0.731_{\pm.023}$	$0.725_{\pm.017}$	$0.732_{\pm.020}$	$0.725_{\pm.027}$	$0.634_{\pm.019}$
GPT-J	**Automatic**	IID	$0.673_{\pm.027}$	$0.710_{\pm.015}$	$0.757_{\pm.016}$	$0.765_{\pm.014}$	$0.788_{\pm.011}$	$0.810_{\pm.029}$	$0.831_{\pm.021}$	$0.830_{\pm.018}$	$0.771_{\pm.019}$
		OOD	$0.685_{\pm.016}$	$0.677_{\pm.012}$	$0.706_{\pm.022}$	$0.746_{\pm.020}$	$0.733_{\pm.011}$	$0.795_{\pm.017}$	$0.812_{\pm.014}$	$0.810_{\pm.026}$	$0.746_{\pm.017}$
	Human	IID	$0.556_{\pm.019}$	$0.660_{\pm.010}$	$0.698_{\pm.015}$	$0.726_{\pm.024}$	$0.759_{\pm.022}$	$0.778_{\pm.013}$	$0.804_{\pm.021}$	$0.850_{\pm.018}$	$0.729_{\pm.018}$
		OOD	$0.451_{\pm.026}$	$0.523_{\pm.013}$	$0.557_{\pm.024}$	$0.605_{\pm.012}$	$0.731_{\pm.011}$	$0.725_{\pm.020}$	$0.733_{\pm.028}$	$0.721_{\pm.023}$	$0.631_{\pm.020}$
GPT-3	**Automatic**	IID	$0.706_{\pm.020}$	$0.716_{\pm.018}$	$0.768_{\pm.019}$	$0.780_{\pm.010}$	$0.769_{\pm.013}$	$0.809_{\pm.017}$	$0.825_{\pm.021}$	$0.826_{\pm.016}$	$0.775_{\pm.017}$
		OOD	$0.681_{\pm.015}$	$0.680_{\pm.014}$	$0.710_{\pm.010}$	$0.753_{\pm.011}$	$0.729_{\pm.026}$	$0.792_{\pm.019}$	$0.813_{\pm.012}$	$0.815_{\pm.024}$	$0.747_{\pm.016}$
	Human	IID	$0.527_{\pm.028}$	$0.645_{\pm.016}$	$0.731_{\pm.010}$	$0.745_{\pm.023}$	$0.782_{\pm.017}$	$0.793_{\pm.026}$	$0.836_{\pm.015}$	$0.897_{\pm.013}$	$0.745_{\pm.019}$
		OOD	$0.468_{\pm.024}$	$0.568_{\pm.018}$	$0.612_{\pm.011}$	$0.619_{\pm.020}$	$0.720_{\pm.013}$	$0.775_{\pm.012}$	$0.756_{\pm.019}$	$0.728_{\pm.014}$	$0.656_{\pm.016}$
GPT-3.5	**Automatic**	IID	$0.711_{\pm.010}$	$0.728_{\pm.012}$	$0.744_{\pm.015}$	$0.780_{\pm.014}$	$0.790_{\pm.018}$	$0.797_{\pm.027}$	$0.827_{\pm.010}$	$0.836_{\pm.021}$	$0.777_{\pm.016}$
		OOD	$0.675_{\pm.016}$	$0.685_{\pm.024}$	$0.709_{\pm.010}$	$0.735_{\pm.017}$	$0.744_{\pm.020}$	$0.790_{\pm.028}$	$0.810_{\pm.016}$	$0.824_{\pm.011}$	$0.747_{\pm.018}$
	Human	IID	$0.586_{\pm.027}$	$0.677_{\pm.012}$	$0.746_{\pm.013}$	$0.797_{\pm.014}$	$0.786_{\pm.018}$	$0.812_{\pm.023}$	$0.821_{\pm.012}$	$0.880_{\pm.011}$	$0.763_{\pm.016}$
		OOD	$0.445_{\pm.019}$	$0.592_{\pm.015}$	$0.721_{\pm.017}$	$0.722_{\pm.026}$	$0.723_{\pm.024}$	$0.791_{\pm.010}$	$0.791_{\pm.016}$	$0.795_{\pm.012}$	$0.698_{\pm.017}$

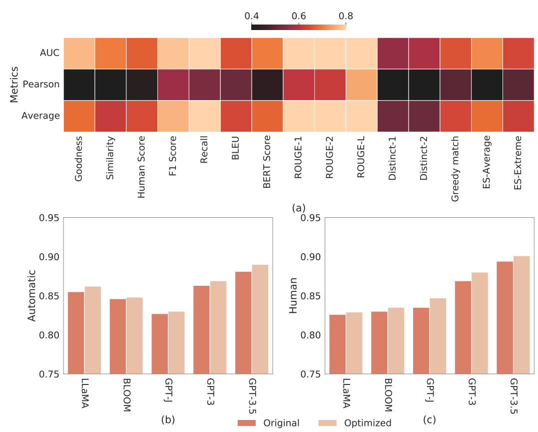

Figure 8: The optimal weights of each metric (a) and the performance of RelD with the original and optimal weights in automatic (b) and human-in-the-loop metrics (c), respectively, on the validation dataset.

Table 8: Performance of RelD with different backbones among LLMs on the validation dataset.

RelD	Metric	LLaMA	BLOOM	GPT-J	GPT-3	GPT-3.5
BERT	Automatic	0.826	0.825	0.800	0.837	0.859
	Human	0.809	0.807	0.819	0.844	0.867
RoBERTa	Automatic	0.848	0.834	0.812	0.839	0.873
	Human	0.821	0.811	0.824	0.852	0.877
DeBERTa	Automatic	0.850	0.842	0.818	0.854	0.878
	Human	0.824	0.815	0.829	0.866	0.893
ELECTRA	Automatic	**0.855**	**0.846**	**0.827**	**0.863**	**0.881**
	Human	**0.826**	**0.830**	**0.835**	**0.869**	**0.894**

of these categories and understand the functioning of RelD, we conduct an exploratory analysis.

Analysis 1: Distribution Analysis To analyze the distributions within each category, we utilize boxplots (Fig.9(a)) to illustrate key statistics such as median, quartiles, and outliers of samples. Additionally, we employ density plots (Fig.9(b)) to visualize the probability distribution of samples within each category. In the

Table 9: Four categories are defined based on the agreement between LLMs' generated answers and RelD's predictions. Q, A, P, and D represent questions, ground-truth answers, LLMs' generated answers, and RelD's predictions, respectively.

Category	Definition	Sample
1	The LLM generates correct answers, and RelD also predicts them as correct.	Q: Strabismus is more commonly known by which one-syllable word? A: squint P: squint D: True
2	The LLM generates correct answers, but RelD predicts them as incorrect.	Q: On which Apollo mission did Armstrong and Aldrin land on the moon? A: apollo 11 P: apollo 11 D: False
3	The LLM generates incorrect answers, but RelD predicts them as correct.	Q: what's the number for the metro pcs customer care line? A: customer care number for metro pcs is 8009016266 P: answer is 611 or 8009016266 or 8888638768 D: True
4	The LLM generates incorrect answers, and RelD also predicts them as incorrect.	Q: When did freestyle skiing first became a sport contested at the World Olympics? A: 1992 P: 1988 as freestyle skiing was first added as event in 1988 winter olympics D: False

first category, the boxplot exhibits a wide range and the density plot shows a concentrated distribution with multiple peaks. This suggests that RelD may have some uncertainties in its predictions for this category. For the second and third categories, the boxplot widths fall between those of the first and fourth categories and the density plots display more dispersed probability distributions. This indicates that RelD is more hesitant in its predictions or has lower proficiency in learning for these types of questions. In contrast, the fourth category exhibits a narrower boxplot and the density plot shows a concentrated probability distribution. It indicates that RelD is more confident in its predictions for this category.

Analysis 2: Clustering Analysis. By applying clustering algorithms to the text data, we investigate whether each category exhibits distinct cluster centers, as illustrated in Fig. 10. For the first category, the data distribution appears clustered and relatively uniform, indicating consistent and accurate performance by RelD within this category. The second category contains an extremely

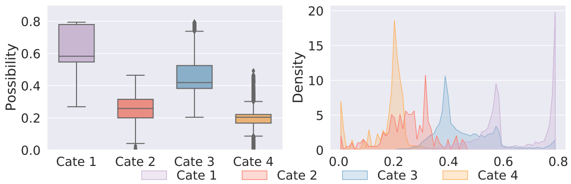

Figure 9: **The distribution of samples from each category with boxplots (a) and density plots (b). Cate: Category (The same below).**

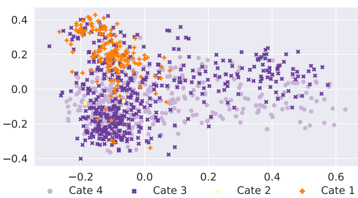

Figure 10: **Results of clustering based on four categories.**

Figure 11: **The vocabulary distribution between correctly predicted samples and incorrectly predicted samples by RelD.**

small number of samples, suggesting that RelD rarely misclassifies the correct answers generated by the LLMs. In the third category, the clustering results reveal significant variability, indicating that errors can occur in various aspects when RelD misclassifies the incorrect answer as correct, such as grammar or comprehension errors. Similarly, the fourth category displays a wide and dispersed clustering distribution, indicating diverse performance by RelD within this category. This suggests the presence of different types of errors that make it challenging for RelD to detect. From the clustering graph, we observe that RelD performs best in the first category. However, for the second, third, and fourth categories, the performance of RelD may be influenced by the complexity and ambiguity of the input contexts or questions.

Analysis 3: Vocabulary Distribution. We can compare the vocabulary distribution between correctly predicted samples and incorrectly predicted samples by RelD, as depicted in Fig. 11. There is a noticeable distinction between the left side (RelD predicts correctly) and the right side (RelD predicts incorrectly). It appears that content related to "story" is relatively easy for RelD to classify correctly, while content related to "country" poses more difficulty for RelD in accurate classification. However, it is important to note that vocabulary alone may not be the sole determining factor for RelD's recognition accuracy. The critical factors might involve underlying semantic relationships, which would necessitate further research and investigation.

5 RELATED WORK

Hallucination detection. Existing research primarily contains statistical metrics [21, 66, 72], model-based metrics (including Information Extraction (IE)-based metric, QA-based metric [25, 57, 60], Natural Language Inference (NLI) Metrics [26, 32, 73], Faithfulness Classification Metrics [25, 41, 79], LM-based Metrics [19, 67]), and human-based evaluations [61, 65]. We list some typical work as follows: Dhingra et al. [15] propose PARENT to measure hallucinations using both the source and target text as references. Goyal and Durrett [20] attempt to identify factual inconsistencies in a more fine-grained manner with a new dependency-level entailment. Liu et al. [41] and Zhou et al. [79] construct syntactic data by automatically inserting hallucinations into training instances. Chen et al. [8] and Nie et al. [47] use finer-grained metrics for intrinsic hallucination and extrinsic hallucination separately. Azaria et al. [2] utilize the internal state and hidden layer activations of LLMs to detect the truthfulness of generated statements. Ye et al. [76] consider that errors in user-generated query input may cause unexpected responses from LLMs.

Hallucination mitigation. There are also some work that focus on mitigating hallucination. For example, Dale et al. [13] and Ji et al. [27] focus on hallucination in machine translation. Pagnoni et al. [50] address hallucination in text summarization. Peng et al. [53] adopt various methods to prompt LLMs, including posting multiple queries. Ouyang et al. [49] propose a method to enhance the content generated by LLMs. Yan et al. [74] introduce an iterative self-evaluating optimization mechanism based on prompt engineering. Park et al. [52] leverage search results corresponding to a user's input query to generate an augmented query.

6 CONCLUSIONS AND FUTURE WORK

Hallucination of LLMs poses a significant challenge. In this paper, we address this issue by proposing a robust discriminator, RelD, trained on the constructed RelQA dataset, which is a bilingual question-answering dialogue dataset along with generated answers by LLMs and a comprehensive set of metrics to effectively detect hallucinations in LLMs' generated answers. Our experimental results demonstrate the effectiveness of RelD in detecting hallucinations in LLMs' generated answers. Moreover, RelD exhibits strong robustness and generalization capabilities, performing well on both in-distribution and out-of-distribution datasets. These findings make a significant contribution to the detection of reliable answers generated by LLMs and hold promising implications for future work in mitigating hallucination.

7 ACKNOWLEDGEMENT

Some of computational resource are partially supported by Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103), Science and Technology Commission of Shanghai Municipality Grant (No. 22511105902), National Key Research and Development Project (No.2020AAA0109302), National Natural Science Foundation of China (No.62072323), Shanghai Science and Technology Innovation Action Plan (No. 22511104700, 22511105902), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103), and Science and Technology Commission of Shanghai Municipality Grant (No. 22511105902).

REFERENCES

[1] Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2023. Can we trust the evaluation on ChatGPT? *arXiv preprint arXiv:2303.12767* (2023).

[2] Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When its Lying. *arXiv preprint arXiv:2304.13734* (2023).

[3] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).

[4] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems* 28 (2015).

[5] Ali Borji. 2023. A categorical archive of ChatGPT failures. *arXiv preprint arXiv:2302.03494* (2023).

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

[7] David J Chalmers. 2023. Could a large language model be conscious? *arXiv preprint arXiv:2303.07103* (2023).

[8] Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. *arXiv preprint arXiv:2104.09061* (2021).

[9] Yongcong Chen, Ting Zeng, Xiaoyi Qian, Jun Zhang, and Xinyue Chen. [n. d.]. Apreliminary STUDY ON THE CAPABILITY BOUNDARY OF LLM AND A NEW IMPLEMENTATION APPROACH FOR AGI. ([n. d.]).

[10] Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations? *arXiv preprint arXiv:2305.01937* (2023).

[11] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. *arXiv preprint arXiv:1808.07036* (2018).

[12] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. https://doi.org/10.48550/ARXIV.2003.10555

[13] David Dale, Elena Voita, Loïc Barrault, and Marta R Costa-jussà. 2022. Detecting and Mitigating Hallucinations in Machine Translation: Model Internal Workings Alone Do Well, Sentence Similarity Even Better. *arXiv preprint arXiv:2212.08597* (2022).

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/ARXIV.1810.04805

[15] Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William W Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. *arXiv preprint arXiv:1906.01081* (2019).

[16] Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754* (2020).

[17] Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021. Evaluating groundedness in dialogue systems: The begin benchmark. *arXiv preprint arXiv:2105.00071* (2021).

[18] Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2214–2220.

[19] Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. *arXiv preprint arXiv:2010.05873* (2020).

[20] Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. *arXiv preprint arXiv:2010.05478* (2020).

[21] Jian Guan and Minlie Huang. 2020. Union: An unreferenced metric for evaluating open-ended story generation. *arXiv preprint arXiv:2009.07602* (2020).

[22] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. https://doi.org/10.48550/ARXIV.2006.03654

[23] Tianxing He, Jingzhao Zhang, Zhiming Zhou, and James Glass. 2019. Exposure Bias versus Self-Recovery: Are Distortions Really Incremental for Autoregressive Text Generation? *arXiv preprint arXiv:1905.10617* (2019).

[24] Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073* (2017).

[25] Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2: Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering. *arXiv preprint arXiv:2104.08202* (2021).

[26] Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839* (2021).

[27] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[28] Peiling Jiang, Jude Rayan, Steven P Dow, and Haijun Xia. 2023. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. *arXiv preprint arXiv:2305.11473* (2023).

[29] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551* (2017).

[30] Raphaël Khoury, Anderson R Avila, Jacob Brunelle, and Baba Mamadou Camara. 2023. How Secure is Code Generated by ChatGPT? *arXiv preprint arXiv:2304.09655* (2023).

[31] Jia Jin Koay, Alexander Roustai, Xiaojin Dai, and Fei Liu. 2021. A sliding-window approach to automatic creation of meeting minutes. *arXiv preprint arXiv:2104.12324* (2021).

[32] Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics* 10 (2022), 163–177.

[33] Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems* 35 (2022), 34586–34599.

[34] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. *arXiv e-prints* (2023), arXiv–2305.

[35] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HELMA: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. *arXiv preprint arXiv:2305.11747* (2023).

[36] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055* (2015).

[37] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355* (2023).

[38] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[39] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023* (2016).

[40] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439* (2023).

[41] Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2021. A token-level reference-free hallucination detection benchmark for free-form text generation. *arXiv preprint arXiv:2104.08704* (2021).

[42] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852* (2023).

[43] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/ARXIV.1907.11692

[44] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052* (2021).

[45] Andrea Madotto, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Language models as few-shot learner for task-oriented dialogue systems. *arXiv preprint arXiv:2008.06239* (2020).

[46] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *choice* 2640 (2016), 660.

[47] Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2673–2679.

[48] Reham Omar, Omij Mangukiya, Panos Kalnis, and Essam Mansour. 2023. Chatgpt versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots. *arXiv preprint arXiv:2302.06466* (2023).

[49] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[50] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346* (2021).

[51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[52] Hyun Jin Park and Changwan Ryu. 2023. Query Augmentation Using Search Engine Results to Improve Answers Generated by Large Language Models. (2023).

[53] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813* (2023).

[54] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066* (2019).

[55] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

[56] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732* (2015).

[57] Clément Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Data-QuestEval: A referenceless metric for data-to-text semantic evaluation. *arXiv preprint arXiv:2104.07555* (2021).

[58] Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics* 7 (2019), 249–266.

[59] Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910* (2020).

[60] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637* (2020).

[61] Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *arXiv preprint arXiv:2110.05456* (2021).

[62] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).

[63] Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693* (2021).

[64] Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT. *arXiv preprint arXiv:2304.08979* (2023).

[65] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567* (2021).

[66] Hui Su, Xiaoyu Shen, Sanqiang Zhao, Xiao Zhou, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. *arXiv preprint arXiv:2005.04346* (2020).

[67] Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P Parikh. 2019. Sticking to the facts: Confident decoding for faithful data-to-text generation. *arXiv preprint arXiv:1910.08684* (2019).

[68] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[69] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830* (2016).

[70] Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228* (2020).

[71] Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax. https://github.com/kingoflolz/mesh-transformer-jax

[72] Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020. Towards faithful neural table-to-text generation with content-matching constraints. *arXiv preprint arXiv:2005.00969* (2020).

[73] Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426* (2017).

[74] Tianqiang Yan and Tiansheng Xu. 2023. Refining the Responses of LLMs by Themselves. *arXiv preprint arXiv:2305.04039* (2023).

[75] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600* (2018).

[76] Wentao Ye, Mingfeng Ou, Tianyi Li, Xuetao Ma, Yifan Yanggong, Sai Wu, Jie Fu, Gang Chen, Junbo Zhao, et al. 2023. Assessing Hidden Risks of LLMs: An Empirical Study on Robustness, Consistency, and Credibility. *arXiv preprint arXiv:2305.10235* (2023).

[77] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[78] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[79] Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Paco Guzman, Luke Zettlemoyer, and Marjan Ghazvininejad. 2020. Detecting hallucinated content in conditional neural sequence generation. *arXiv preprint arXiv:2011.02593* (2020).