HalluCounter: Reference-free LLM Hallucination Detection in the Wild!

Ashok Urlana[1,2] **Gopichand Kanumolu**[2] **Charaka Vinayak Kumar** [2]
Bala Mallikarjunarao Garlapati[2] **Rahul Mishra**[1,3]
IIIT Hyderabad[1] TCS Research, Hyderabad[2] University of Oslo, Norway[3]
ashok.u@research.iiit.ac.in, rahul.mishra@iiit.ac.in,
{ashok.urlana, gopichand.kanumolu, charaka.v, balamallikarjuna.g}@tcs.com

Abstract

Response consistency-based, reference-free hallucination detection (RFHD) methods do not depend on internal model states, such as generation probabilities or gradients, which Grey-box models typically rely on but are inaccessible in closed-source LLMs. However, their inability to capture query-response alignment patterns often results in lower detection accuracy. Additionally, the lack of large-scale benchmark datasets spanning diverse domains remains a challenge, as most existing datasets are limited in size and scope. To this end, we propose ***HalluCounter***, a novel reference-free hallucination detection method that utilizes both response-response and query-response consistency and alignment patterns. This enables the training of a classifier that detects hallucinations and provides a confidence score and an optimal response for user queries. Furthermore, we introduce ***HalluCounterEval***, a benchmark dataset comprising both synthetically generated and human-curated samples across multiple domains. Our method outperforms state-of-the-art approaches by a significant margin, achieving over 90% average confidence in hallucination detection across datasets[1].

1 Introduction

Reference-free hallucination detection (RFHD) is gaining significant traction in the research community (Manakul et al., 2023; Zhang et al., 2023; Yehuda et al., 2024), as it obviates the need for reference texts or external knowledge bases (KBs) to identify potential hallucinations. This enhances the scalability and applicability of RFHD across a broader range of tasks and scenarios, which would otherwise be constrained by reference- or KB-dependent approaches (Hu et al., 2024a; Liu et al., 2024). In the literature, RFHD approaches can be broadly categorized into two major classes. The first category, known as black-box approaches, relies on analyzing multiple responses generated by LLMs to assess consistency and alignment among them, thereby detecting hallucinations in the output (Manakul et al., 2023).

On the other hand, grey-box models leverage internal states of the models, such as decoder generation probabilities (Sennrich et al., 2024; Waldendorf et al., 2024), final-layer gradients (Hu et al., 2024b; Snyder et al., 2024), and entropy of the generated tokens (Farquhar et al., 2024) to identify hallucinations. While grey-box models achieve higher detection accuracy than black-box models, they are computationally more demanding and cannot be applied to closed-source models due to restricted access to internal states. Conversely, black-box models, though computationally simpler, tend to perform less effectively (Deutsch et al., 2022). Additionally, we observe a significant lack of suitable and sufficiently large benchmark datasets spanning multiple domains to facilitate the evaluation and development of future RFHD methods (Sahoo et al., 2024a).

In this paper, we propose HalluCounter, a novel method that enhances response-consistency-based approaches by incorporating both response-response and query-response interactions. By leveraging consistency and alignment scores, HalluCounter learns a robust hallucination detection classifier. Response consistency-based approaches aim to detect hallucination in LLMs by generating multiple responses for the same input query and analyzing the variation in these responses (Manakul et al., 2023). Significant inconsistencies or contradictions across the generated responses signal potential hallucinations. Unlike prior methods, HalluCounter does not evaluate hallucination at the level of individual responses; rather, it assesses the self-consistency of an LLM when generating multiple responses to the same query. The core objective is to determine whether the LLM exhibits a ten-

[1]Code and data are publicly available for research purposes: https://github.com/rahulOm9/HalluCounter

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 352–383
December 20-24, 2025 ©2025 Association for Computational Linguistics

dency to hallucinate for a given query, rather than making a binary decision about a single response. Our model not only achieves higher detection accuracy compared to popular baselines but also provides a confidence score indicating how certain it is about its decision. Additionally, HalluCounter suggests the optimal response for users, regardless of whether the original generation contains hallucinations. Furthermore, we introduce a large-scale, multi-domain dataset for the RFHD task, comprising both synthetic and human-annotated samples. Unlike other existing datasets, this dataset poses significantly greater challenges for RFHD methods. It includes samples that demand domain knowledge across diverse fields, ranging from factual queries to those requiring reasoning and mathematical skills, which could be a good test bench for further RFHD explorations.

The key contributions of this work are: 1) We introduce HalluCounter[2], a novel approach for the RFHD task. 2) We present a large-scale, multi-domain benchmark dataset for RFHD, featuring both synthetic and human-annotated samples. 3) We conduct extensive experiments exploring various feature combinations, labeling strategies, classifiers, and LLMs across different sizes and families. 4) We perform a rigorous human evaluation of the model's selected optimal responses and carry out a thorough error analysis to uncover its potential limitations.

2 HalluCounterEval dataset creation

This section describes the creation of the HalluCounterEval dataset, which consists of various synthetic and human-annotated datasets for training and testing.

2.1 Raw data collection and processing

HalluCounterEval consists of two different training datasets. To create the first one, we obtain the raw data from an American television game show 'Jeopardy' (Jeopardy) and filter the dataset, which is highly diverse by including question-answer pairs related to six major domains and 22 sub-domains as detailed in Table 9. Moreover, the dataset includes factoid-based QA pairs, where many questions are not straightforward to answer. These questions often contain indirect hints, which increase their complexity and challenge the LLM's ability to handle ambiguity. The second dataset is the combination of multiple datasets obtained from Kaggle including Scientific QA (ScientificQA), MathQA (MathQA), Math QSA (MathQSA), and General Knowledge (GK) QA pairs as shown in Table 10. In the Kaggle dataset, scientific and GK questions test the LLMs' ability to extract factual knowledge. Whereas, MathQA and MathQSA questions assess the LLMs' logical reasoning and familiarization capabilities with mathematical notations. Both datasets undergo rule-based filtration steps as detailed in Appendix A to maintain the high quality. In accordance with Gebru et al. (2021)'s recommendation, we include a data sheet in Appendix L.

String matching	Qwen2.5-32B	Llama3-70B	GPT-4o
69.4%	89.4%	89.6%	89.8%

Table 1: Proportion of samples where the classification aligns with the human-annotated dataset.

2.2 Training dataset creation

The creation of training datasets consists of two stages 1) generation of sample responses, and 2) data labeling.

2.2.1 Sample responses generation

We utilize six different LLMs, including TinyLLaMA-1.1B (Zhang et al., 2024), Phi-3.5-B-mini (Abdin et al., 2024), Mistral-7B-instruct (Jiang et al., 2023), LLaMA-3-instruct 8B and 70B (Dubey et al., 2024), and Gemma-7B-instruct (Team et al., 2024) models to generate *'k'* responses[3] for each query by prompting each model *'k'* times. Due to limited compute, we use the 8-bit quantized version of the LLaMA-3-instruct-70B model for the inference, whereas other models are non-quantized versions. Further, as depicted in Appendix B Figure 3, we notice that TinyLLaMA-1.1B has the highest number of unique responses (lowest self-consistency) followed by Mistral-7B-instruct. All the prompts and corresponding inference configurations can be found in Appendix D.

2.2.2 Data Labeling

Data labeling aims to classify each LLM-generated sample response as either accurate (0) or hallucinated (1). The labeling can be achieved either through an LLM as a judge approach or a search-based string-matching method.

[2]We plan to make the code and dataset public.

[3]LLM generated 'responses' interchangeably refereed as 'sample responses'

(1) LLM as a judge. Prompt an LLM by providing the question, LLM response, and gold answer to classify whether the LLM response is accurate (0) or hallucinated (1).
(2) Exact-match. A search-based string-matching approach classifies an LLM's response as non-hallucinated if it matches the gold answer; otherwise, it is labeled as hallucinated.
Pilot study. To find the appropriate approach for the data labeling, we create a human-annotated dataset of 500 samples with the help of three expert annotators. To perform the annotation, we provide the question, gold answer, and LLM-generated response and ask the annotators to classify whether the LLM-generated response is hallucinated.
Selection of best labeling strategy. To find out the appropriate labeling strategy, we generate the labels by prompting GPT-4o mini (Achiam et al., 2023) (closed source), LLaMA3-70B and Qwen2.5-32B (Yang et al., 2024) (open source), and string-based matching methods and compare the percentage of labels match with the human-annotated dataset. As illustrated in Table 1, all three LLM-based labeling strategies perform similarly, with only minor variations when compared to human-annotated labels. However, we choose the Qwen2.5-32B for the entire training dataset labeling to reduce the compute requirements and encourage reproducibility by utilizing open-source models. The corresponding prompt for the labeling method is mentioned in Appendix E Table 12.

2.3 Test datasets creation

The HalluCounterEval dataset consists of 16 test datasets. Out of these, 14 are synthetically generated and two are human-annotated test sets. To create these test sets, we leverage both LLM and human annotation strategies.
Synthetic test sets. To create each test set, we follow the similar steps detailed for the training dataset creation (see Section 2.2). We obtain the test sets corresponding to Jeopardy and Kaggle datasets for TinyLLaMA-1.1B (*TL-1.1B-Gen*), Phi-3.5-B-mini (*PHI-3.5B-Gen*), Mistral-7B-instruct (*MST-7B-Gen*), LLaMA-3-instruct 8B (*LL-7B-Gen*) and 70B (*LL-70B-Gen*), Gemma-7B-instruct (*GM-7B-Gen*) and 'ensemble' (*ENSB-Gen*) models. The 'ensemble' test set consists of an equal number of samples assigned to different LLMs to generate the sample responses. In the rest of the paper, we report all the results on the test sets with corresponding acronyms of each LLM.

Human-annotated test set (HA-Test) is a manually curated dataset consisting of 1,956 samples or queries, with 956 sourced from Jeopardy and 1,000 from Kaggle datasets. For each query, we generate 10 responses, resulting in a total dataset size of 19,560 query-response pairs. Similar to the 'ensemble' test set, the HA-Test consists of LLM-generated responses from various LLMs. We classify the sample responses with the help of three expert annotators. Where, we provide a question, gold answer, and LLM response to the annotator and ask them to label it as either hallucinated (1) or non-hallucinated (0). We measure the Inter Annotator Agreement (IAA) between the annotators and obtain the Fleiss[4] kappa score of 0.83, which indicates an almost perfect agreement.

3 Methodology

3.1 Task formulation

We prompt a query Q to an LLM and collect *'k'* responses, denoted as $R = R_1, R_2, \ldots, R_k$, by querying the model *'k'* times with the same prompt. The query and its corresponding *'k'* responses are then processed by the proposed HalluCounter pipeline, which performs three key tasks: 1) determines whether LLM makes the hallucination for the given query, 2) provides a confidence score for the classifier's overall prediction, and 3) identifies the least hallucinated response among the *'k'* responses, referred as the optimal response.

3.2 HalluCounter Approach

The HalluCounter pipeline consists of three stages: 1) Extracting the NLI features, 2) Classification of the responses, and 3) Optimal response generation, and confidence score calculation. The following is a detailed description of each stage.

3.2.1 Extracting NLI features

We extract the NLI features between the Query-Response (Q-R) and Response-Response (R-R) pairs using the DeBERTa-v3-large (He et al., 2021) based cross-encoder model, fine-tuned on MNLI (Williams et al., 2018). We measure the NLI scores by concatenating the query with the LLM response or between the sample responses. The outputs from the NLI model are the logits associated with entailment, neutral, and contradiction.
Query-Response NLI features. To understand whether the generated response is relevant to the

[4] https://en.wikipedia.org/wiki/Fleiss%27_kappa

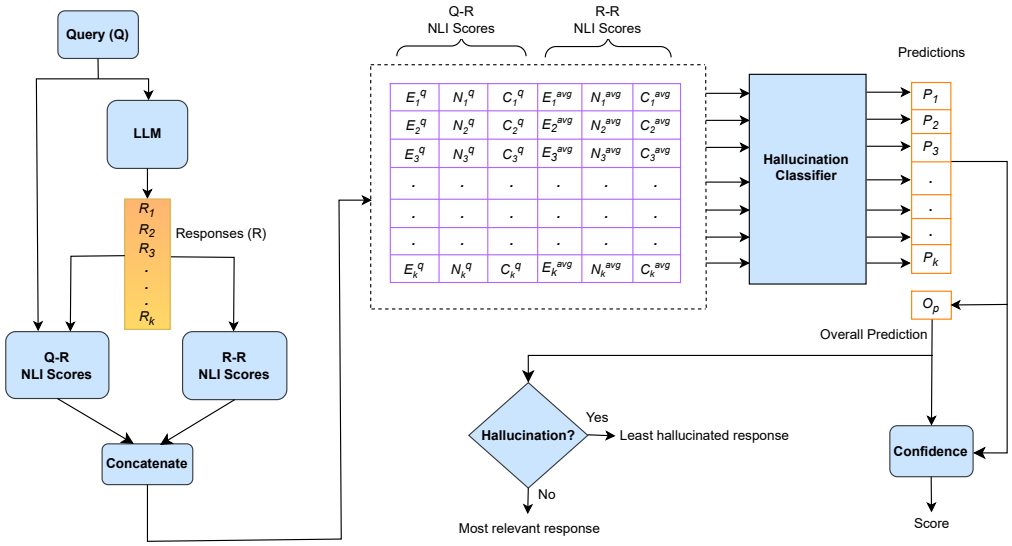

Figure 1: ***HalluCounter:*** A reference-free Hallucination Detection Pipeline for LLMs with three key components, 1) Extracting NLI features for query-response and response-response pairs, 2) A hallucination classifier that predicts hallucinations, and 3) Aggregating the final prediction, confidence score, and optimal response.

query or not, we obtain the NLI scores between the query and each response among all the *'k'* responses. As shown in Figure 1, the corresponding NLI scores indicated as: (E_i^q, N_i^q, C_i^q) for $i = 1, 2, \ldots, k$. We adopted the use of Q-R NLI scores following prior research (Fortier-Dubois and Rosati, 2023), which highlights the effectiveness of contradiction-based reasoning in improving QA models.

Response-Response NLI features. To verify the consistency among the sample responses, each response in the *R* is compared with other responses and obtains the corresponding NLI scores. We average the entailment, neutral, and contradiction features for each response. For a response R_i,

$$\text{Avg NLI scores} = \begin{cases} \mathrm{E}_i^{avg} = \frac{1}{k-1} \sum_{i=1, j \neq i}^{k} e_{ij} \\ \mathrm{N}_i^{avg} = \frac{1}{k-1} \sum_{i=1, j \neq i}^{k} n_{ij} \\ \mathrm{C}_i^{avg} = \frac{1}{k-1} \sum_{i=1, j \neq i}^{k} c_{ij} \end{cases} \tag{1}$$

Where e_{ij}, n_{ij}, c_{ij} are the entailment, neutral and contradiction scores between i^{th} and j^{th} responses.

3.2.2 Hallucination detection classifier

We build a classifier to classify whether the generated response contains hallucination or not. It takes the input as NLI feature values and generates binary output '1' for hallucination and '0' for non-hallucination. We built two different classifiers using statistical and BERT-based approaches.

Statistical Method. We utilize the ensemble of the Decision Trees, XGBoost, gradient-boosted Decision Trees (GBDT), and a voting classifier to design an ensemble classifier.

BERT classifier. We use the bert-base-uncased (Devlin et al., 2019) model to fine-tune the classifier by converting all the numerical features into textual features. Additional experimental details can be found in Appendix H. Furthermore, our pipeline yields the following three key outcomes.

1. Overall prediction: Let the k predictions be denoted as $p_1, p_2, \ldots, p_k$, where each $p_i \in \{0, 1\}$. We define the overall prediction $\hat{y}$ as:

$$\hat{y} = \begin{cases} 1 & \text{if } \sum_{i=1}^{k} p_i \geq \frac{k}{2} \\ 0 & \text{if } \sum_{i=1}^{k} p_i < \frac{k}{2} \end{cases} \tag{2}$$

2. Optimal response: We select the optimal response based on the overall prediction ($\hat{y}$) of the classifier. If the overall prediction is hallucinated, we choose all sample responses categorized as hallucination and among them pick the sample with the lowest contradiction score, whereas if the overall prediction is non-hallucinated, we select all the corresponding sample responses and among them pick the sample with the highest entailment score. This process ensures an optimal response to user queries. The optimal response R^* is selected as

Combination	Q-R			R-R			Text	
	E	C	N	E	C	N	Query (q)	Response (r)
C-C		✓			✓			
EC-EC	✓	✓		✓	✓			
Q-R	✓	✓	✓					
R-R				✓	✓	✓		
(Q-R)+(R-R)	✓	✓	✓	✓	✓	✓		
q-r+(Q-R)+(R-R)	✓	✓	✓	✓	✓	✓	✓	✓

Table 2: NLI features combinations; E, C, N denote Entailment, Contradiction, and Neutral features.

follows:

$$\mathrm{R}^* = \begin{cases} \arg\min_{R_i \in R} \left(\epsilon_1 \cdot \left(c_i^{\mathrm{q}}\right) + \epsilon_2 \cdot \left(c_i^{\mathrm{avg}}\right)\right) & \hat{y} = 1 \\ \arg\max_{R_i \in R} \left(\epsilon_1 \cdot \left(e_i^{\mathrm{q}}\right) + \epsilon_2 \cdot \left(e_i^{\mathrm{avg}}\right)\right) & \hat{y} = 0 \end{cases} \tag{3}$$

Where $R = [R_1, R_2, \ldots, R_k]$ represents the set of responses, ϵ_1 and ϵ_2 values indicate the weightage given to the Q-R and R-R feature values. After experimenting with various combinations of ϵ_1 and ϵ_2 values, we set ϵ_1 = 0.3 and ϵ_2 = 0.7.

3. Confidence score (CS): The confidence score is measured using all *'k'* responses predictions and the overall prediction. Let's take the *'k'* responses individual classifier predictions are $\{p_1, p_2 \ldots p_k\}$ and $\hat{y}$ is the overall prediction for the given query, then the confidence score is measured using Equation 4.

$$\mathrm{CS} = \begin{cases} \frac{1}{k} \sum_{i=1}^{k} p_i & \hat{y} = 1 \\ 1 - \frac{1}{k} \sum_{i=1}^{k} p_i & \hat{y} = 0 \end{cases} \tag{4}$$

4 Experiments and Results

This section presents the experimental results of the proposed pipeline and corresponding analysis. We report the F1-Score, AUC, and Balanced accuracy scores to evaluate the hallucination classifier performance.

4.1 NLI features combinations

We obtain various combinations of NLI features to train different classifiers. In total, we obtain eight features for a given query, out of them 6 are numerical features (three from each query-response (Q-R) and response-response (R-R) pairs NLI scores) and two are textual features ('query' & 'LLM response'). Using these features, we built several classifiers by combining them as shown in Table 2.

4.2 Jeopardy and Kaggle results analysis

We conduct experiments on Jeopardy and Kaggle datasets, by training various classifiers using statistical and BERT-based models on the 16 test sets. All the combinations of the experiments conducted are listed in Table 23. As shown in Table 18, for the Jeopardy dataset, the BERT classifier trained on a combination of numerical and textual features (q-r+Q-R-R-R) outperforms all other models, except for the HA-Test. Whereas on HA-test the model trained using a statistical classifier with EC-EC feature combination performs better than others. Additionally, as detailed in Table 20, we conduct experiments to evaluate the performance of the hallucination classifier across six sub-categories present in the Jeopardy dataset.

Similarly, we conduct experiments with the Kaggle test sets and listed the results in Table 8. Given the variations, such as mathematical formulations, present in the Kaggle test sets, we notice that the classifier trained on EC-EC feature combination performs comparably or even surpasses the 'q-r+Q-R-R-R' combination. Moreover, we report the results from all four datasets within the Kaggle dataset in Table 21. Appendix C presents the hallucination classifier results for all the combinations listed in Table 23 and Appendix J describes HalluCounter's performance on responses generated by GPT-4o (Hurst et al., 2024). We recommend using the 'q-r+Q-R-R-R' feature combination with a BERT classifier as a strong starting point when applying HalluCounter to new datasets. This combination has shown robust performance across multiple test sets, making it a reliable default choice.

4.3 Ablation study

Impact on the varying number of responses. We experiment with different numbers of sample responses (k = 3, 5, 7, 10) and notice the variations in the pipeline's prediction confidence values and hallucination rates. As detailed in Table 5, we find that as the number of sample responses increases, both the hallucination rate and the confidence of the hallucination classifier slightly decrease. However, despite changing the number of responses, our pipeline exhibits more than 90% confidence across different test sets, which indicates that the proposed pipeline is independent of the number of responses and the best results can be obtained with three sample responses as well. Moreover, as shown in Table 3 the pipeline exhibits stable per-

		TL-1.1B-Gen			PHI-3.5B-Gen			MST-7B-Gen			LL-8B-Gen			GM-7B-Gen			LL-70B-Gen			ENSB-Gen		
		3	5	10	3	5	10	3	5	10	3	5	10	3	5	10	3	5	10	3	5	10
Jeopardy	F1	0.75	0.75	0.75	0.71	0.71	0.71	0.68	0.68	0.68	0.82	0.82	0.81	0.63	0.63	0.62	0.54	0.54	0.54	0.74	0.74	0.73
	B-ACC	0.93	0.93	0.93	0.75	0.75	0.75	0.82	0.82	0.82	0.80	0.79	0.79	0.67	0.67	0.67	0.44	0.44	0.44	0.84	0.85	0.84
	ROC	0.74	0.74	0.75	0.78	0.78	0.79	0.75	0.76	0.76	0.89	0.88	0.88	0.70	0.69	0.70	0.60	0.60	0.60	0.83	0.83	0.83
Kaggle	F1	0.83	0.84	0.83	0.70	0.70	0.70	0.54	0.54	0.54	0.75	0.75	0.75	0.66	0.66	0.66	0.79	0.79	0.79	0.75	0.75	0.75
	B-ACC	0.92	0.93	0.93	0.63	0.61	0.60	0.65	0.65	0.65	0.63	0.64	0.65	0.72	0.72	0.72	0.70	0.70	0.68	0.80	0.79	0.80
	ROC	0.68	0.67	0.68	0.66	0.65	0.64	0.54	0.55	0.55	0.70	0.69	0.70	0.66	0.66	0.66	0.77	0.77	0.76	0.72	0.72	0.73

Table 3: HalluCounter performance with varying the number of sample responses.

Method	HaluEval Datasets		
	Summarization	QA	Dialogue
SelfCheckGPT	0.610	0.572	0.69
EigenScore	0.501	0.495	0.498
Perplexity	0.499	0.484	0.507
LN-Entropy	0.494	0.494	0.507
LexicalSimilarity	0.494	0.501	0.506
HalluCounter	**0.7**	**0.78**	**0.93**

Table 4: HalluCounter performance on HaluEval; all values represent accuracies.

formance across different 'k' values.

Performance on non-QA tasks. To verify the efficacy of HalluCounter on other than factoid QA datasets, we tested the HalluCounter on HaluEval (Li et al., 2023) dataset. Which consists of summarization, knowledge-grounded dialogue, and QA tasks. The HalluCounter performance on the HaluEval dataset are reported in Table 4.

4.4 Comparison with state-of-the-art

We compare our approach with two popularly known reference-free hallucination detection approaches in LLMs, which are SelfCheckGPT (Manakul et al., 2023) and InterrogateLLM (Yehuda et al., 2024), and uncertainty-based approaches, namely Perplexity (Ren et al.), Length Normalized entropy (Malinin and Gales, 2021), and Lexical similarity (Lin et al., 2022). Moreover, we also compared with two reference-based approaches HaloScope (Du et al., 2024), SAPLMA (Azaria and Mitchell, 2023) and a training-free approach Eigenscore (Chen et al., 2024). As detailed in Table 6, HalluCounter outperforms current state-of-the-art methods by a significant average margin of 10% with SelfCheckGPT and 21% with InterrogateLLM. Our study proves that consistency among only generated responses is insufficient to perform the RFHD task, the proposed approach outperforms state-of-the-art approaches by incorporating both response-response and query-response interactions. In contrast to existing works, our pipeline provides a confidence score and optimal response as well. Further details on the comparison study experimental setup can be found in Appendix F.

	Test set	Hallucination rate				Confidence score			
		K=3	K=5	K=7	K=10	K=3	K=5	K=7	K=10
Jeopardy	TL-1.1B-Gen	86	88	88	87	91	89	88	88
	PHI-3.5B-Gen	53	53	53	51	92	91	90	90
	LL-8B-Gen	29	28	28	26	94	93	93	93
	MST-7B-Gen	59	59	58	55	88	86	84	84
	GM-7B-Gen	38	37	37	36	95	94	93	93
	LL-70B-Gen	17	17	17	17	100	100	100	100
	ENSB-Gen	53	53	53	51	91	90	89	88
	HA-Test	53	53	54	52	87	84	83	82
Kaggle	TL-1.1B-Gen	87	87	87	86	96	95	95	95
	PHI-3.5B-Gen	67	67	67	66	96	95	95	95
	LL-8B-Gen	63	63	64	62	93	92	92	92
	MST-7B-Gen	76	76	76	75	95	94	93	93
	GM-7B-Gen	73	73	73	72	95	94	93	93
	LL-70B-Gen*	68	67	67	66	95	94	93	93
	ENSB-Gen	53	53	53	51	91	90	89	88
	HA-Test	65	67	68	66	88	85	84	84

Table 5: HalluCounter pipeline results by varying number of sample responses (*'K'*); The results of best-performing model for each test is reported. * denotes the quantized version. All the values are in percentages.

4.5 Human evaluation

We conduct a human evaluation on 500 samples each from the Jeopardy and Kaggle datasets to assess whether the pipeline-selected response is optimal. These samples are taken from the Human annotated test set. For this analysis, we choose the optimal responses from the *'k'* sample responses for each query. We instruct the expert evaluators to indicate whether they agree or disagree with the pipeline-selected optimal response, based on the classification label (hallucinated or non-hallucinated). In the HA-test, for the Jeopardy dataset, we achieve 82.4% agreement, whereas for the Kaggle dataset, the agreement is 84%. Moreover, on the LL-70B-Gen test set, we obtain 75.8%, and 86% scores for Jeopardy and Kaggle datasets.

		Jeopardy			Kaggle		
Type of method	**Approach**	**F1**	**B-ACC**	**AUC**	**F1**	**B-ACC**	**AUC**
Response-consistency	SelfCheckGPT	0.651	0.687	0.811	0.674	0.674	0.755
	InterrogateLLM	0.427	0.697	0.813	0.671	0.671	**0.899**
Uncertainty-based	Perplexity	0.487	0.504	0.505	0.678	0.678	0.606
	LN-Entropy	0.441	0.499	0.70	0.707	0.707	0.567
	LexicalSimilarity	0.442	0.498	0.498	0.711	0.711	0.558
Training-based	HaloScope	0.323	0.524	0.584	0.402	0.402	0.600
	SAPLMA	0.668	0.679	0.725	0.716	0.716	0.773
Training-free	EigenScore	0.437	0.496	0.496	0.658	0.658	0.501
Ours	**HalluCounter**	**0.743**	**0.83**	**0.82**	**0.782**	**0.782**	0.76

Table 6: Comparison with state-of-the-art approaches.

4.6 Error analysis

We perform the error analysis to understand the effectiveness of the proposed HalluCounter approach. We manually verify 500 samples each from HA-Test and LL-70B-Gen. Each category error analysis details are outlined in Table 7. The following are the major error categories, where the proposed pipeline might exhibit sub-standard performance.

1. Misclassification. The HalluCounter pipeline makes incorrect predictions, due to *a). Complete inconsistency* among the sample responses, which is against the core principle of the design of the HalluCounter approach. *b). Partial inconsistency.* The number of incorrect responses is greater than correct responses in total sample responses, *c). Pipeline inefficiency.* The HalluCounter pipeline might fail due to the inefficacy of one or more components including measuring NLI scores, classifier prediction, or optimal response selection .

2. Answer denial. *a). Insufficient context.* LLMs refuse to answer the query either due to insufficient context or ambiguous information present in the query. *b). Problematic context.* Presence of misleading, violent, or contradictory information in the query. The corresponding examples for all the error categories are illustrated in Appendix I Table 14.

5 Discussion and Insights

Performance across various domains. As shown in Table 5, all LLMs exhibit a higher tendency to hallucinate on the Kaggle test sets compared to the Jeopardy test sets. Specifically, Figure 2 reveals that LLMs experience the highest hallucination rates on questions related to "MathQA", "arts and humanity", followed by "language and communication", with the lowest rates occurring in the "GK" and "Geography and travel" categories. It is evident from our study that, the majority of LLMs face significant challenges with queries demanding mathematical reasoning (Srivatsa and Kochmar, 2024; Ahn et al., 2024) and scientific factual knowledge (Yang and Zhao, 2024).

High resiliency. The confidence score in HalluCounter reflects the level of resiliency in determining whether a response is hallucinated. As presented in Table 5, despite the slight variations in the hallucination rates with varying numbers of sample responses, the proposed pipeline consistently achieves an average confidence score above 90% across both the Jeopardy and Kaggle test sets. From this result, it is evident that the performance of the HalluCounter pipeline remains largely unchanged regardless of the number of sample responses.

LLMs hallucination rate. To assess which LLMs are highly prone to hallucination, we compare overall prediction with the actual label. As shown in Table 5, we find that for the Jeopardy dataset TinyLLaMA-1.1B and Mistral-7B models are more likely to generate hallucinated responses, and LLaMA-3-70B produces the least percentage of hallucinations. Whereas in the case of Kaggle datasets TinyLLaMA-1.1B, Mistral-7B, and Gemma-7B models are prone to higher hallucination. The models that failed on the Jeopardy dataset lack logical reasoning capabilities because most of the Jeopardy dataset consists of hint-based general knowledge questions.

NLI model robustness. We notice that often the NLI model assigns high scores to longer LLM response sequences and unseen premise-hypothesis pairs (Yang, 2024), which leads to high entailment and contradiction scores. In such cases, the classifier might exhibit mediocre performance.

Assessing the ambiguity. Since most of the Jeop-

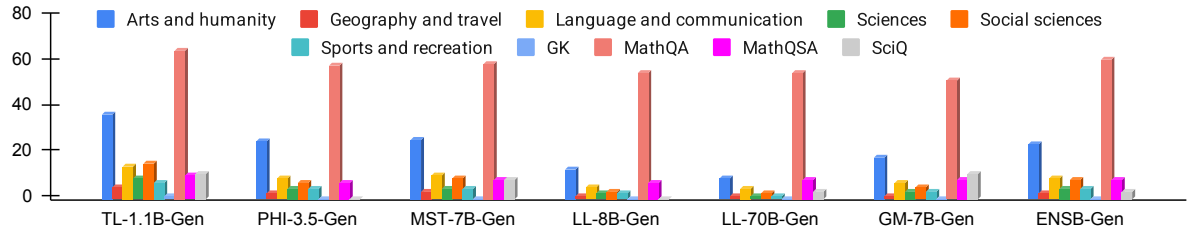

Figure 2: Hallucination rates across different sub-domains in various test sets of the Jeopardy and Kaggle datasets.

ardy dataset questions are hint-based, there is a possibility of providing a biased answer to an ambiguous question that could have multiple correct answers (Park and Kim, 2025). In such cases, the HalluCounter pipeline might struggle to classify it as either accurate or hallucinated. Similarly, in a few instances, the labeling model Qwen2.5-32B fails to perform accurate semantic matching.

6 Background on Hallucination detection

Hallucinations in LLMs remain an enduring challenge across text, image, audio, and video (Sahoo et al., 2024b; Li et al., 2024), and detecting them is crucial, especially when no external reference or ground truth is available.

Self-consistency approaches gained a lot of attention in detecting the factual correctness in the LLM-generated responses. Approaches such as SelfCheckGPT (Manakul et al., 2023), which relies on the principle of self-consistency among the stochastically generated responses and detects the hallucination based on whether the generated responses support the original answer. SAC^3 (Zhang et al., 2023) detect hallucination by analyzing cross-model consistency and cross-rephrased queries. InterrogateLLM (Yehuda et al., 2024), detects hallucination by asking the reverse question and verifies whether the original question can be generated.

		Misclassification			Answer Denial	
		C1	C2	C3	C4	C5
LL-70B-Gen	Jeopardy	21.4	0	2	0	0
	Kaggle	5.2	6.2	2.8	0	0
HA-Test	Jeopardy	8.4	3.2	2.6	1.4	1
	Kaggle	11.4	0.8	3.6	3.8	0

Table 7: Error analysis of 500 samples for the following error categories, C1) Complete inconsistency, C2) Partial inconsistency, C3) Pipeline inefficiency, C4) Insufficient context, C5) Problematic context; Each value represents percentages of error instances.

LogicCheckGPT (Wu et al., 2024), asks LLMs questions with logical correlations to detect hallucination. SELF-FAMILIARITY (Luo et al., 2024) focuses on evaluating the model's familiarity with the concepts present in the instruction.

Several approaches leverage LLM's internal representations to detect hallucination, by training a classifier using the LLM's hidden representations (Azaria and Mitchell, 2023), weighting LLMs' expertise (Wei et al., 2024), by calculating the probability of each token in the given text (Liu et al., 2022), measuring the semantic consistency across various generations in embedding space (Chen et al., 2024). Additionally, uncertainty-based estimation approaches based on aleatoric and epistemic uncertainty have been studied to detect hallucination in auto-regressive generation (Xiao and Wang, 2021; Malinin and Gales, 2021). However, these approaches are limited to white-box models.

We draw inspiration from the SelfCheckGPT, which uses the normalized scores of entailment and contradiction NLI scores between the responses to detect the hallucinations. In contrast, our approach leverages query-response and response-response consistency and alignment patterns to train a hallucination detection classifier. Additionally, unlike existing methods, our pipeline provides the least hallucinated response among all the responses along with overall prediction and the corresponding confidence score.

7 Conclusion

In this work, we propose HalluCounter, a novel method for RFHD in LLMs. This method improves response consistency-based hallucination detection methods and generates confidence scores and optimal responses along with hallucination detection. We introduce a large-scale HalluCounterEval dataset, which consists of a large set of synthetic and human-annotated samples across diverse domains. Through extensive experiments and ab-

		QR			RR			EC-EC			CC			QR-RR			q-r+Q-R+R-R		
Test Data	**Classifier**	**F1**	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**
TL-1.1B-Gen	Statistical	0.71	0.60	0.88	0.80	0.61	0.92	0.82	0.68	0.93	0.73	0.62	0.90	0.83	0.68	0.93	-	-	-
	BERT	0.82	0.60	0.88	0.63	0.61	0.88	0.85	0.70	0.94	0.74	0.64	0.91	0.85	0.70	0.94	**0.86**	0.76	0.94
PHI-3.5B-Gen	Statistical	0.58	0.50	0.49	0.66	0.63	0.60	0.68	0.62	0.59	0.61	0.54	0.51	0.70	0.64	0.60	-	-	-
	BERT	0.68	0.65	0.62	0.66	0.52	0.50	0.70	0.65	0.61	0.66	0.55	0.51	0.71	0.65	0.62	**0.77**	0.71	0.65
LL-8B-Gen	Statistical	0.56	0.53	0.51	0.73	0.69	0.64	0.75	0.70	0.65	0.63	0.60	0.56	0.75	0.70	0.65	-	-	-
	BERT	0.76	0.72	0.66	0.65	0.56	0.52	**0.77**	0.73	0.67	0.72	0.65	0.61	**0.77**	0.72	0.66	**0.77**	0.75	0.69
MST-7B-Gen	Statistical	0.53	0.53	0.66	0.56	0.49	0.64	0.54	0.47	0.62	0.54	0.55	0.66	0.54	0.55	0.65	-	-	-
	BERT	0.55	0.52	0.64	0.53	0.51	0.64	0.53	0.54	0.65	0.53	0.45	0.61	0.53	0.51	0.63	0.56	0.68	0.74
GM-7B-Gen	Statistical	0.60	0.53	0.62	**0.67**	0.67	0.72	0.66	0.66	0.71	0.63	0.60	0.67	**0.67**	0.67	0.72	-	-	-
	BERT	**0.67**	0.68	0.73	0.68	0.55	0.63	0.64	0.68	0.73	**0.67**	0.62	0.68	0.66	0.67	0.71	0.65	0.70	0.75
LL-70B-Gen	Statistical	0.55	0.49	0.48	0.79	0.77	0.71	0.80	0.78	0.72	0.61	0.60	0.58	0.79	0.76	0.68	-	-	-
	BERT	0.83	0.80	0.73	0.65	0.55	0.52	**0.84**	0.81	0.74	0.72	0.65	0.60	0.82	0.80	0.72	0.80	0.80	0.73
ENSB-Gen	Statistical	0.60	0.53	0.65	0.73	0.72	0.80	0.76	0.72	0.80	0.66	0.60	0.72	0.75	0.73	0.81	-	-	-
	BERT	0.77	0.74	0.78	0.64	0.54	0.65	0.78	0.76	0.82	0.69	0.63	0.73	0.79	0.75	0.82	**0.80**	0.83	0.86
HA-Test	Statistical	0.65	0.51	0.70	0.76	0.66	0.82	**0.78**	0.69	0.82	0.70	0.59	0.77	0.77	0.70	0.82	-	-	-
	BERT	0.23	0.50	0.68	0.59	0.50	0.68	0.23	0.50	0.68	0.59	0.50	0.68	0.23	0.50	0.68	0.68	0.76	0.81

Table 8: Hallucination classifier results on various test sets from Kaggle dataset, **AUC:** Area Under Curve, **B-ACC:** Balanced Accuracy. All the values are the average scores of four Kaggle datasets, with the best result in **bold**.

lations, we evaluate various NLI feature combinations, classifiers, and labeling strategies. Additionally, we offer a detailed error analysis, key insights, and takeaways from our method and benchmark dataset.

8 Limitations

This paper proposes a novel reference-free hallucination detection pipeline, despite the best efforts, our paper still has several limitations. (1) Synthetic datasets creation: To create synthetic train and test sets, we experiment with zero-shot prompting only, and to increase the quality of the datasets further studies can experiment with few-shot and Chain-of-thought prompting strategies as well. (2) Cross-encoder module sensitivity towards longer sequences: The classifier heavily relies on the cross-encoder module to obtain NLI logit values, however the cross-encode module is prone to provide high entailment values for longer sequences, which might lead to inaccurate classifier prediction. (3) Inconsistency among sample responses: Our approach works on the principle of self-consistency among the sample responses, we face challenges if all the responses are hallucinated in that case our approach may exhibit mediocre performance. (4) Computational complexity: Despite HalluCounter's superior performance compared to state-of-the-art approaches, it is quite computationally heavy, which could be addressed in future work to be made more efficient.

9 Ethics Statement

In this work, we utilize only the publicly available datasets. We make all the synthetic and human-annotated datasets public to encourage reproducibility. Moreover, by tackling the issue of hallucinations in LLMs, this work points out that undetected hallucinations could lead to misinformation.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237, St. Julian's, Malta. Association for Computational Linguistics.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xuefeng Du, Chaowei Xiao, and Sharon Li. 2024. Haloscope: Harnessing unlabeled llm generations for hallucination detection. *Advances in Neural Information Processing Systems*, 37:102948–102972.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Etienne Fortier-Dubois and Domenic Rosati. 2023. Using contradictions improves question answering systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 827–840, Toronto, Canada. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12).

GK. https://www.kaggle.com/datasets/ilyaryabov/general-knowledge-qa. Online; accessed 1-November-2024.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024a. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models. *arXiv preprint arXiv:2405.14486*.

Xiaomeng Hu, Yiming Zhang, Ru Peng, Haozhe Zhang, Chenwei Wu, Gang Chen, and Junbo Zhao. 2024b. Embedding and gradient say wrong: A white-box method for hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1959.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jeopardy. https://www.reddit.com/r/datasets/comments/1uyd0t/200000_jeopardy_questions_in_a_json_file/?rdt=35719. Online; accessed 1-December-2024.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Qing Li, Jiahui Geng, Chenyang Lyu, Derui Zhu, Maxim Panov, and Fakhri Karray. 2024. Reference-free hallucination detection for large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4542–4551, Miami, Florida, USA. Association for Computational Linguistics.

Zi Lin, Jeremiah Zhe Liu, and Jingbo Shang. 2022. Towards collaborative neural-symbolic graph semantic parsing via uncertainty. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4160–4173, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

Xin Liu, Muhammad Khalifa, and Lu Wang. 2024. Litcab: Lightweight language model calibration over short-and long-form responses. In *The Twelfth International Conference on Learning Representations*.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2024. Zero-resource hallucination prevention for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3586–3602, Miami, Florida, USA. Association for Computational Linguistics.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

MathQA. https://www.kaggle.com/datasets/thedevastator/dataset-for-solving-math-word-problems. Online; accessed 1-November-2024.

MathQSA. https://www.kaggle.com/datasets/awsaf49/math-qsa-dataset. Online; accessed 1-November-2024.

Hancheol Park and Geonmin Kim. 2025. Where do LLMs encode the knowledge to assess the ambiguity? In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 445–452, Abu Dhabi, UAE. Association for Computational Linguistics.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024a. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024b. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.

ScientificQA. Sciq: A dataset for science question answering. https://www.kaggle.com/datasets/thedevastator/sciq-a-dataset-for-science-question-\answering. Kaggle; accessed 1 November 2024.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 21–33.

Ben Snyder, Marius Moisescu, and Muhammad Bilal Zafar. 2024. On early detection of hallucinations in factual question answering. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2721–2732.

Kv Aditya Srivatsa and Ekaterina Kochmar. 2024. What makes math word problems challenging for LLMs? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1138–1148, Mexico City, Mexico. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Jonas Waldendorf, Barry Haddow, and Alexandra Birch. 2024. Contrastive decoding reduces hallucinations in large multilingual machine translation models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2539.

Jiaheng Wei, Yuanshun Yao, Jean-Francois Ton, Hongyi Guo, Andrew Estornell, and Yang Liu. 2024. Measuring and reducing llm hallucination without gold-standard answers via expertise-weighting. *arXiv preprint arXiv:2402.10412*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. 2024. Logical closed loop: Uncovering object hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6944–6962, Bangkok, Thailand. Association for Computational Linguistics.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Dongjie Yang and Hai Zhao. 2024. Are LLMs aware that some questions are not open-ended? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2142–2152, Miami, Florida, USA. Association for Computational Linguistics.

Zijiang Yang. 2024. Improving the natural language inference robustness to hard dataset by data augmentation and preprocessing. *arXiv preprint arXiv:2412.07108*.

Yakir Yehuda, Itzik Malkiel, Oren Barkan, Jonathan Weill, Royi Ronen, and Noam Koenigstein. 2024. InterrogateLLM: Zero-resource hallucination detection in LLM-generated answers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9333–9347, Bangkok, Thailand. Association for Computational Linguistics.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023. SAC^3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15445–15458, Singapore. Association for Computational Linguistics.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

A HalluCounterEval dataset filtration details

The datasets present in the HaluCounterEval (Jeopardy and Kaggle) undergo rule-based filtering stages to ensure quality and consistency before being split into training and test sets. The following filtration steps are common to all the training and testing datasets.

- Initial Dataset: The raw dataset consists of question-answer pairs collected from their respective sources.
- Removal of URLs: Questions containing URLs in the text are filtered out.
- Exclusion of "Fill-in-the-Blank" Questions: Questions with dashes (representing blanks) are excluded from the dataset.
- Elimination of Short Questions: Questions with fewer than five words are removed to maintain sufficient context.

B Train and test dataset details

The training and testing dataset statistics of Jeopardy and Kaggle are detailed in Table 9 and 10. All the values are in Table 9 and 10 corresponding to total number of unique queries. We generate 10 samples per each query and obtain 10 times of the total unique samples for the purposes of training and testing. Moreover, the jeopardy dataset comprises of 6 major categories and 22 sub-categories of various domains of data. Whereas, the Kaggle dataset consists of four different datasets including scientific, general knowledge, and mathematical domain factoid question-answer pairs. Further,

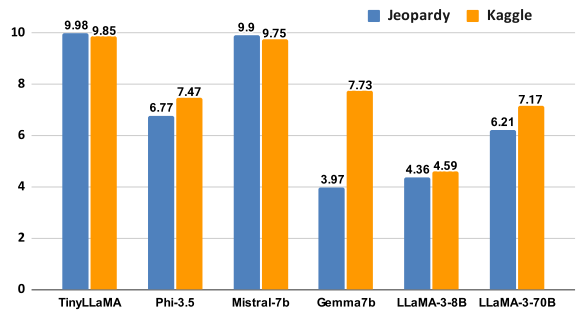

Figure 3: Number of unique responses generated by each LLM out of 10 responses for Jeopardy and Kaggle datasets. The lower the number represents the higher the consistency.

Main category	Sub-category	Train	Test
Arts and Humanities	Authors	843	94
	Books	997	111
	Culture	300	33
	Literature	1370	152
	Movies	1426	159
	Music	2581	287
	TV	2272	253
Geography and travel	Geography	1245	138
	Rivers	320	35
	Travel	535	60
Language and communication	Language	526	58
	Words	3424	380
Sciences	Animals	550	61
	Physics	189	21
	Science	1819	202
Social sciences	Education	137	15
	History	3245	361
	Law	233	26
	Politics	259	29
	Presidents	547	61
Sports and recreation	Awards	335	37
	Sports	1512	168
	Total	**24665**	**2741**

Table 9: Jeopardy dataset statistics.

	MathQA	MathQSA	SciQ	GK	**Total**
Train	32980	4956	12102	657	**50695**
Test	3665	550	1345	73	**5633**

Table 10: Kaggle dataset statistics.

as shown in Figure 3, TinyLLaMA-1.1B has the highest number of unique responses followed by Mistral-7B model.

C More results for the Hallucination classifier

We perform a series of experiments across multiple test sets, using different classifiers and labeling strategies for both the Jeopardy and Kaggle datasets.

C.1 Results on Jeopardy dataset

We built various classifiers to detect hallucination in LLMs. For all the best-performing models, hallucination classifier results are detailed in Table 18. Moreover, we report the results of the statistical approach-based hallucination classifier trained on the Jeopardy dataset with labels obtained from the exact-match approach in Table 24, LLM-based approach in Table 28. Similarly, the BERT classifier is trained on the Jeopardy dataset with labels obtained from the exact-match approach in Table 28, LLM-based approach in Table 30. Additionally, we report each category-wise result for the Jeopardy dataset in Table 20.

Model	Source
TinyLlama-1.1B	https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0
Gemma-7B	https://huggingface.co/google/gemma-7b-it
Mistral-7B	https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1
Phi-3.5B	https://huggingface.co/microsoft/Phi-3.5-mini-instruct
Llama-8B	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
Llama-70B	https://huggingface.co/Groq/Llama-3-Groq-70B-Tool-Use
Qwen-32B	https://huggingface.co/Qwen/Qwen2.5-32B-Instruct

Table 11: Source of Huggingface models.

C.2 Results on Kaggle dataset

We report the results of the statistical approach-based hallucination classifier trained on the Kaggle dataset with labels obtained from the exact-match approach in Table 25, LLM-based approach in Table 29. Similarly, the BERT classifier is trained on the Kaggle dataset with labels obtained from the exact-match approach in Table 27, LLM-based approach in Table 31. Additionally, we test the efficiency of the classifier on four different Kaggle datasets, and the corresponding results are mentioned in Table 21.

C.3 Experiments with additional features

We additionally include two token-based features for training the classifier: the length of the LLM-generated response and the number of punctuation marks it contains. Incorporating these features alongside the NLI-based features yields a modest improvement in overall classifier accuracy. Experimental results on the HA-Test dataset are presented in Table 19.

D Sample responses generation

As mentioned in Table 22, we use the same prompt *'k'* times to generate *'k'* responses each time to avoid the mismatch in the total number of sample responses for each query. We did the inference with various LLMs by using the same prompt. While generating the data for training, we set the *'k'* value to 10.

Role	Content
System	You are Qwen, created by Alibaba Cloud. You are a helpful assistant.
User	You are a helpful assistant tasked with evaluating whether a model-generated response is hallucinated or not. Here is the context: Question: `{question}` Correct Answer: `{gold_answer}` Model Response: `{llm_response}` Your task is as follows: 1. Check if the correct answer or its meaningful variations (e.g., initials, abbreviations, synonyms) appear in the model response. 2. If the correct answer (or a variation) is present, even partially, and the essence of correctness is captured, label it as '0' (not hallucinated). 3. If the correct answer or meaningful variations are completely absent or contradicted, label it as '1' (hallucinated). 4. Provide only the label (1 or 0) as your output. Do not include any additional information.

Table 12: Prompt for classifying whether LLM generated response is hallucinated or not

Dataset	Category	ENSB-Gen	GM-7B-Gen	LL-70B-Gen	LL-8B-Gen	MST-7B-Gen	PHI-3.5B-Gen	TL-1.1B-Gen
Jeopardy	Arts and humanity	24	18	9	13	26	25	37
	Geography and travel	2	1	1	1	3	2	5
	Language and communication	9	7	4	5	10	9	14
	Sciences	4	3	1	2	4	4	9
	Social sciences	8	5	2	3	9	7	15
	Sports and recreation	4	3	1	2	4	4	7
Kaggle	GK	0	0	0	0	0	0	1
	MathQA	61	52	55	55	59	58	65
	MathQSA	8	8	8	7	8	7	10
	SciQ	3	11	3	0	8	0	11

Table 13: Hallucination rate for each category in Jeopardy and Kaggle datasets across various test sets generated by LLMs; all the values are in percentages.

D.1 LLM inference configuration details

We did the inference with various small and large language models. Across all the models we use the max_new_tokens=32, top_k=50, top_p=0.95, and temperature=1. Additionally, we did the necessary response parsing to obtain only the relevant information related to the given query.

E Labeling using Qwen2.5-32B Model

We perform the labeling using the Qwen2.5-32B (Yang et al., 2024) to classify whether each LLM response is hallucinated or non-hallucinated. We used the prompt mentioned in Table 12 to perform the labeling.

F Comparison experiments details

We compare our approach with two popularly known reference-free hallucination detection approaches, which are SelfCheckGPT (Manakul et al., 2023) and InterrogateLLM (Yehuda et al., 2024).

SelfCheckGPT. To compare with the SelfCheckGPT approach, we utilize the prompt variant approach, where by providing the context, sentence and instruct the Qwen2.5-32B (Yang et al., 2024) LLM to whether the sentence is supported by the context or not. The final inconsistency score is computed by averaging the sentence scores.

InterrogateLLM. To compare with the InterrogateLLM approach, first, we create a few-shot prompt with question and answer pairs. In the forward pass, we generate an answer to each question and in the back-ward pass obtain the 10 questions to the same answer by modifying the few-shot prompt. In the end, by measuring the average cosine similarity between the original question and generated questions, we classify the question with more than 0.91 threshold as non-hallucinated. In the forward and backward process, we utilize the LLaMA3-8B model for inference.

G Generalization experiments

To verify the generalizability of the HalluCounter approach, we train the HalluCounter on Jeopardy, test on Kaggle, and perform the vice-versa experiments, and the corresponding results are detailed in Table 16.

G.1 Hallucounter performance with varying number of sample responses

We conduct experiments to analyze the performance of HalluCounter while varying the number of sample responses obtained from the LLM and

Error category	Examples
Complete inconsistency	**Question:** A record from years ago that's still worth listening to is an oldie but this **Correct answer:** Goodie **Sample responses:** ['Goldfinger.', 'Goldfinger', 'Goldfinger.', 'Gold.', 'gold.', 'goldmine.', 'Goldfinger.', 'Goldfish Crackers', 'gold.', 'Goldfinger.'] **Optimal response:** Goldfinger **Classification:** Non-hallcuinated
Partial inconsistency	**Question:** Type of machine you can use to send & receive letters or photos over the telephone line **Correct answer:** A fax **Sample responses:** ['Modem.', 'Modem.', 'Modem (Modulator-Demodulator).', 'Modem.', 'Modem.', 'Modem.', 'Facsimile machine or- Fax machine.', 'Fax machine.', 'Modem.', 'Modem.'] **Optimal response:** Modem. **Classification:** Non-hallucinated
Pipeline failure	**Question:** What is the innermost layer of the sun called? **Correct answer:** The core **Sample responses:** ['The core.', 'Core.', 'Core.', 'The core.', 'Core.', 'Core.', 'The inner most layer of the sun is called the core.', 'The core.', 'The inner most layer of the sun is called the Core.', 'The core.'] **Optimal response:** The core **Classification:** Hallucinated
Insufficient context	**Question:** Could go in an "nct" category. Period **Correct answer:** succinct **Sample responses:** ['The answer is: Not applicable to the context.', 'Okay.', 'The category you requested is not relevant to the given question, so I am unable to provide an answer.', 'Alright.', 'Certainly.', 'Alright.', 'Okay.', 'Insufficient context provided for a response.', 'Insufficient information provided to categorize.', 'Okay.'] **Optimal response:** The answer is: Not applicable to the context. **Classification:** Hallucinated
Problematic context	**Question:** "I didn't kill my wife!" **Correct answer:** The Fugitive **Sample responses:** ['I am unable to provide responses that suggest innocence or guilt in criminal matters.', 'I am unable to provide responses that suggest or deny involvement in criminal activities.', 'I am unable to provide responses that suggest or imply criminal actions or deception.', 'I am unable to provide responses that may contain misleading or potentially incriminating information.', 'I am unable to provide responses that may contain misleading or contradictory information.', 'I am unable to provide responses that suggest innocence in criminal activities or potentially misleading information.', 'I am unable to provide responses that may be misleading or potentially conceal criminal activity.', 'I am unable to provide responses that suggest or imply criminal activity or harmful actions.', 'I am unable to provide responses that may provide misleading or potentially incriminating information.', 'I am unable to provide subjective information or opinions, including personal claims of innocence.'] **Optimal response:** I am unable to provide responses that may contain misleading or contradictory information. **Classification:** Hallucinated

Table 14: Examples for different error categories.

	HA-Test	
	Jeopardy	Kaggle
Hallucination rate	13.5	22.6
Confidence score	90	83

Table 15: HalluCounter performance on GPT-4o-mini generated sample responses; all the values are in percentages.

Trained	Tested	F1-Score	B-ACC	AUC
Jeopardy	Jeopardy	0.73	0.86	0.82
	Kaggle	0.77	0.61	0.80
Kaggle	Kaggle	0.66	0.82	0.76
	Jeopardy	0.68	0.82	0.79

Table 16: Cross comparison experiments results.

the corresponding results are outlined in Table 3. From the results, it is evident that despite varying the K values, there is no significant variation in the accuracies across various tests for both the Jeopardy and Kaggle datasets. This indicates that our proposed HalluCounter pipeline is stable across different K values.

H Experimental setup

We conduct all experiments using two Nvidia GeForce RTX A6000 (48GB) GPUs. We do not perform the hyperparameter search. The maximum sequence length for classifier training with various feature combinations is set to 200, except for the 'q-r+Q-R-R-R', where it is set to 512. All other configurations follow the default settings of the Hugging Face trainer[5]. The huggingface models used in the experiments along with their sources are detailed in Table 11.

H.1 Conversion of numerical to textual features

The following template is used to convert the numerical features into textual features for training the classifier. The template takes into account the question, response, and several scores related to query-response and response-response entailment, neutrality, and contradiction.

- **Question**: The given question is the text input represented by `Question`.
- **Response**: The given response from the model is represented by `Response`.
- **Query-Response Entailment Score**: The numerical score indicating the entailment score obtained between the query to response, represented by `feature_1`.
- **Query-Response Neutral Score**: The numerical score representing the neutral score obtained between the query to response, represented by `feature_2`.
- **Query-Response Contradiction Score**: The numerical score representing the contradiction score obtained between the query to response, represented by `feature_3`.
- **Response-Response Entailment Score**: The numerical score indicating the entailment score obtained between the response to response, represented by `feature_4`.
- **Response-Response Neutral Score**: The numerical score representing the neutral score obtained between the response to response, represented by `feature_5`.
- **Response-Response Contradiction Score**: The numerical score representing the contradiction score obtained between the response to response, represented by `feature_6`.

This conversion process generates a structured textual feature that combines the question, response, and scores in the following format:

> "The given question is `{Question}` and the corresponding answer is `{Response}`, and they got the query-response entailment score: `{feature_1}`, neutral score: `{feature_2}`, and contradiction score: `{feature_3}`. And they got the response-response entailment score: `{feature_4}`, neutral score: `{feature_5}`, contradiction score: `{feature_6}`."

This textual feature is used as input for the classifier.

I Error analysis examples

We observe various error cases, where our HalluCounter pipeline fails to do the accurate classification and optimal response selection. The corresponding examples are detailed in Table 14.

[5] https://huggingface.co/docs/transformers/main_classes/trainer

Dataset	Category	ENSB-Gen	GM-7B-Gen	LL-70B-Gen	LL-8B-Gen	MST-7B-Gen	PHI-3.5B-Gen	TL-1.1B-Gen
Jeopardy	Arts and humanity	88	92	100	92	85	90	90
	Geography and travel	88	96	100	95	81	90	79
	Language and communication	87	92	100	92	85	89	91
	Sciences	89	94	100	93	83	89	84
	Social sciences	89	94	100	94	81	90	85
	Sports and recreation	89	94	100	92	83	90	89
Kaggle	GK	93	98	93	96	89	96	85
	MathQA	96	92	94	90	96	95	100
	MathQSA	93	96	92	89	91	87	98
	SciQ	91	93	89	96	87	97	80

Table 17: Confidence Score for each category in Jeopardy and Kaggle datasets across various test sets generated by LLMs. All the values are in percentages.

			QR			RR			EC-EC			CC			QR-RR			q-r+Q-R+R-R		
Test Data	**Classifier**	**Labeling**	**F1**	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**
TL-1.1B-Gen	Statistical	Exact-match	0.68	0.58	0.90	0.74	0.57	0.90	0.76	0.64	0.92	0.69	0.59	0.91	0.76	0.63	0.91	-	-	-
		LLM-based	0.65	0.67	0.90	0.73	0.68	0.90	0.75	0.75	0.93	0.64	0.62	0.89	0.75	0.75	0.93	-	-	-
	BERT	Exact-match	0.81	0.59	0.90	0.64	0.62	0.92	0.82	0.67	0.92	0.70	0.61	0.91	0.81	0.66	0.92	**0.88**	0.90	0.99
		LLM-based	0.77	0.70	0.91	0.55	0.67	0.91	0.79	0.78	0.94	0.53	0.61	0.89	0.78	0.78	0.94	0.82	0.84	0.96
PHI-3.5B-Gen	Statistical	Exact-match	0.55	0.59	0.69	0.63	0.69	0.77	0.65	0.71	0.79	0.58	0.61	0.72	0.65	0.71	0.79	-	-	-
		LLM-based	0.53	0.58	0.56	0.69	0.77	0.74	0.71	0.79	0.76	0.58	0.63	0.61	0.71	0.79	0.75	-	-	-
	BERT	Exact-match	0.81	0.59	0.90	0.64	0.62	0.92	0.82	0.67	0.92	0.70	0.61	0.91	0.81	0.66	0.92	**0.88**	0.90	0.99
		LLM-based	0.72	0.80	0.77	0.56	0.63	0.58	0.71	0.81	0.78	0.63	0.68	0.66	0.71	0.81	0.78	0.79	0.86	0.84
LL-8B-Gen	Statistical	Exact-match	0.51	0.57	0.57	0.68	0.76	0.76	0.69	0.76	0.76	0.65	0.69	0.70	0.68	0.75	0.75	-	-	-
		LLM-based	0.55	0.64	0.47	0.81	0.88	0.78	0.82	0.88	0.78	0.75	0.80	0.68	0.81	0.88	0.79	-	-	-
	BERT	Exact-match	0.69	0.78	0.79	0.58	0.62	0.61	0.68	0.79	0.79	0.66	0.74	0.74	0.69	0.79	0.79	0.73	0.82	0.82
		LLM-based	0.82	0.90	0.81	0.69	0.69	0.51	0.82	0.90	0.81	0.78	0.84	0.73	0.82	0.90	0.81	**0.84**	0.90	0.83
MST-7B-Gen	Statistical	Exact-match	0.58	0.58	0.76	0.63	0.66	0.80	0.65	0.68	0.82	0.59	0.59	0.77	0.65	0.68	0.82	-	-	-
		LLM-based	0.58	0.63	0.74	0.66	0.73	0.78	0.69	0.76	0.82	0.57	0.62	0.73	0.68	0.76	0.82	-	-	-
	BERT	Exact-match	0.70	0.69	0.82	0.57	0.58	0.76	0.7	0.72	0.83	0.64	0.62	0.79	0.69	0.70	0.82	0.81	0.89	0.95
		LLM-based	0.70	0.76	0.80	0.56	0.65	0.75	0.72	0.80	0.84	0.53	0.64	0.75	0.71	0.79	0.84	**0.81**	0.89	0.93
GM-7B-Gen	Statistical	Exact-match	0.56	0.48	0.68	0.59	0.63	0.77	0.58	0.61	0.75	0.59	0.57	0.74	0.55	0.61	0.74	-	-	-
		LLM-based	0.54	0.59	0.57	0.62	0.69	0.66	0.63	0.70	0.67	0.63	0.66	0.64	0.62	0.70	0.67	-	-	-
	BERT	Exact-match	0.68	0.78	0.79	0.59	0.62	0.61	0.68	0.79	0.79	0.66	0.74	0.74	0.69	0.79	0.79	0.68	0.72	0.65
		LLM-based	0.61	0.70	0.67	0.56	0.62	0.60	0.61	0.71	0.68	0.60	0.68	0.65	0.61	0.71	0.68	**0.70**	0.78	0.76
LL-70B-Gen	Statistical	Exact-match	0.51	0.59	0.61	0.45	0.56	0.53	0.45	0.57	0.56	0.53	0.55	0.56	0.47	0.58	0.55	-	-	-
		LLM-based	0.52	0.61	0.49	0.53	0.54	0.38	0.53	0.60	0.43	0.62	0.58	0.47	0.54	0.60	0.44	-	-	-
	BERT	Exact-match	0.34	0.52	0.53	0.59	0.62	0.63	0.37	0.57	0.58	0.46	0.55	0.58	0.36	0.55	0.57	0.71	0.80	0.83
		LLM-based	0.52	0.48	0.34	0.66	0.67	0.54	0.52	0.56	0.40	0.6	0.58	0.48	0.52	0.53	0.38	**0.72**	0.78	0.71
ENSB-Gen	Statistical	Exact-match	0.58	0.60	0.76	0.67	0.75	0.84	0.68	0.76	0.85	0.62	0.66	0.79	0.62	0.66	0.79	-	-	-
		LLM-based	0.57	0.63	0.69	0.72	0.81	0.82	0.73	0.83	0.85	0.63	0.69	0.74	0.73	0.83	0.84	-	-	-
	BERT	Exact-match	0.73	0.78	0.86	0.59	0.64	0.77	0.73	0.80	0.87	0.66	0.72	0.82	0.74	0.79	0.87	0.79	0.89	0.94
		LLM-based	0.76	0.84	0.85	0.58	0.66	0.71	0.76	0.86	0.87	0.61	0.74	0.77	0.76	0.85	0.86	**0.82**	0.90	0.92
HA-Test	Statistical	Exact-match	0.56	0.64	0.71	0.71	0.80	0.82	**0.74**	0.82	0.83	0.63	0.69	0.74	0.74	0.82	0.83	-	-	-
		LLM-based	0.56	0.63	0.69	0.73	0.83	0.83	**0.74**	0.84	0.84	0.64	0.71	0.75	**0.74**	0.84	0.84	-	-	-
	BERT	Exact-match	0.28	0.50	0.55	0.40	0.50	0.55	0.40	0.50	0.55	0.40	0.50	0.55	0.40	0.50	0.55	0.73	0.82	0.86
		LLM-based	0.32	0.50	0.52	0.35	0.50	0.52	0.32	0.50	0.52	0.35	0.50	0.52	0.35	0.50	0.52	0.64	0.82	0.83

Table 18: Hallucination classifier results on various test sets created using the Jeopardy dataset samples, **AUC:** Area Under Curve, **B-Acc:** Balanced Accuracy. The best result highlighted in **bold**.

Classifier	Feature Combination	F1	AUC	B-ACC
Jeopardy	EC-EC+TokenCounts	0.74	0.84	0.84
	QR-RR+TokenCounts	0.74	0.85	0.85
Kaggle	EC-EC+TokenCounts	0.79	0.73	0.84
	QR-RR+TokenCounts	0.80	0.74	0.83

Table 19: Classifier results with combination of NLI features, TokenCounts (total tokens, and special tokens count)

J HalluCounter performance on GPT4

To understand the efficiency of the HalluCounter pipeline on closed-source models, we ran our pipeline on the samples generated using the GPT4o-mini (Achiam et al., 2023) LLM. We utilized the queries from the Human annotated dataset and generated 10 responses to each query and obtained the corresponding NLI scores. As shown in Table 15, the GPT4o-mini model exhibits 13.5% hallucination rate on Jeopardy and 22.6% on the Kaggle dataset queries. Moreover, our HalluCounter pipeline exhibits more than 80% prediction confidence.

K Category-wise hallucination rates and confidence scores

We perform the category-wise results analysis to understand the category-wise hallucination rates for all test sets corresponding to the Jeopardy and Kaggle datasets. All the hallucination rates details are mentioned in Table 13 and corresponding confidence scores are listed in Table 17.

Test set	Sub-category	(Statistical, Exact-match)				(Statistical, LLM-based)				(BERT, Exact-match)				(BERT, LLM-based)			
		ACC	F1	AUC	B-ACC	ACC	F1	AUC	B-ACC	ACC	F1	AUC	B-ACC	ACC	F1	AUC	B-ACC
TL-1.1B-Gen	Arts and humanity	0.77	0.80	0.62	0.93	0.76	0.78	0.74	0.94	0.90	**0.90**	0.89	0.99	0.86	0.85	0.85	0.97
	Geography and travel	0.70	0.70	0.65	0.85	0.71	0.71	0.76	0.90	0.83	**0.84**	0.93	0.98	0.78	0.79	0.84	0.94
	Language and communication	0.70	0.76	0.61	0.94	0.70	0.74	0.70	0.94	0.90	**0.90**	0.87	0.98	0.86	0.85	0.82	0.97
	Sciences	0.74	0.76	0.65	0.92	0.72	0.74	0.75	0.93	0.84	**0.86**	0.89	0.98	0.79	0.79	0.83	0.95
	Social sciences	0.75	0.77	0.63	0.92	0.75	0.76	0.76	0.92	0.87	**0.88**	0.92	0.99	0.82	0.82	0.86	0.96
	Sports and recreation	0.77	0.79	0.64	0.92	0.78	0.79	0.77	0.94	0.90	**0.90**	0.92	0.99	0.85	0.84	0.84	0.96
PHI-3.5B-Gen	Arts and humanity	0.69	0.70	0.73	0.89	0.69	0.70	0.78	0.86	0.80	**0.81**	0.88	0.96	0.80	0.80	0.88	0.92
	Geography and travel	0.63	0.63	0.69	0.66	0.75	0.74	0.81	0.64	0.75	0.75	0.86	0.86	0.82	**0.82**	0.87	0.76
	Language and communication	0.61	0.62	0.69	0.81	0.64	0.63	0.75	0.76	0.73	**0.74**	0.83	0.91	0.73	0.73	0.82	0.83
	Sciences	0.64	0.64	0.71	0.73	0.74	0.74	0.82	0.69	0.72	0.72	0.81	0.81	0.84	**0.83**	0.89	0.81
	Social sciences	0.64	0.64	0.71	0.77	0.75	0.75	0.81	0.74	0.75	0.75	0.85	0.89	0.81	**0.80**	0.89	0.83
	Sports and recreation	0.68	0.69	0.74	0.87	0.68	0.68	0.75	0.82	0.81	**0.81**	0.91	0.96	0.75	0.75	0.83	0.87
LL-8B-Gen	Arts and humanity	0.71	0.71	0.78	0.82	0.81	0.81	0.88	0.84	0.75	0.75	0.82	0.84	0.84	**0.84**	0.90	0.88
	Geography and travel	0.75	0.73	0.75	0.64	0.89	0.88	0.92	0.72	0.80	0.78	0.86	0.80	0.91	**0.90**	0.92	0.78
	Language and communication	0.66	0.66	0.76	0.82	0.74	0.73	0.84	0.81	0.71	0.72	0.82	0.86	0.77	**0.76**	0.85	0.84
	Sciences	0.69	0.68	0.74	0.73	0.83	0.83	0.89	0.77	0.70	0.69	0.79	0.77	0.85	**0.85**	0.89	0.81
	Social sciences	0.69	0.67	0.74	0.74	0.85	0.85	0.91	0.79	0.71	0.70	0.81	0.80	0.88	**0.87**	0.92	0.83
	Sports and recreation	0.71	0.70	0.76	0.81	0.80	0.80	0.88	0.81	0.74	0.74	0.84	0.86	0.83	**0.83**	0.90	0.86
MST-7B-Gen	Arts and humanity	0.68	0.69	0.68	0.88	0.69	0.70	0.74	0.86	0.83	0.84	0.90	0.97	0.82	**0.82**	0.90	0.95
	Geography and travel	0.63	0.62	0.68	0.70	0.71	0.71	0.79	0.77	0.79	0.79	0.91	0.94	0.80	**0.80**	0.89	0.88
	Language and communication	0.61	0.64	0.65	0.85	0.63	0.65	0.72	0.83	0.79	**0.80**	0.87	0.95	0.79	0.79	0.87	0.93
	Sciences	0.65	0.65	0.71	0.80	0.69	0.69	0.78	0.81	0.79	0.79	0.89	0.94	0.82	**0.82**	0.90	0.92
	Social sciences	0.64	0.64	0.67	0.80	0.69	0.69	0.77	0.81	0.79	0.80	0.89	0.95	0.82	**0.82**	0.91	0.93
	Sports and recreation	0.66	0.68	0.70	0.87	0.69	0.69	0.77	0.86	0.82	0.82	0.90	0.97	0.81	**0.81**	0.89	0.94
GM-7B-Gen	Arts and humanity	0.71	0.71	0.78	0.82	0.60	0.60	0.70	0.79	0.74	**0.74**	0.81	0.84	0.69	0.70	0.77	0.86
	Geography and travel	0.75	**0.73**	0.75	0.64	0.70	0.66	0.68	0.55	0.74	0.71	0.76	0.67	0.73	0.70	0.76	0.66
	Language and communication	0.66	0.66	0.76	0.82	0.59	0.59	0.69	0.76	0.70	**0.71**	0.73	0.74	0.66	0.67	0.76	0.83
	Sciences	0.69	0.68	0.74	0.73	0.68	0.66	0.71	0.57	0.70	0.68	0.76	0.77	0.72	**0.71**	0.79	0.70
	Social sciences	0.69	0.67	0.74	0.74	0.68	0.65	0.69	0.64	0.69	0.66	0.76	0.76	0.73	**0.72**	0.79	0.75
	Sports and recreation	0.70	**0.70**	0.77	0.82	0.63	0.62	0.70	0.69	0.70	0.69	0.79	0.83	0.68	0.68	0.78	0.78
LL-70B-Gen	Arts and humanity	0.53	0.52	0.56	0.67	0.57	0.54	0.59	0.56	0.71	**0.70**	0.81	0.88	0.70	0.68	0.81	0.83
	Geography and travel	0.64	0.57	0.59	0.53	0.78	0.69	0.68	0.36	0.76	0.75	0.85	0.83	0.80	**0.76**	0.72	0.54
	Language and communication	0.52	0.51	0.58	0.65	0.58	0.55	0.51	0.49	0.68	**0.69**	0.75	0.82	0.69	0.67	0.72	0.68
	Sciences	0.59	0.53	0.57	0.51	0.67	0.64	0.54	0.38	0.74	0.73	0.80	0.81	0.79	**0.75**	0.80	0.69
	Social sciences	0.55	0.54	0.55	0.51	0.67	0.65	0.60	0.39	0.73	0.71	0.81	0.81	0.81	**0.77**	0.85	0.76
	Sports and recreation	0.54	0.54	0.61	0.65	0.69	0.67	0.62	0.62	0.70	**0.69**	0.80	0.83	0.73	**0.69**	0.78	0.77
ENSB-Gen	Arts and humanity	0.71	0.72	0.78	0.90	0.74	0.75	0.83	0.89	0.83	**0.84**	0.90	0.96	0.83	0.83	0.90	0.94
	Geography and travel	0.67	0.67	0.73	0.74	0.77	0.77	0.86	0.79	0.79	0.78	0.91	0.92	0.83	**0.82**	0.92	0.88
	Language and communication	0.63	0.65	0.73	0.87	0.66	0.66	0.78	0.85	0.78	**0.79**	0.86	0.94	0.77	0.77	0.86	0.91
	Sciences	0.69	0.70	0.77	0.83	0.75	0.75	0.85	0.83	0.76	0.76	0.87	0.92	0.83	**0.83**	0.92	0.91
	Social sciences	0.67	0.67	0.75	0.84	0.76	0.76	0.85	0.84	0.78	0.78	0.89	0.94	0.84	**0.84**	0.93	0.93
	Sports and recreation	0.70	0.71	0.78	0.90	0.72	0.73	0.82	0.87	0.80	0.81	0.90	0.96	0.81	**0.81**	0.88	0.93
HA-Test	Arts and humanity	0.76	**0.76**	0.82	0.88	0.74	0.75	0.82	0.88	0.68	0.68	0.79	0.90	0.62	0.61	0.81	0.90
	Geography and travel	0.76	**0.76**	0.84	0.76	0.77	**0.76**	0.84	0.75	0.77	**0.76**	0.84	0.82	0.73	0.70	0.85	0.82
	Language and communication	0.69	0.70	0.79	0.85	0.69	0.70	0.81	0.87	0.72	**0.73**	0.80	0.87	0.65	0.65	0.76	0.85
	Sciences	0.78	**0.78**	0.86	0.84	0.76	0.76	0.87	0.84	0.72	0.72	0.81	0.83	0.67	0.64	0.82	0.83
	Social sciences	0.73	0.73	0.82	0.79	0.77	**0.77**	0.86	0.84	0.76	0.75	0.83	0.85	0.71	0.69	0.87	0.89
	Sports and recreation	0.73	**0.74**	0.83	0.86	0.73	0.73	0.82	0.85	0.72	0.71	0.83	0.87	0.63	0.62	0.76	0.83

Table 20: Category wise results on Jeopardy test sets; **(Statistical, Exact-match)** - Statistical classifier trained on Exact-match based labels, **(Statistical, LLM-based)** - Statistical classifier trained on LLM-based labels, **(BERT, Exact-match)** - BERT classifier trained on Exact-match based labels, **(BERT, LLM-based)** - BERT classifier trained on LLM-based labels; we report the best classifier combination results for each LLM. The best result highlighted in **bold**.

Test set	Sub-category	(Statistical, Exact-match)				(Statistical, LLM-based)				(BERT, Exact-match)				(BERT, LLM-based)			
		ACC	F1	AUC	B-ACC	ACC	F1	AUC	B-ACC	ACC	F1	AUC	B-ACC	ACC	F1	AUC	B-ACC
TL-1.1B-Gen	GK	0.60	0.61	0.71	0.83	0.78	0.77	0.82	0.87	0.75	0.73	0.74	0.83	0.79	**0.77**	0.86	0.90
	MathQA	0.73	0.80	0.53	0.95	0.92	0.95	0.66	1	0.88	0.90	0.83	0.99	0.99	**0.99**	0.74	1
	MathQSA	0.71	0.76	0.56	0.93	0.88	0.92	0.55	0.99	0.79	0.82	0.54	0.93	0.97	**0.97**	0.69	0.99
	SciQ	0.59	0.63	0.62	0.86	0.71	0.70	0.68	0.84	0.73	**0.74**	0.65	0.87	0.71	0.72	0.73	0.87
PHI-3.5B-Gen	GK	0.71	0.70	0.69	0.46	0.75	0.76	0.64	0.34	0.74	0.70	0.66	0.50	0.85	**0.82**	0.64	0.37
	MathQA	0.63	0.73	0.53	0.95	0.65	0.67	0.56	0.83	0.89	**0.91**	0.89	0.99	0.82	0.80	0.75	0.91
	MathQSA	0.61	0.69	0.50	0.91	0.63	0.65	0.64	0.82	0.73	**0.78**	0.70	0.96	0.74	0.73	0.72	0.86
	SciQ	0.57	0.57	0.58	0.48	0.69	0.71	0.67	0.38	0.61	0.59	0.64	0.54	0.75	**0.75**	0.68	0.39
LL-8B-Gen	GK	0.73	0.71	0.66	0.48	0.82	**0.82**	0.71	0.38	0.73	0.68	0.70	0.49	0.82	**0.82**	0.68	0.38
	MathQA	0.67	0.73	0.62	0.93	0.77	0.76	0.67	0.88	0.78	**0.82**	0.81	0.97	0.82	0.81	0.78	0.92
	MathQSA	0.62	0.64	0.66	0.85	0.71	0.70	0.69	0.82	0.73	0.74	0.70	0.86	0.76	**0.76**	0.77	0.86
	SciQ	0.62	0.61	0.70	0.65	0.73	0.72	0.72	0.51	0.62	0.58	0.74	0.70	0.77	**0.75**	0.76	0.55
MST-7B-Gen	GK	0.46	0.46	0.41	0.37	0.47	0.49	0.46	0.32	0.62	**0.61**	0.62	0.56	0.35	0.21	0.66	0.52
	MathQA	0.93	**0.91**	0.52	0.95	0.93	0.90	0.55	0.94	0.90	**0.91**	0.86	0.99	0.93	0.90	0.76	0.98
	MathQSA	0.89	0.86	0.53	0.91	0.91	0.87	0.55	0.93	0.82	0.84	0.75	0.96	0.91	**0.87**	0.68	0.95
	SciQ	0.50	0.50	0.49	0.55	0.51	0.51	0.49	0.37	0.62	**0.62**	0.70	0.74	0.39	0.25	0.62	0.50
GM-7B-Gen	GK	0.67	0.63	0.66	0.61	0.72	**0.68**	0.65	0.52	0.65	0.59	0.68	0.66	0.72	0.66	0.65	0.54
	MathQA	0.60	0.68	0.48	0.90	0.66	0.72	0.58	0.92	0.70	0.76	0.72	0.96	0.73	**0.77**	0.51	0.89
	MathQSA	0.60	0.64	0.51	0.84	0.66	0.68	0.52	0.83	0.59	0.64	0.67	0.89	0.74	**0.73**	0.54	0.83
	SciQ	0.58	0.56	0.65	0.67	0.68	0.66	0.66	0.51	0.50	0.49	0.50	0.57	0.71	**0.68**	0.68	0.53
LL-70B-Gen	GK	0.72	0.72	0.70	0.51	0.88	**0.88**	0.84	0.57	0.73	0.68	0.76	0.58	0.89	**0.88**	0.83	0.56
	MathQA	0.63	0.70	0.67	0.93	0.82	0.81	0.74	0.91	0.84	**0.85**	0.73	0.94	0.86	**0.85**	0.78	0.92
	MathQSA	0.62	0.65	0.71	0.87	0.78	0.76	0.79	0.88	0.78	0.78	0.77	0.89	0.83	**0.82**	0.84	0.90
	SciQ	0.63	0.62	0.68	0.63	0.74	0.75	0.76	0.51	0.65	0.61	0.75	0.71	0.80	**0.79**	0.77	0.56
ENSB-Gen	GK	0.65	0.63	0.73	0.69	0.77	0.76	0.76	0.69	0.73	0.71	0.77	0.74	0.79	**0.79**	0.78	0.71
	MathQA	0.68	0.76	0.57	0.95	0.78	0.80	0.70	0.94	0.85	**0.88**	0.85	0.99	0.89	**0.88**	0.84	0.97
	MathQSA	0.63	0.70	0.55	0.91	0.77	0.79	0.68	0.92	0.70	0.75	0.72	0.95	0.85	**0.85**	0.81	0.96
	SciQ	0.58	0.58	0.68	0.71	0.70	0.70	0.76	0.67	0.66	0.65	0.74	0.75	0.74	**0.74**	0.80	0.68
HA-Test	GK	0.70	0.70	0.75	0.61	0.77	**0.77**	0.78	0.65	0.69	0.59	0.74	0.60	0.71	0.63	0.71	0.55
	MathQA	0.71	0.80	0.66	0.98	0.80	0.87	0.67	0.98	0.97	**0.95**	0.50	0.97	0.97	**0.95**	0.50	0.97
	MathQSA	0.65	0.76	0.63	0.97	0.79	0.84	0.61	0.97	0.96	**0.93**	0.50	0.96	0.96	**0.93**	0.50	0.96
	SciQ	0.62	0.61	0.70	0.65	**0.65**	0.65	0.71	0.67	0.56	0.43	0.74	0.72	0.56	0.44	0.71	0.70

Table 21: Dataset-wise results on Kaggle test sets; **(Statistical, Exact-match)** - Statistical classifier trained on Exact-match based labels, **(Statistical, LLM-based)** - Statistical classifier trained on LLM-based labels, **(BERT, Exact-match)** - BERT classifier trained on Exact-match based labels, **(BERT, LLM-based)** - BERT classifier trained on LLM-based labels; we report the best classifier combination results for each LLM. The best result highlighted in **bold**.

Role	Content
System	You are a helpful AI assistant. Provide the answer to the question, do not provide any extra information.
User	`{question}`

Table 22: Prompt for response generation to a query, we used the same prompt for all the different LLMs inference

Test set	Jeopardy			Kaggle		
	Labeling strategy	Feature combination	Classifier	Labeling strategy	Feature combination	Classifier
TL-1.1B-Gen	Exact-match	q-r+(Q-R)+(R-R)	BERT	LLM-based	q-r+(Q-R)+(R-R)	BERT
PHI-3.5B-Gen	Exact-match	q-r+(Q-R)+(R-R)	BERT	LLM-based	q-r+(Q-R)+(R-R)	BERT
LL-8B-Gen	LLM-based	q-r+(Q-R)+(R-R)	BERT	Exact-match	q-r+(Q-R)+(R-R)	BERT
MST-7B-Gen	LLM-based	q-r+(Q-R)+(R-R)	BERT	Exact-match	q-r+(Q-R)+(R-R)	BERT
GM-7B-Gen	LLM-based	q-r+(Q-R)+(R-R)	BERT	LLM-based	(Q-R) + (R-R)	BERT
LL-70B-Gen	LLM-based	q-r+(Q-R)+(R-R)	BERT	LLM-based	q-r+(Q-R)+(R-R)	BERT
ENSB-Gen	LLM-based	q-r+(Q-R)+(R-R)	BERT	LLM-based	q-r+(Q-R)+(R-R)	BERT
HA-Test	Human-annotated	EC-EC	Statistical	Human-annotated	EC-EC	Statistical

Table 23: Best feature combination for each test set, including the associated classifier and labeling strategy.

Test set	Sub-category	QR			RR			EC-EC			C-C			QR+RR		
		F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC
TL-1.1B-Gen	Arts and humanity	0.71	0.59	0.93	0.78	0.57	0.92	0.79	0.63	0.94	0.71	0.58	0.93	**0.80**	0.62	0.93
	Geography and travel	0.62	0.60	0.83	0.65	0.57	0.82	**0.70**	0.66	0.86	0.64	0.60	0.84	**0.70**	0.65	0.85
	Language and communication	0.71	0.58	0.93	0.75	0.56	0.92	**0.76**	0.60	0.94	0.70	0.60	0.94	**0.76**	0.61	0.94
	Sciences	0.66	0.58	0.90	0.72	0.60	0.90	0.74	0.65	0.92	0.68	0.60	0.90	**0.76**	0.65	0.92
	Social Sciences	0.70	0.59	0.91	0.76	0.58	0.90	**0.77**	0.65	0.92	0.69	0.59	0.91	**0.77**	0.63	0.92
	Sports and recreation	0.70	0.57	0.91	0.76	0.57	0.91	0.78	0.64	0.92	0.70	0.59	0.92	**0.79**	0.64	0.92
	Average	0.68	0.59	0.90	0.74	0.58	0.90	0.76	0.64	0.92	0.69	0.59	0.91	**0.76**	0.63	0.91
PHI-3.5B-Gen	Arts and humanity	0.61	0.59	0.82	0.66	0.69	0.87	**0.70**	0.73	0.89	0.64	0.64	0.85	**0.70**	0.73	0.89
	Geography and travel	0.51	0.57	0.54	**0.63**	0.69	0.66	**0.63**	0.68	0.66	0.54	0.58	0.57	0.62	0.69	0.66
	Language and communication	0.56	0.57	0.74	0.59	0.66	0.80	**0.62**	0.69	0.81	0.55	0.58	0.75	0.60	0.67	0.80
	Sciences	0.51	0.58	0.59	**0.64**	0.71	0.70	**0.64**	0.71	0.71	0.57	0.64	0.63	**0.64**	0.71	0.73
	Social sciences	0.51	0.56	0.65	0.62	0.70	0.75	**0.64**	0.71	0.77	0.56	0.61	0.69	0.63	0.71	0.77
	Sports and recreation	0.60	0.64	0.83	0.66	0.70	0.85	0.69	0.74	0.87	0.60	0.63	0.82	**0.70**	0.76	0.88
	Average	0.55	0.59	0.70	0.63	0.69	0.77	**0.65**	0.71	0.79	0.58	0.61	0.72	**0.65**	0.71	0.79
LL-8B-Gen	Arts and humanity	0.52	0.57	0.59	0.69	0.78	0.81	**0.71**	0.78	0.82	0.66	0.72	0.76	**0.71**	0.78	0.82
	Geography and travel	0.48	0.56	0.41	0.72	0.75	0.64	**0.73**	0.75	0.64	0.69	0.68	0.57	0.69	0.73	0.61
	Language and communication	0.52	0.58	0.68	**0.66**	0.76	0.82	0.64	0.77	0.82	0.64	0.71	0.78	**0.66**	0.76	0.82
	Sciences	0.52	0.59	0.55	**0.68**	0.74	0.73	0.67	0.75	0.73	0.64	0.67	0.65	0.67	0.74	0.72
	Social sciences	0.51	0.56	0.56	0.65	0.76	0.74	**0.67**	0.74	0.74	0.63	0.68	0.67	0.66	0.74	0.73
	Sports and recreation	0.52	0.59	0.65	0.66	0.77	0.80	0.70	0.76	0.81	0.62	0.70	0.75	**0.70**	0.77	0.82
	Average	0.51	0.58	0.57	0.68	0.76	0.76	**0.69**	0.76	0.76	0.65	0.69	0.70	0.68	0.75	0.75
MST-7B-Gen	Arts and humanity	0.61	0.56	0.82	0.67	0.65	0.86	**0.69**	0.67	0.87	0.63	0.59	0.84	**0.69**	0.68	0.88
	Geography and travel	0.54	0.59	0.64	0.59	0.65	0.68	0.61	0.68	0.71	0.55	0.60	0.65	**0.62**	0.68	0.70
	Language and communication	0.60	0.55	0.80	0.63	0.64	0.85	**0.64**	0.66	0.85	0.60	0.56	0.81	**0.64**	0.65	0.85
	Sciences	0.58	0.60	0.72	0.59	0.66	0.77	0.64	0.70	0.79	0.56	0.58	0.73	**0.65**	0.71	0.80
	Social sciences	0.59	0.58	0.75	0.63	0.65	0.79	0.63	0.67	0.80	0.60	0.60	0.78	**0.64**	0.67	0.80
	Sports and recreation	0.59	0.58	0.82	**0.68**	0.68	0.86	0.67	0.69	0.87	0.61	0.60	0.83	**0.68**	0.70	0.87
	Average	0.59	0.58	0.76	0.63	0.66	0.80	**0.65**	0.68	0.82	0.59	0.59	0.77	**0.65**	0.68	0.82
GM-7B-Gen	Arts and humanity	0.52	0.57	0.59	0.69	0.78	0.81	**0.71**	0.78	0.82	0.66	0.72	0.76	**0.71**	0.78	0.82
	Geography and travel	0.48	0.56	0.41	0.72	0.75	0.64	**0.73**	0.75	0.64	0.69	0.68	0.57	0.69	0.73	0.61
	Language and communication	0.52	0.58	0.68	**0.66**	0.76	0.82	0.64	0.77	0.82	0.64	0.71	0.78	**0.66**	0.76	0.82
	Sciences	0.52	0.59	0.55	0.68	0.74	0.73	**0.67**	0.75	0.73	0.64	0.67	0.65	**0.67**	0.74	0.72
	Social sciences	0.51	0.56	0.56	0.65	0.76	0.74	**0.67**	0.74	0.74	0.63	0.68	0.67	0.66	0.74	0.73
	Sports and recreation	0.52	0.59	0.65	0.66	0.77	0.80	**0.70**	0.76	0.81	0.62	0.70	0.75	**0.70**	0.77	0.82
	Average	0.51	0.58	0.57	0.68	0.76	0.76	**0.69**	0.76	0.76	0.65	0.69	0.70	0.68	0.75	0.75
LL-70B-Gen	Arts and humanity	**0.52**	0.56	0.67	0.34	0.54	0.59	0.38	0.53	0.64	0.51	0.56	0.65	0.44	0.56	0.64
	Geography and travel	0.47	0.57	0.51	0.54	0.58	0.45	**0.57**	0.59	0.53	0.56	0.53	0.44	0.52	0.58	0.45
	Language and communication	**0.51**	0.58	0.65	0.40	0.54	0.62	0.41	0.56	0.62	0.49	0.53	0.62	0.45	0.59	0.64
	Sciences	0.49	0.62	0.62	**0.53**	0.57	0.51	0.46	0.58	0.49	0.51	0.55	0.51	0.50	0.59	0.54
	Social sciences	0.51	0.59	0.54	0.49	0.57	0.50	0.49	0.58	0.51	**0.54**	0.55	0.51	0.50	0.57	0.49
	Sports and recreation	**0.54**	0.61	0.65	0.39	0.55	0.52	0.40	0.57	0.57	**0.54**	0.56	0.64	0.42	0.59	0.57
	Average	0.51	0.59	0.61	0.45	0.56	0.53	0.45	0.57	0.56	**0.53**	0.55	0.56	0.47	0.58	0.56
ENSB-Gen	Arts and humanity	0.60	0.58	0.79	0.71	0.76	0.88	**0.72**	0.78	0.90	0.64	0.67	0.84	0.64	0.67	0.84
	Geography and travel	0.55	0.61	0.64	**0.67**	0.73	0.74	0.66	0.73	0.74	0.61	0.67	0.69	0.61	0.67	0.69
	Language and communication	0.60	0.58	0.80	**0.65**	0.73	0.87	**0.65**	0.72	0.87	0.61	0.62	0.82	0.61	0.62	0.82
	Sciences	0.57	0.62	0.73	0.65	0.75	0.82	**0.70**	0.77	0.83	0.60	0.66	0.76	0.60	0.66	0.76
	Social sciences	0.54	0.57	0.73	**0.67**	0.76	0.84	**0.67**	0.75	0.84	0.60	0.64	0.78	0.60	0.64	0.78
	Sports and recreation	0.60	0.64	0.84	0.69	0.75	0.88	**0.71**	0.78	0.90	0.65	0.69	0.86	0.65	0.69	0.86
	Average	0.58	0.60	0.76	0.67	0.75	0.84	**0.69**	0.76	0.85	0.62	0.66	0.79	0.62	0.66	0.79
HA-Test	Arts and humanity	0.58	0.60	0.76	0.73	0.80	0.87	0.75	0.82	0.89	0.64	0.69	0.82	**0.76**	0.82	0.88
	Geography and travel	0.55	0.66	0.60	0.74	0.80	0.73	0.72	0.82	0.74	0.69	0.76	0.67	**0.76**	0.84	0.76
	Language and communication	0.56	0.60	0.73	0.67	0.79	0.86	**0.70**	0.79	0.85	0.61	0.64	0.76	**0.70**	0.78	0.85
	Sciences	0.58	0.67	0.71	0.71	0.80	0.80	**0.78**	0.86	0.84	0.62	0.69	0.72	0.74	0.83	0.82
	Social sciences	0.53	0.62	0.66	0.71	0.81	0.80	0.72	0.81	0.81	0.61	0.65	0.69	**0.73**	0.82	0.79
	Sports and recreation	0.57	0.68	0.77	0.71	0.80	0.83	**0.74**	0.83	0.86	0.63	0.71	0.78	0.72	0.81	0.85
	Average	0.56	0.64	0.71	0.71	0.80	0.82	**0.74**	0.82	0.83	0.63	0.69	0.74	**0.74**	0.82	0.83

Table 24: Hallucination detection with statistical classifier results for various models trained on labels obtained from Exact-match based approach on Jeopardy test sets. The best result highlighted in **bold**.

Test set	Sub-category	QR			RR			EC-EC			C-C			QR+RR		
		F1	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**
TL-1.1B-Gen	GK	0.50	0.54	0.71	0.59	0.69	0.82	**0.61**	0.71	0.83	0.54	0.62	0.76	0.60	0.72	0.82
	MathQA	0.75	0.52	0.94	0.74	0.52	0.95	0.80	0.53	0.95	0.74	0.54	0.95	**0.81**	0.54	0.95
	MathQSA	0.70	0.54	0.92	0.71	0.52	0.92	0.75	0.55	0.93	0.68	0.52	0.92	**0.76**	0.56	0.93
	SciQ	0.55	0.54	0.82	**0.63**	0.62	0.86	0.60	0.63	0.86	0.56	0.56	0.84	0.62	0.64	0.87
	Average	0.63	0.54	0.85	0.67	0.59	0.89	0.69	0.61	0.89	0.63	0.56	0.87	**0.70**	0.62	0.89
PHI-3.5B-Gen	GK	0.53	0.49	0.28	0.64	0.61	0.4	0.69	0.70	0.42	0.57	0.54	0.35	**0.70**	0.69	0.46
	MathQA	**0.73**	0.53	0.95	0.63	0.54	0.95	0.69	0.57	0.95	0.71	0.55	0.95	0.70	0.57	0.95
	MathQSA	**0.69**	0.50	0.91	0.65	0.55	0.92	0.65	0.51	0.91	0.68	0.50	0.91	0.64	0.52	0.91
	SciQ	0.50	0.47	0.39	0.56	0.60	0.50	**0.57**	0.58	0.48	0.50	0.50	0.42	0.56	0.57	0.47
	Average	0.61	0.50	0.63	0.62	0.58	0.69	**0.65**	0.59	0.69	0.62	0.52	0.66	**0.65**	0.59	0.70
LL-8B-Gen	GK	0.54	0.57	0.33	0.66	0.67	0.45	0.70	0.74	0.52	0.63	0.60	0.42	**0.71**	0.66	0.48
	MathQA	0.70	0.54	0.91	0.69	0.61	0.93	**0.73**	0.62	0.93	0.69	0.57	0.92	**0.73**	0.62	0.93
	MathQSA	0.60	0.53	0.78	0.60	0.64	0.84	**0.64**	0.66	0.85	0.59	0.56	0.79	0.60	0.63	0.83
	SciQ	0.51	0.50	0.47	**0.61**	0.70	0.65	0.59	0.66	0.62	0.57	0.60	0.55	0.59	0.67	0.62
	Average	0.59	0.54	0.62	0.64	0.66	0.72	**0.67**	0.67	0.73	0.62	0.58	0.67	0.66	0.65	0.72
MST-7B-Gen	GK	0.27	0.53	0.46	0.44	0.46	0.39	**0.46**	0.41	0.37	0.44	0.41	0.38	0.45	0.47	0.41
	MathQA	**0.91**	0.52	0.95	0.28	0.49	0.94	0.60	0.44	0.93	0.27	0.43	0.93	0.14	0.51	0.94
	MathQSA	**0.86**	0.53	0.91	0.29	0.49	0.90	0.66	0.47	0.89	0.39	0.39	0.88	0.27	0.49	0.90
	SciQ	0.42	0.51	0.57	0.34	0.48	0.54	**0.50**	0.49	0.55	0.42	0.49	0.55	0.39	0.51	0.57
	Average	**0.62**	0.52	0.72	0.34	0.48	0.69	0.56	0.45	0.69	0.38	0.43	0.69	0.31	0.50	0.71
GM-7B-Gen	GK	0.42	0.46	0.46	**0.63**	0.66	0.61	0.58	0.65	0.58	0.61	0.66	0.63	0.53	0.65	0.54
	MathQA	**0.68**	0.48	0.90	0.60	0.56	0.92	0.60	0.55	0.92	0.64	0.52	0.91	0.60	0.55	0.91
	MathQSA	**0.64**	0.51	0.84	0.58	0.64	0.88	0.59	0.63	0.88	0.60	0.55	0.85	0.57	0.62	0.87
	SciQ	0.50	0.49	0.53	**0.56**	0.65	0.67	0.54	0.60	0.61	0.51	0.55	0.57	0.52	0.62	0.62
	Average	0.56	0.49	0.68	**0.59**	0.63	0.77	0.58	0.61	0.75	**0.59**	0.57	0.74	0.56	0.61	0.74
LL-70B-Gen	GK	0.47	0.52	0.31	**0.72**	0.70	0.51	**0.72**	0.70	0.49	0.58	0.61	0.44	0.69	0.66	0.48
	MathQA	0.66	0.51	0.89	0.69	0.65	0.93	**0.70**	0.67	0.93	0.66	0.54	0.90	0.69	0.65	0.93
	MathQSA	0.58	0.49	0.75	**0.65**	0.71	0.87	0.63	0.67	0.85	0.59	0.56	0.79	0.63	0.68	0.86
	SciQ	0.49	0.48	0.44	**0.62**	0.68	0.63	0.60	0.65	0.61	0.53	0.58	0.53	0.59	0.63	0.59
	Average	0.55	0.50	0.60	**0.67**	0.69	0.74	0.66	0.67	0.72	0.59	0.57	0.67	0.65	0.66	0.72
ENSB-Gen	GK	0.53	0.52	0.46	**0.63**	0.71	0.68	**0.63**	0.73	0.69	0.58	0.64	0.62	0.58	0.64	0.62
	MathQA	0.72	0.54	0.94	0.71	0.55	0.94	**0.76**	0.57	0.95	0.72	0.56	0.94	0.72	0.56	0.94
	MathQSA	0.67	0.54	0.90	0.68	0.52	0.90	**0.70**	0.55	0.91	0.66	0.53	0.90	0.66	0.53	0.90
	SciQ	0.49	0.47	0.54	**0.58**	0.68	0.71	0.57	0.65	0.69	0.52	0.55	0.59	0.52	0.55	0.59
	Average	0.60	0.52	0.71	0.65	0.62	0.81	**0.67**	0.63	0.81	0.62	0.57	0.76	0.62	0.57	0.76
HA-Test	GK	0.57	0.52	0.36	0.65	0.72	0.59	**0.70**	0.75	0.61	0.58	0.64	0.53	**0.70**	0.72	0.60
	MathQA	0.76	0.54	0.98	0.75	0.59	0.98	0.78	0.62	0.98	0.75	0.62	0.98	**0.80**	0.66	0.98
	MathQSA	0.72	0.61	0.97	0.73	0.58	0.97	**0.76**	0.63	0.97	0.71	0.56	0.96	0.75	0.63	0.97
	SciQ	0.50	0.48	0.47	0.60	0.70	0.66	**0.61**	0.68	0.64	0.53	0.55	0.51	**0.61**	0.70	0.65
	Average	0.64	0.54	0.70	0.68	0.65	0.80	0.71	0.67	0.80	0.64	0.59	0.75	**0.72**	0.68	0.80

Table 25: Hallucination detection with statistical classifier results for various models trained on labels obtained from Exact-match based approach on Kaggle test sets. The best result highlighted in **bold**.

Test set	Sub-category	QR			RR			EC-EC			C-C			QR+RR		
		F1	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**	**F1**	**AUC**	**B-ACC**
TL-1.1B-Gen	Arts and humanity	0.67	0.66	0.92	0.76	0.67	0.92	**0.78**	0.74	0.94	0.67	0.63	0.92	**0.78**	0.74	0.94
	Geography and travel	0.63	0.68	0.86	0.68	0.70	0.86	**0.71**	0.76	0.90	0.62	0.64	0.85	**0.71**	0.76	0.90
	Language and communication	0.66	0.65	0.93	0.73	0.64	0.91	**0.74**	0.71	0.94	0.64	0.60	0.91	**0.74**	0.70	0.94
	Sciences	0.64	0.65	0.89	0.72	0.69	0.89	**0.74**	0.75	0.93	0.63	0.62	0.88	0.73	0.75	0.92
	Social sciences	0.66	0.68	0.90	0.72	0.67	0.89	0.75	0.76	0.92	0.64	0.62	0.88	**0.76**	0.76	0.92
	Sports and recreation	0.66	0.67	0.91	0.76	0.71	0.92	0.78	0.76	0.94	0.64	0.58	0.89	**0.79**	0.77	0.94
	Average	0.65	0.67	0.90	0.73	0.68	0.90	**0.75**	0.75	0.93	0.64	0.62	0.89	**0.75**	0.75	0.93
PHI-3.5B-Gen	Arts and humanity	0.54	0.56	0.71	0.69	0.77	0.86	**0.70**	0.78	0.86	0.60	0.65	0.78	**0.70**	0.78	0.86
	Geography and travel	0.54	0.59	0.40	0.73	0.78	0.60	**0.75**	0.81	0.64	0.58	0.64	0.46	0.74	0.81	0.64
	Language and communication	0.54	0.58	0.63	0.62	0.74	0.76	**0.63**	0.74	0.76	0.52	0.56	0.61	**0.63**	0.75	0.76
	Sciences	0.50	0.57	0.43	0.74	0.81	0.69	**0.76**	0.82	0.71	0.61	0.67	0.52	0.74	0.82	0.69
	Social sciences	0.57	0.63	0.52	0.73	0.79	0.72	**0.75**	0.81	0.74	0.58	0.64	0.58	**0.75**	0.81	0.74
	Sports and recreation	0.51	0.55	0.65	0.63	0.70	0.80	0.66	0.76	0.82	0.60	0.62	0.72	**0.68**	0.75	0.82
	Average	0.53	0.58	0.56	0.69	0.77	0.74	**0.71**	0.79	0.76	0.58	0.63	0.61	**0.71**	0.79	0.75
LL-8B-Gen	Arts and humanity	0.53	0.62	0.52	0.80	0.88	0.83	**0.81**	0.88	0.84	0.72	0.79	0.74	0.80	0.88	0.84
	Geography and travel	0.61	0.66	0.32	**0.89**	0.93	0.73	0.88	0.91	0.70	0.84	0.83	0.60	0.88	0.92	0.72
	Language and communication	0.54	0.63	0.60	0.72	0.83	0.80	**0.73**	0.84	0.81	0.67	0.74	0.71	0.71	0.83	0.80
	Sciences	0.54	0.63	0.41	0.82	0.87	0.74	**0.83**	0.89	0.77	0.74	0.79	0.64	**0.83**	0.88	0.75
	Social sciences	0.52	0.63	0.42	0.83	0.90	0.77	**0.85**	0.90	0.77	0.78	0.82	0.66	**0.85**	0.91	0.79
	Sports and recreation	0.57	0.66	0.56	0.79	0.87	0.81	**0.80**	0.88	0.81	0.73	0.81	0.74	**0.80**	0.88	0.81
	Average	0.55	0.64	0.47	0.81	0.88	0.78	**0.82**	0.88	0.78	0.75	0.80	0.68	0.81	0.88	0.79
ENSB-Gen	Arts and humanity	0.57	0.59	0.73	0.73	0.81	0.87	**0.75**	0.83	0.89	0.63	0.69	0.80	0.74	0.82	0.88
	Geography and travel	0.58	0.67	0.62	0.76	0.84	0.75	**0.77**	0.86	0.79	0.68	0.74	0.67	0.76	0.85	0.78
	Language and communication	0.56	0.61	0.74	**0.65**	0.76	0.83	**0.65**	0.78	0.85	0.59	0.64	0.76	0.66	0.78	0.85
	Sciences	0.55	0.64	0.64	0.73	0.84	0.81	0.74	0.85	0.83	0.61	0.69	0.69	**0.75**	0.85	0.83
	Social sciences	0.56	0.64	0.68	0.74	0.83	0.82	**0.76**	0.85	0.84	0.63	0.71	0.72	**0.76**	0.85	0.84
	Sports and recreation	0.57	0.64	0.75	0.69	0.79	0.85	**0.73**	0.82	0.87	0.61	0.68	0.78	0.72	0.81	0.87
	Average	0.57	0.63	0.69	0.72	0.81	0.82	**0.73**	0.83	0.85	0.63	0.69	0.74	**0.73**	0.83	0.84
GM-7B-Gen	Arts and humanity	0.58	0.61	0.72	0.59	0.69	0.79	**0.60**	0.70	0.79	**0.60**	0.66	0.77	**0.60**	0.70	0.79
	Geography and travel	0.55	0.57	0.43	**0.66**	0.67	0.53	**0.66**	0.65	0.52	0.65	0.65	0.49	**0.66**	0.68	0.55
	Language and communication	0.55	0.59	0.70	0.56	0.71	0.77	0.56	0.72	0.78	**0.59**	0.69	0.76	0.56	0.70	0.77
	Sciences	0.53	0.60	0.46	**0.66**	0.70	0.57	**0.66**	0.71	0.57	**0.66**	0.67	0.54	**0.66**	0.71	0.57
	Social sciences	0.51	0.58	0.52	**0.65**	0.69	0.64	**0.65**	0.70	0.65	**0.65**	0.66	0.60	**0.65**	0.69	0.64
	Sports and recreation	0.50	0.56	0.60	0.59	0.69	0.68	**0.62**	0.70	0.69	0.61	0.65	0.66	0.61	0.71	0.69
	Average	0.54	0.59	0.57	0.62	0.69	0.66	**0.63**	0.70	0.67	0.63	0.66	0.64	0.62	0.70	0.67
MST-7B-Gen	Arts and humanity	0.60	0.61	0.79	0.67	0.71	0.84	**0.70**	0.74	0.86	0.59	0.62	0.80	0.69	0.74	0.86
	Geography and travel	0.58	0.67	0.67	0.65	0.74	0.70	0.70	0.79	0.77	0.55	0.63	0.65	**0.71**	0.79	0.77
	Language and communication	0.57	0.60	0.76	0.63	0.71	0.81	**0.65**	0.72	0.83	0.56	0.58	0.76	0.63	0.71	0.83
	Sciences	0.58	0.64	0.70	0.66	0.74	0.76	0.68	0.78	0.80	0.56	0.61	0.69	**0.69**	0.78	0.81
	Social sciences	0.57	0.64	0.72	0.65	0.72	0.77	**0.69**	0.77	0.81	0.57	0.64	0.73	0.68	0.77	0.81
	Sports and recreation	0.58	0.63	0.77	0.67	0.74	0.82	**0.69**	0.77	0.86	0.58	0.62	0.77	0.68	0.76	0.85
	Average	0.58	0.63	0.74	0.66	0.73	0.78	**0.69**	0.76	0.82	0.57	0.62	0.73	0.68	0.76	0.82
LL-70B-Gen	Arts and humanity	0.52	0.59	0.57	0.38	0.54	0.54	0.37	0.57	0.55	**0.54**	0.59	0.56	0.40	0.60	0.58
	Geography and travel	0.52	0.66	0.42	0.68	0.56	0.26	**0.69**	0.67	0.34	**0.69**	0.59	0.36	**0.69**	0.68	0.36
	Language and communication	0.53	0.57	0.49	0.44	0.49	0.41	0.44	0.55	0.45	**0.55**	0.51	0.49	0.47	0.55	0.47
	Sciences	0.49	0.62	0.46	0.60	0.52	0.32	0.59	0.60	0.35	**0.64**	0.54	0.38	0.60	0.57	0.34
	Social sciences	0.56	0.63	0.44	0.62	0.59	0.35	0.62	0.59	0.36	**0.65**	0.60	0.39	0.63	0.60	0.40
	Sports and recreation	0.51	0.60	0.55	0.44	0.52	0.41	0.44	0.64	0.52	**0.67**	0.62	0.62	0.44	0.61	0.51
	Average	0.52	0.61	0.49	0.53	0.54	0.38	0.53	0.60	0.43	**0.62**	0.58	0.47	0.54	0.60	0.44
HA-Test	Arts and humanity	0.57	0.58	0.73	0.73	0.82	0.88	**0.75**	0.83	0.89	0.65	0.70	0.81	**0.75**	0.82	0.88
	Geography and travel	0.56	0.65	0.59	**0.76**	0.84	0.75	0.75	0.85	0.75	0.70	0.77	0.67	0.75	0.85	0.75
	Language and communication	0.57	0.62	0.75	**0.70**	0.81	0.87	0.69	0.82	0.87	0.57	0.65	0.76	0.69	0.81	0.87
	Sciences	0.53	0.63	0.68	0.74	0.86	0.83	**0.76**	0.87	0.84	0.65	0.74	0.74	0.75	0.86	0.84
	Social sciences	0.58	0.66	0.69	0.75	0.83	0.81	0.76	0.85	0.84	0.61	0.69	0.72	**0.77**	0.86	0.84
	Sports and recreation	0.57	0.65	0.72	0.67	0.79	0.82	0.72	0.82	0.85	0.63	0.72	0.78	**0.73**	0.82	0.85
	Average	0.56	0.63	0.69	0.73	0.83	0.83	**0.74**	0.84	0.84	0.64	0.71	0.75	**0.74**	0.84	0.84

Table 26: Hallucination detection with statistical classifier results for various models trained on labels obtained from LLM-based approach on Jeopardy test sets. The best result highlighted in **bold**.

Test set	Sub-category	QR			RR			EC-EC			C-C			QR+RR		
		F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC
TL-1.1B-Gen	GK	0.55	0.58	0.72	0.70	0.79	0.84	0.73	0.81	0.87	0.62	0.67	0.79	**0.77**	0.82	0.87
	MathQA	0.87	0.65	1	0.91	0.48	0.99	**0.95**	0.63	1	0.88	0.65	1	**0.95**	0.66	1
	MathQSA	0.83	0.57	0.99	0.90	0.49	0.99	**0.92**	0.59	0.99	0.84	0.60	0.99	**0.92**	0.55	0.99
	SciQ	0.57	0.58	0.79	**0.70**	0.68	0.84	0.69	0.70	0.85	0.57	0.57	0.80	0.69	0.69	0.85
	Average	0.71	0.60	0.88	0.80	0.61	0.92	0.82	0.68	0.93	0.73	0.62	0.90	**0.83**	0.68	0.93
PHI-3.5B-Gen	GK	0.53	0.41	0.13	0.72	0.65	0.35	0.74	0.59	0.32	0.62	0.53	0.18	**0.76**	0.64	0.34
	MathQA	**0.67**	0.56	0.83	0.60	0.58	0.84	0.65	0.61	0.85	0.66	0.56	0.84	0.66	0.62	0.85
	MathQSA	0.60	0.52	0.76	0.63	0.63	0.82	0.64	0.63	0.81	0.62	0.55	0.77	**0.65**	0.64	0.82
	SciQ	0.53	0.51	0.23	0.67	0.67	0.39	0.70	0.65	0.37	0.54	0.53	0.25	**0.71**	0.67	0.38
	Average	0.58	0.50	0.49	0.66	0.63	0.60	0.68	0.62	0.59	0.61	0.54	0.51	**0.70**	0.64	0.60
LL-8B-Gen	GK	0.48	0.52	0.18	0.78	0.67	0.33	0.81	0.71	0.37	0.62	0.58	0.25	**0.82**	0.71	0.38
	MathQA	0.67	0.54	0.83	0.74	0.66	0.88	**0.76**	0.67	0.88	0.69	0.59	0.85	**0.76**	0.67	0.88
	MathQSA	0.60	0.54	0.73	**0.70**	0.71	0.83	**0.70**	0.70	0.82	0.61	0.58	0.75	**0.70**	0.69	0.82
	SciQ	0.49	0.53	0.30	0.71	0.73	0.51	**0.72**	0.73	0.51	0.60	0.63	0.40	**0.72**	0.72	0.51
	Average	0.56	0.53	0.51	0.73	0.69	0.64	**0.75**	0.70	0.65	0.63	0.60	0.56	**0.75**	0.70	0.65
ENSB-Gen	GK	0.47	0.50	0.38	0.70	0.76	0.66	**0.76**	0.76	0.69	0.65	0.68	0.60	0.72	0.76	0.69
	MathQA	0.73	0.56	0.91	0.78	0.67	0.93	**0.80**	0.70	0.94	0.74	0.60	0.92	**0.80**	0.70	0.94
	MathQSA	0.70	0.53	0.89	0.77	0.67	0.92	0.78	0.68	0.92	0.71	0.56	0.89	**0.79**	0.68	0.92
	SciQ	0.50	0.51	0.42	0.68	0.77	0.68	0.69	0.75	0.66	0.54	0.57	0.47	**0.70**	0.76	0.67
	Average	0.60	0.53	0.65	0.73	0.72	0.80	**0.76**	0.72	0.80	0.66	0.60	0.72	0.75	0.73	0.81
GM-7B-Gen	GK	0.52	0.56	0.37	**0.68**	0.65	0.52	0.64	0.62	0.48	0.60	0.65	0.46	0.66	0.64	0.51
	MathQA	**0.72**	0.51	0.90	0.69	0.64	0.93	0.69	0.64	0.93	**0.72**	0.58	0.92	0.69	0.65	0.93
	MathQSA	**0.68**	0.52	0.83	0.64	0.70	0.90	0.66	0.72	0.91	0.65	0.57	0.86	0.65	0.71	0.91
	SciQ	0.47	0.53	0.36	**0.66**	0.68	0.53	**0.66**	0.65	0.50	0.53	0.58	0.42	**0.66**	0.66	0.51
	Average	0.60	0.53	0.62	**0.67**	0.67	0.72	0.66	0.66	0.71	0.63	0.60	0.67	**0.67**	0.67	0.72
MST-7B-Gen	GK	0.16	0.56	0.38	**0.49**	0.46	0.32	0.16	0.40	0.28	0.17	0.60	0.38	0.16	0.59	0.35
	MathQA	**0.90**	0.52	0.94	0.61	0.48	0.93	**0.90**	0.49	0.93	**0.90**	0.53	0.94	**0.90**	0.55	0.94
	MathQSA	**0.87**	0.53	0.92	0.63	0.52	0.92	**0.87**	0.49	0.91	**0.87**	0.54	0.92	**0.87**	0.55	0.93
	SciQ	0.20	0.51	0.38	**0.51**	0.49	0.37	0.21	0.49	0.37	0.20	0.52	0.39	0.21	0.52	0.38
	Average	0.53	0.53	0.66	**0.56**	0.49	0.64	0.54	0.47	0.62	0.54	0.55	0.66	0.54	0.55	0.65
LL-70B-Gen	GK	0.46	0.41	0.11	0.84	0.81	0.52	**0.88**	0.84	0.57	0.58	0.69	0.36	0.83	0.79	0.42
	MathQA	0.67	0.52	0.83	**0.81**	0.73	0.91	**0.81**	0.74	0.91	0.69	0.57	0.85	**0.81**	0.73	0.91
	MathQSA	0.60	0.51	0.72	**0.76**	0.78	0.88	**0.76**	0.79	0.88	0.62	0.56	0.76	**0.76**	0.78	0.88
	SciQ	0.47	0.52	0.26	**0.75**	0.76	0.51	0.74	0.73	0.50	0.53	0.59	0.34	0.74	0.74	0.50
	Average	0.55	0.49	0.48	0.79	0.77	0.71	**0.80**	0.78	0.72	0.61	0.60	0.58	0.79	0.76	0.68
HA-Test	GK	0.49	0.51	0.36	0.71	0.77	0.64	**0.77**	0.78	0.65	0.67	0.72	0.61	0.73	0.78	0.65
	MathQA	0.81	0.50	0.97	0.83	0.58	0.98	0.86	0.67	0.98	0.82	0.63	0.98	**0.87**	0.67	0.98
	MathQSA	0.78	0.56	0.97	0.83	0.56	0.96	**0.84**	0.61	0.97	0.78	0.47	0.95	**0.84**	0.62	0.97
	SciQ	0.50	0.48	0.48	**0.65**	0.74	0.70	**0.65**	0.71	0.67	0.51	0.55	0.53	**0.65**	0.71	0.67
	Average	0.65	0.51	0.70	0.76	0.66	0.82	**0.78**	0.69	0.82	0.70	0.59	0.77	0.77	0.70	0.82

Table 27: Hallucination detection with statistical classifier results for various models trained on labels obtained from LLM-based approach on Kaggle test sets. The best result highlighted in **bold**.

Test set	Sub-category	QR			RR			EC-EC			CC			QR-RR			q-r+Q-R+R-R		
		F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC
TL-1.1B-Test	Arts and humanity	0.67	0.61	0.94	0.85	0.58	0.93	0.85	0.66	0.94	0.65	0.62	0.94	0.85	0.64	0.94	**0.90**	0.89	0.99
	Geography and travel	0.62	0.64	0.86	0.70	0.61	0.84	0.75	0.70	0.86	0.63	0.62	0.85	0.72	0.68	0.86	**0.84**	0.93	0.98
	Language and communication	0.62	0.59	0.94	0.83	0.56	0.93	0.82	0.63	0.94	0.62	0.60	0.94	0.81	0.63	0.94	**0.90**	0.87	0.98
	Sciences	0.64	0.61	0.91	0.81	0.61	0.90	0.81	0.68	0.92	0.65	0.60	0.91	0.80	0.67	0.92	**0.86**	0.89	0.98
	Social sciences	0.65	0.62	0.92	0.81	0.59	0.91	0.82	0.69	0.93	0.64	0.61	0.91	0.81	0.66	0.92	**0.88**	0.92	0.99
	Sports and recreation	0.66	0.63	0.93	0.84	0.59	0.91	0.84	0.68	0.93	0.64	0.60	0.92	0.84	0.65	0.92	**0.90**	0.92	0.99
PHI-3.5B-Gen	Arts and humanity	0.60	0.63	0.84	0.69	0.73	0.89	0.74	0.78	0.91	0.65	0.70	0.87	0.73	0.78	0.91	**0.81**	0.88	0.96
	Geography and travel	0.56	0.61	0.57	0.64	0.69	0.68	0.65	0.72	0.69	0.61	0.62	0.59	0.66	0.72	0.69	**0.75**	0.86	0.86
	Language and communication	0.53	0.63	0.77	0.57	0.72	0.83	0.61	0.74	0.84	0.59	0.65	0.79	0.59	0.73	0.84	**0.74**	0.83	0.91
	Sciences	0.56	0.61	0.62	0.66	0.73	0.75	0.67	0.75	0.76	0.63	0.67	0.68	0.67	0.75	0.76	**0.72**	0.81	0.81
	Social sciences	0.54	0.62	0.69	0.63	0.71	0.78	0.67	0.74	0.81	0.61	0.65	0.74	0.66	0.73	0.80	**0.75**	0.85	0.89
	Sports and recreation	0.58	0.65	0.82	0.64	0.71	0.86	0.70	0.76	0.88	0.61	0.67	0.85	0.69	0.76	0.88	**0.81**	0.91	0.96
LL-8B-Gen	Arts and humanity	0.57	0.60	0.62	0.74	0.80	0.84	0.74	0.81	0.84	0.69	0.76	0.79	0.74	0.81	0.84	**0.75**	0.82	0.84
	Geography and travel	0.62	0.62	0.46	0.71	0.76	0.67	0.69	0.77	0.68	0.72	0.73	0.63	0.70	0.77	0.68	**0.78**	0.86	0.80
	Language and communication	0.55	0.63	0.70	0.65	0.80	0.85	0.65	0.81	0.86	0.62	0.75	0.81	0.66	0.81	0.86	**0.72**	0.82	0.86
	Sciences	0.59	0.63	0.58	0.67	0.75	0.76	0.67	0.77	0.78	0.66	0.74	0.72	0.67	0.76	0.77	**0.69**	0.79	0.77
	Social sciences	0.56	0.59	0.59	0.65	0.77	0.77	0.66	0.76	0.76	0.67	0.73	0.73	0.66	0.76	0.76	**0.70**	0.81	0.80
	Sports and recreation	0.59	0.65	0.69	0.69	0.79	0.82	0.68	0.81	0.83	0.66	0.74	0.78	0.69	0.80	0.83	**0.74**	0.84	0.86
MST-7B-Gen	Arts and humanity	0.59	0.57	0.82	0.75	0.69	0.87	0.75	0.72	0.89	0.67	0.62	0.85	0.74	0.71	0.88	**0.84**	0.90	0.97
	Geography and travel	0.57	0.58	0.65	0.64	0.68	0.69	0.67	0.72	0.72	0.58	0.63	0.68	0.66	0.70	0.71	**0.79**	0.91	0.94
	Language and communication	0.56	0.57	0.80	0.70	0.68	0.86	0.68	0.68	0.85	0.68	0.60	0.82	0.66	0.67	0.85	**0.80**	0.87	0.95
	Sciences	0.57	0.59	0.72	0.71	0.72	0.80	0.70	0.74	0.82	0.62	0.62	0.76	0.70	0.73	0.81	**0.79**	0.89	0.94
	Social sciences	0.57	0.58	0.76	0.70	0.68	0.81	0.70	0.71	0.83	0.66	0.63	0.79	0.69	0.69	0.82	**0.80**	0.89	0.95
	Sports and recreation	0.58	0.61	0.83	0.72	0.68	0.86	0.71	0.73	0.88	0.67	0.64	0.85	0.71	0.72	0.87	**0.82**	0.90	0.97
GM-7B-Gen	Arts and humanity	0.58	0.60	0.62	0.73	0.80	0.84	0.73	0.81	0.84	0.69	0.76	0.79	**0.74**	0.81	0.84	0.67	0.67	0.62
	Geography and travel	0.62	0.63	0.47	0.71	0.76	0.67	0.70	0.77	0.68	**0.72**	0.73	0.63	0.70	0.77	0.68	0.69	0.75	0.53
	Language and communication	0.56	0.63	0.70	0.65	0.81	0.85	0.64	0.81	0.86	0.63	0.75	0.81	0.66	0.81	0.85	**0.71**	0.73	0.74
	Sciences	0.60	0.63	0.59	0.67	0.76	0.77	0.67	0.76	0.78	0.66	0.74	0.72	**0.68**	0.76	0.77	0.67	0.72	0.67
	Social sciences	0.56	0.59	0.59	0.65	0.76	0.76	0.66	0.76	0.76	**0.67**	0.73	0.73	0.66	0.76	0.76	0.66	0.72	0.65
	Sports and recreation	0.60	0.64	0.68	**0.69**	0.79	0.83	0.68	0.81	0.83	0.67	0.74	0.78	0.68	0.80	0.83	0.68	0.72	0.69
LL-70B-Gen	Arts and humanity	0.58	0.59	0.69	0.24	0.52	0.59	0.30	0.58	0.68	0.48	0.57	0.66	0.27	0.57	0.67	**0.70**	0.81	0.88
	Geography and travel	0.62	0.64	0.54	0.45	0.61	0.55	0.46	0.61	0.54	0.60	0.51	0.45	0.46	0.59	0.51	**0.75**	0.85	0.83
	Language and communication	0.56	0.62	0.68	0.26	0.51	0.59	0.30	0.53	0.63	0.46	0.53	0.64	0.28	0.51	0.61	**0.69**	0.75	0.82
	Sciences	0.62	0.65	0.63	0.38	0.44	0.42	0.42	0.57	0.53	0.60	0.58	0.58	0.39	0.54	0.50	**0.73**	0.80	0.81
	Social sciences	0.59	0.59	0.55	0.40	0.51	0.50	0.42	0.54	0.51	0.56	0.56	0.53	0.41	0.53	0.50	**0.71**	0.81	0.81
	Sports and recreation	0.59	0.63	0.68	0.32	0.52	0.54	0.34	0.58	0.61	0.53	0.57	0.63	0.32	0.56	0.61	**0.69**	0.80	0.83
ENSB-Gen	Arts and humanity	0.60	0.61	0.80	0.76	0.80	0.90	0.78	0.82	0.91	0.66	0.71	0.86	0.78	0.81	0.91	**0.84**	0.90	0.96
	Geography and travel	0.59	0.64	0.67	0.69	0.76	0.76	0.70	0.79	0.78	0.65	0.72	0.73	0.71	0.78	0.78	**0.78**	0.91	0.92
	Language and communication	0.57	0.61	0.81	0.68	0.75	0.88	0.67	0.77	0.89	0.62	0.68	0.84	0.69	0.76	0.89	**0.79**	0.86	0.94
	Sciences	0.60	0.64	0.73	0.75	0.80	0.85	0.75	0.82	0.86	0.65	0.73	0.80	**0.76**	0.81	0.85	**0.76**	0.87	0.92
	Social sciences	0.56	0.62	0.75	0.73	0.78	0.85	0.73	0.79	0.86	0.64	0.71	0.81	0.73	0.79	0.86	**0.78**	0.89	0.94
	Sports and recreation	0.62	0.70	0.86	0.74	0.79	0.89	0.74	0.81	0.90	0.68	0.75	0.88	0.76	0.81	0.91	**0.81**	0.90	0.96
HA-Test	Arts and humanity	0.52	0.50	0.66	0.17	0.50	0.66	0.52	0.50	0.66	0.66	0.50	0.66	0.52	0.50	0.66	**0.68**	0.79	0.90
	Geography and travel	0.23	0.50	0.40	0.45	0.50	0.40	0.23	0.50	0.40	0.40	0.50	0.40	0.23	0.50	0.40	**0.76**	0.84	0.82
	Language and communication	0.50	0.50	0.64	0.19	0.50	0.64	0.50	0.50	0.64	0.64	0.50	0.64	0.50	0.50	0.64	**0.73**	0.80	0.87
	Sciences	0.35	0.50	0.52	0.32	0.50	0.52	0.35	0.50	0.52	0.52	0.50	0.52	0.35	0.50	0.52	**0.72**	0.81	0.83
	Social sciences	0.35	0.50	0.51	0.32	0.50	0.51	0.35	0.50	0.51	0.51	0.50	0.51	0.35	0.50	0.51	**0.75**	0.83	0.85
	Sports and recreation	0.42	0.50	0.58	0.25	0.50	0.58	0.42	0.50	0.58	0.58	0.50	0.58	0.42	0.50	0.58	**0.71**	0.83	0.87

Table 28: Hallucination detection with BERT classifier results for various models trained on labels obtained from Exact-match based approach on Jeopardy test sets. The best result is highlighted in **bold**.

Test set	Sub-category	QR			RR			EC-EC			CC			QR-RR			q-r+Q-R+R-R		
		F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC
TL-1.1B-Gen	SciQ	0.33	0.55	0.85	**0.74**	0.65	0.87	0.68	0.68	0.89	0.48	0.58	0.86	0.63	0.66	0.88	0.54	0.73	0.91
	MathQA	0.73	0.54	0.95	0.85	0.51	0.95	**0.87**	0.55	0.95	0.79	0.55	0.95	**0.87**	0.54	0.95	0.90	0.83	0.99
	MathQSA	0.66	0.54	0.92	0.80	0.52	0.93	**0.82**	0.56	0.93	0.72	0.55	0.93	**0.82**	0.54	0.93	0.80	0.72	0.97
	GK	0.33	0.58	0.76	**0.73**	0.74	0.83	**0.73**	0.77	0.86	0.64	0.68	0.81	0.68	0.77	0.85	0.55	0.77	0.86
PHI-3.5B-Gen	SciQ	0.50	0.46	0.41	**0.59**	0.64	0.54	0.56	0.59	0.50	0.52	0.52	0.44	0.55	0.58	0.49	0.50	0.71	0.65
	MathQA	0.69	0.54	0.95	0.54	0.55	0.95	0.64	0.57	0.95	0.70	0.57	0.95	0.68	0.58	0.95	**0.91**	0.89	0.99
	MathQSA	0.63	0.47	0.90	0.62	0.52	0.91	0.64	0.52	0.92	0.67	0.53	0.92	0.66	0.52	0.92	**0.78**	0.70	0.96
	GK	0.65	0.58	0.38	0.68	0.62	0.44	**0.70**	0.66	0.50	0.68	0.63	0.46	0.69	0.70	0.51	0.64	0.66	0.50
LL-8B-Gen	SciQ	0.48	0.46	0.46	**0.58**	0.74	0.70	0.55	0.71	0.67	0.55	0.68	0.62	0.55	0.68	0.65	0.44	0.75	0.72
	MathQA	0.72	0.54	0.91	0.79	0.64	0.93	0.80	0.65	0.94	0.74	0.61	0.93	**0.82**	0.64	0.93	**0.82**	0.81	0.97
	MathQSA	0.62	0.54	0.79	0.71	0.70	0.87	0.71	0.71	0.86	0.62	0.60	0.81	**0.74**	0.70	0.86	0.58	0.68	0.85
	GK	0.58	0.60	0.36	0.64	0.74	0.51	0.66	0.72	0.51	**0.68**	0.70	0.49	0.67	0.68	0.50	0.60	0.66	0.55
MST-7B-Gen	SciQ	0.40	0.50	0.56	0.42	0.50	0.56	0.41	0.50	0.55	0.40	0.49	0.54	0.40	0.50	0.56	**0.62**	0.70	0.74
	MathQA	**0.91**	0.45	0.93	**0.91**	0.54	0.95	**0.91**	0.52	0.94	**0.91**	0.48	0.94	**0.91**	0.49	0.94	**0.91**	0.86	0.99
	MathQSA	**0.86**	0.43	0.88	0.85	0.58	0.92	0.85	0.56	0.92	**0.86**	0.45	0.89	**0.86**	0.48	0.90	0.84	0.75	0.96
	GK	0.25	0.45	0.37	0.32	0.56	0.48	0.28	0.50	0.41	0.26	0.46	0.39	0.25	0.45	0.38	**0.61**	0.62	0.56
GM-7B-Gen	SciQ	**0.49**	0.50	0.57	0.47	0.67	0.70	0.45	0.65	0.67	0.46	0.60	0.63	0.48	0.62	0.66	0.35	0.71	0.74
	MathQA	0.64	0.45	0.89	0.60	0.58	0.92	0.57	0.56	0.91	0.59	0.55	0.92	0.65	0.56	0.92	**0.76**	0.72	0.96
	MathQSA	0.63	0.53	0.83	0.51	0.68	0.88	0.52	0.67	0.88	0.57	0.63	0.87	**0.64**	0.67	0.89	0.53	0.65	0.87
	GK	0.54	0.59	0.52	0.53	0.68	0.68	0.48	0.71	0.66	**0.59**	0.68	0.66	0.48	0.70	0.65	0.44	0.74	0.72
LL-70B-Gen	SciQ	0.50	0.46	0.45	0.61	0.75	0.71	0.58	0.71	0.67	0.56	0.64	0.58	**0.59**	0.67	0.65	0.46	0.74	0.72
	MathQA	0.60	0.50	0.89	**0.85**	0.73	0.94	0.82	0.72	0.94	0.66	0.58	0.91	0.82	0.72	0.94	0.82	0.74	0.95
	MathQSA	0.61	0.53	0.78	**0.78**	0.77	0.89	0.76	0.76	0.89	0.61	0.60	0.81	**0.78**	0.76	0.89	0.64	0.69	0.85
	GK	0.59	0.60	0.36	**0.68**	0.76	0.58	0.66	0.72	0.53	**0.68**	0.73	0.55	0.66	0.75	0.54	0.58	0.73	0.60
ENSB-Gen	SciQ	0.41	0.44	0.53	**0.65**	0.74	0.75	0.60	0.71	0.73	0.48	0.59	0.62	0.59	0.71	0.72	0.52	0.79	0.81
	MathQA	0.70	0.54	0.94	0.78	0.55	0.94	0.80	0.59	0.95	0.74	0.58	0.95	0.81	0.59	0.95	**0.88**	0.85	0.99
	MathQSA	0.64	0.54	0.90	0.73	0.52	0.90	**0.75**	0.55	0.91	0.66	0.52	0.90	0.74	0.55	0.91	**0.75**	0.72	0.95
	GK	0.49	0.52	0.47	0.70	0.77	0.72	**0.71**	0.77	0.74	0.64	0.71	0.66	0.69	0.76	0.72	0.61	0.75	0.76
HA-Test	SciQ	0.30	0.50	0.47	0.37	0.50	0.47	0.37	0.50	0.47	0.30	0.50	0.47	0.37	0.50	0.47	**0.43**	0.74	0.72
	MathQA	**0.95**	0.50	0.97	0	0.50	0.97	0	0.50	0.97	**0.95**	0.50	0.97	0	0.50	0.97	0.86	0.80	0.99
	MathQSA	**0.93**	0.50	0.96	0	0.50	0.96	0	0.50	0.96	**0.93**	0.50	0.96	0	0.50	0.96	0.75	0.75	0.98
	GK	0.16	0.49	0.32	0.55	0.49	0.32	0.55	0.49	0.32	0.16	0.50	0.32	0.55	0.49	0.32	**0.59**	0.74	0.60

Table 29: Hallucination detection with BERT classifier results for various models trained on labels obtained from Exact-match based approach on Kaggle test sets. The best result highlighted in **bold**.

Test set	Sub-category	QR			RR			EC-EC			CC			QR-RR			q-r+Q-R+R-R		
		F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC
TL-1.1B-Gen	Arts and humanity	0.57	0.66	0.92	0.80	0.69	0.92	0.81	0.77	0.95	0.57	0.63	0.92	0.80	0.77	0.95	**0.85**	0.85	0.97
	Geography and travel	0.56	0.70	0.87	0.73	0.72	0.87	0.77	0.81	0.92	0.54	0.64	0.85	0.76	0.80	0.92	**0.79**	0.84	0.94
	Language and communication	0.51	0.65	0.93	0.77	0.65	0.92	0.76	0.74	0.95	0.47	0.59	0.91	0.75	0.74	0.95	**0.85**	0.82	0.97
	Sciences	0.55	0.67	0.90	0.77	0.71	0.90	0.78	0.79	0.93	0.49	0.60	0.88	0.77	0.79	0.93	**0.79**	0.83	0.95
	Social sciences	0.55	0.69	0.90	0.75	0.70	0.90	0.79	0.80	0.94	0.56	0.62	0.88	0.79	0.80	0.94	**0.82**	0.86	0.96
	Sports and recreation	0.55	0.65	0.91	0.81	0.73	0.92	0.82	0.79	0.95	0.57	0.59	0.90	0.81	0.80	0.95	**0.84**	0.84	0.96
PHI-3.5B-Gen	Arts and humanity	0.50	0.60	0.73	0.71	0.80	0.87	0.71	0.81	0.88	0.60	0.69	0.81	0.71	0.81	0.88	**0.80**	0.88	0.92
	Geography and travel	0.66	0.67	0.45	0.75	0.81	0.65	0.75	0.83	0.67	0.73	0.71	0.53	0.75	0.82	0.67	**0.82**	0.87	0.76
	Language and communication	0.49	0.62	0.65	0.62	0.76	0.78	0.61	0.77	0.79	0.49	0.63	0.66	0.61	0.76	0.78	**0.73**	0.82	0.83
	Sciences	0.66	0.62	0.47	0.79	0.85	0.74	0.79	0.85	0.75	0.70	0.70	0.56	0.79	0.84	0.74	**0.83**	0.89	0.81
	Social sciences	0.60	0.64	0.53	0.75	0.82	0.76	0.75	0.83	0.77	0.68	0.69	0.63	0.76	0.83	0.77	**0.80**	0.89	0.83
	Sports and recreation	0.45	0.60	0.67	0.67	0.75	0.82	0.66	0.77	0.84	0.58	0.65	0.76	0.66	0.77	0.84	**0.75**	0.83	0.87
LL-8B-Gen	Arts and humanity	0.61	0.65	0.54	0.82	0.89	0.86	0.82	0.89	0.86	0.75	0.83	0.78	0.82	0.89	0.85	**0.84**	0.90	0.88
	Geography and travel	0.79	0.70	0.35	0.88	0.93	0.74	0.88	0.93	0.74	0.88	0.88	0.68	0.88	0.93	0.75	**0.90**	0.92	0.78
	Language and communication	0.60	0.67	0.63	0.73	0.86	0.83	0.73	0.86	0.84	0.66	0.78	0.74	0.73	0.86	0.83	**0.76**	0.85	0.84
	Sciences	0.70	0.66	0.44	0.84	0.90	0.79	0.84	0.89	0.80	0.79	0.84	0.70	0.84	0.89	0.79	**0.85**	0.89	0.81
	Social sciences	0.74	0.72	0.50	0.85	0.92	0.80	0.86	0.92	0.81	0.82	0.86	0.71	0.86	0.93	0.81	**0.87**	0.92	0.83
	Sports and recreation	0.68	0.71	0.61	0.80	0.89	0.84	0.81	0.89	0.83	0.76	0.84	0.77	0.81	0.89	0.82	**0.83**	0.90	0.86
MST-7B-Gen	Arts and humanity	0.53	0.62	0.79	0.72	0.74	0.85	0.73	0.77	0.87	0.51	0.63	0.81	0.72	0.77	0.87	**0.82**	0.90	0.95
	Geography and travel	0.63	0.70	0.70	0.73	0.79	0.73	0.76	0.83	0.80	0.60	0.67	0.69	0.74	0.82	0.80	**0.80**	0.89	0.88
	Language and communication	0.49	0.61	0.76	0.66	0.74	0.83	0.66	0.76	0.85	0.44	0.62	0.77	0.64	0.75	0.84	**0.79**	0.87	0.93
	Sciences	0.59	0.66	0.71	0.71	0.78	0.78	0.74	0.82	0.83	0.54	0.62	0.70	0.74	0.82	0.83	**0.82**	0.90	0.92
	Social sciences	0.58	0.66	0.74	0.69	0.75	0.77	0.73	0.80	0.83	0.56	0.66	0.74	0.72	0.80	0.82	**0.82**	0.91	0.93
	Sports and recreation	0.54	0.64	0.78	0.71	0.75	0.83	0.72	0.79	0.87	0.51	0.64	0.79	0.71	0.79	0.87	**0.81**	0.89	0.94
GM-7B-Gen	Arts and humanity	0.53	0.64	0.73	0.58	0.69	0.79	0.59	0.70	0.79	0.59	0.69	0.79	0.58	0.71	0.79	**0.70**	0.77	0.86
	Geography and travel	0.60	0.58	0.45	0.65	0.68	0.55	0.64	0.69	0.56	0.64	0.65	0.52	0.64	0.70	0.56	**0.70**	0.76	0.66
	Language and communication	0.49	0.62	0.71	0.55	0.73	0.79	0.54	0.74	0.79	0.52	0.71	0.77	0.54	0.74	0.79	**0.67**	0.76	0.83
	Sciences	0.60	0.62	0.49	0.65	0.71	0.58	0.65	0.71	0.58	0.65	0.70	0.56	0.65	0.71	0.58	**0.71**	0.79	0.70
	Social sciences	0.59	0.63	0.55	0.65	0.70	0.65	0.65	0.70	0.65	0.64	0.67	0.61	0.65	0.70	0.65	**0.72**	0.79	0.75
	Sports and recreation	0.54	0.64	0.66	0.59	0.71	0.68	0.61	0.72	0.70	0.57	0.68	0.67	0.60	0.72	0.70	**0.68**	0.78	0.78
LL-70B-Gen	Arts and humanity	0.58	0.62	0.60	0.36	0.47	0.47	0.36	0.57	0.55	0.44	0.57	0.56	0.36	0.54	0.54	**0.68**	0.81	0.83
	Geography and travel	0.75	0.76	0.47	0.68	0.49	0.23	0.68	0.58	0.28	0.74	0.57	0.36	0.68	0.54	0.26	**0.76**	0.72	0.54
	Language and communication	0.56	0.61	0.54	0.43	0.42	0.37	0.43	0.46	0.39	0.50	0.51	0.48	0.43	0.46	0.39	**0.67**	0.72	0.68
	Sciences	0.69	0.69	0.52	0.59	0.46	0.26	0.59	0.59	0.33	0.66	0.61	0.44	0.59	0.51	0.29	**0.75**	0.80	0.69
	Social sciences	0.69	0.66	0.44	0.61	0.53	0.32	0.61	0.56	0.35	0.66	0.62	0.42	0.61	0.55	0.34	**0.77**	0.85	0.76
	Sports and recreation	0.66	0.68	0.64	0.44	0.50	0.40	0.44	0.58	0.47	0.58	0.62	0.62	0.44	0.57	0.45	**0.69**	0.78	0.77
ENSB-Gen	Arts and humanity	0.54	0.62	0.74	0.78	0.84	0.89	0.77	0.85	0.90	0.60	0.73	0.82	0.77	0.85	0.90	**0.83**	0.90	0.94
	Geography and travel	0.66	0.69	0.64	0.78	0.87	0.79	0.79	0.88	0.82	0.69	0.77	0.70	0.79	0.88	0.81	**0.82**	0.92	0.88
	Language and communication	0.51	0.63	0.75	0.68	0.79	0.85	0.67	0.81	0.87	0.50	0.67	0.77	0.66	0.80	0.86	**0.77**	0.86	0.91
	Sciences	0.61	0.67	0.66	0.79	0.87	0.84	0.80	0.88	0.86	0.63	0.75	0.73	0.79	0.88	0.85	**0.83**	0.92	0.91
	Social sciences	0.61	0.68	0.70	0.78	0.86	0.84	0.79	0.88	0.87	0.62	0.75	0.76	0.79	0.87	0.86	**0.84**	0.93	0.93
	Sports and recreation	0.56	0.68	0.77	0.75	0.82	0.87	0.74	0.84	0.89	0.63	0.74	0.82	0.74	0.84	0.89	**0.81**	0.88	0.93
HA-Test	Arts and humanity	0.52	0.50	0.66	0.17	0.50	0.66	0.17	0.50	0.66	0.52	0.50	0.66	0.52	0.50	0.66	**0.61**	0.81	0.90
	Geography and travel	0.23	0.50	0.40	0.45	0.50	0.40	0.45	0.50	0.40	0.23	0.50	0.40	0.23	0.50	0.40	**0.70**	0.85	0.82
	Language and communication	0.50	0.50	0.64	0.19	0.50	0.64	0.19	0.50	0.64	0.50	0.50	0.64	0.50	0.50	0.64	**0.65**	0.76	0.85
	Sciences	0.35	0.50	0.52	0.32	0.50	0.52	0.32	0.50	0.52	0.35	0.50	0.52	0.35	0.50	0.52	**0.64**	0.82	0.83
	Social sciences	0.35	0.50	0.51	0.32	0.50	0.51	0.32	0.50	0.51	0.35	0.50	0.51	0.35	0.50	0.51	**0.69**	0.87	0.89
	Sports and recreation	0.42	0.50	0.58	0.25	0.50	0.58	0.25	0.50	0.58	0.42	0.50	0.58	0.42	0.50	0.58	**0.62**	0.76	0.83

Table 30: Hallucination detection with BERT classifier results for various models trained on labels obtained from the LLM-based approach on the Jeopardy test sets.

Test set	Sub-category	QR			RR			EC-EC			CC			QR-RR			q-r+Q-R+R-R		
		F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC	F1	AUC	B-ACC
TL-1.1B-Gen	SciQ	0.45	0.60	0.80	**0.72**	0.68	0.82	**0.72**	0.73	0.87	0.54	0.59	0.81	**0.72**	0.73	0.86	0.66	0.75	0.89
	MathQA	0.84	0.67	1	0.93	0.47	0.99	0.95	0.62	1	0.89	0.65	1	0.97	0.62	1	**0.99**	0.74	1
	MathQSA	0.79	0.58	0.99	0.92	0.54	0.99	0.94	0.59	0.99	0.85	0.59	0.99	0.95	0.58	0.99	**0.97**	0.69	0.99
	GK	0.45	0.59	0.72	0.72	0.72	0.72	0.77	0.86	0.90	0.68	0.72	0.82	0.77	0.85	0.89	**0.80**	0.86	0.89
PHI-3.5B-Gen	SciQ	0.65	0.52	0.24	0.74	0.71	0.41	**0.75**	0.68	0.39	0.64	0.55	0.25	0.73	0.68	0.39	0.74	0.72	0.45
	MathQA	0.66	0.56	0.83	0.55	0.60	0.85	0.61	0.62	0.86	0.68	0.58	0.84	0.65	0.62	0.86	**0.80**	0.75	0.91
	MathQSA	0.60	0.54	0.77	0.65	0.65	0.83	0.65	0.66	0.83	0.62	0.55	0.77	0.67	0.65	0.82	**0.73**	0.72	0.86
	GK	0.72	0.46	0.15	0.77	0.64	0.40	0.80	0.64	0.34	0.70	0.53	0.18	0.80	0.63	0.39	**0.82**	0.64	0.37
LL-8B-Gen	SciQ	0.56	0.54	0.31	**0.75**	0.76	0.54	**0.75**	0.76	0.55	0.70	0.70	0.44	**0.75**	0.75	0.54	0.69	0.76	0.57
	MathQA	0.71	0.56	0.84	0.77	0.67	0.87	0.77	0.69	0.89	0.72	0.60	0.86	0.78	0.69	0.89	**0.81**	0.78	0.92
	MathQSA	0.61	0.54	0.73	0.72	0.71	0.81	0.73	0.74	0.84	0.64	0.60	0.76	0.73	0.73	0.84	**0.76**	0.77	0.86
	GK	0.70	0.58	0.21	0.81	0.72	0.40	0.81	0.72	0.40	**0.82**	0.68	0.38	0.81	0.72	0.37	0.81	0.68	0.41
MST-7B-Gen	SciQ	0.20	0.50	0.37	0.22	0.51	0.38	0.20	0.51	0.37	0.20	0.48	0.36	0.20	0.51	0.37	**0.25**	0.62	0.50
	MathQA	**0.90**	0.50	0.93	**0.90**	0.52	0.94	**0.90**	0.52	0.93	**0.90**	0.47	0.92	**0.90**	0.51	0.93	**0.90**	0.76	0.98
	MathQSA	**0.87**	0.52	0.92	**0.87**	0.51	0.91	**0.87**	0.52	0.92	**0.87**	0.47	0.90	**0.87**	0.49	0.91	**0.87**	0.68	0.95
	GK	0.16	0.53	0.33	0.19	0.54	0.34	0.16	0.60	0.37	0.16	0.38	0.26	0.16	0.52	0.32	**0.21**	0.66	0.52
GM-7B-Gen	SciQ	0.55	0.55	0.39	**0.68**	0.68	0.53	0.65	0.68	0.53	0.65	0.62	0.46	0.66	0.67	0.51	0.57	0.69	0.56
	MathQA	**0.77**	0.51	0.89	0.69	0.66	0.93	0.66	0.66	0.93	0.70	0.61	0.92	0.68	0.66	0.93	0.74	0.71	0.94
	MathQSA	**0.73**	0.54	0.83	0.64	0.72	0.91	0.61	0.73	0.91	0.65	0.64	0.88	0.64	0.71	0.91	0.65	0.71	0.90
	GK	**0.66**	0.60	0.40	**0.66**	0.65	0.54	0.64	0.66	0.53	**0.66**	0.61	0.46	0.64	0.64	0.50	0.62	0.68	0.58
LL-70B-Gen	SciQ	0.55	0.53	0.27	**0.79**	0.78	0.55	**0.79**	0.77	0.56	0.70	0.65	0.37	0.78	0.76	0.54	0.71	0.77	0.57
	MathQA	0.70	0.53	0.83	**0.85**	0.75	0.90	**0.85**	0.78	0.92	0.70	0.59	0.86	0.84	0.77	0.92	0.84	0.78	0.92
	MathQSA	0.62	0.52	0.73	0.80	0.81	0.88	**0.82**	0.84	0.90	0.65	0.61	0.77	0.80	0.82	0.89	0.81	0.81	0.89
	GK	0.72	0.62	0.23	0.87	0.84	0.60	**0.88**	0.83	0.56	0.82	0.76	0.39	0.86	0.84	0.54	0.85	0.84	0.55
ENSB-Gen	SciQ	0.53	0.50	0.42	**0.74**	0.80	0.68	0.73	0.79	0.70	0.58	0.60	0.49	0.73	0.79	0.70	0.70	0.82	0.76
	MathQA	0.72	0.57	0.91	0.79	0.70	0.93	0.81	0.72	0.94	0.76	0.62	0.92	0.82	0.72	0.94	**0.88**	0.84	0.97
	MathQSA	0.69	0.54	0.89	0.79	0.67	0.91	0.79	0.71	0.93	0.72	0.57	0.90	0.81	0.72	0.93	**0.85**	0.81	0.96
	GK	0.62	0.53	0.39	0.75	0.78	0.61	0.78	0.80	0.72	0.71	0.71	0.61	**0.79**	0.78	0.71	0.77	0.83	0.74
HA-Test	SciQ	0.30	0.50	0.47	0.37	0.50	0.47	0.37	0.50	0.47	0.30	0.50	0.47	0.37	0.50	0.47	**0.44**	0.71	0.70
	MathQA	**0.95**	0.50	0.97	0	0.50	0.97	0	0.50	0.97	**0.95**	0.50	0.97	0	0.50	0.97	0.88	0.82	0.99
	MathQSA	**0.93**	0.50	0.96	0	0.50	0.96	0	0.50	0.96	**0.93**	0.50	0.96	0	0.50	0.96	0.75	0.79	0.98
	GK	0.16	0.49	0.32	0.55	0.50	0.32	0.55	0.51	0.33	0.16	0.49	0.32	0.55	0.49	0.32	**0.63**	0.71	0.55

Table 31: Hallucination detection with BERT classifier results for various models trained on labels obtained from LLM-based approach on Kaggle test sets. The best result highlighted in **bold**.

L Datasheet for HaluCounterEval

L.1 Motivation

Q: For what purpose was the dataset created? (Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.)
A: This dataset is developed to facilitate research on reference-free hallucination detection in Large Language Models (LLMs). We observe a significant lack of suitable and sufficiently large benchmark datasets spanning multiple domains for reference-free hallucination detection. It will benefit the research community by enabling the development of hallucination detection pipelines and evaluating their robustness using this dataset.

Q: Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
A: The authors of this research paper created both the synthetic and human-annotated datasets.

Q: Who funded the creation of the dataset?
A: It is not disclosed to adhere to the anonymity policy.

Q: Any other comments? **A:** No.

L.2 Composition

Q: What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? (Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.)
A: Each instance in the dataset contains a question, an actual answer, responses generated by an LLM, and a label for each response indicating hallucination (1) or not hallucination (0).

Q: How many instances are there in total (of each type, if appropriate)?
A: The synthetic datasets contain 27,406 instances from the Jeopardy dataset and 56,328 instances from the Kaggle dataset. Refer to Tables 9 and 10 for more information. Meanwhile, the human-annotated test set consists of a total of 19,560 instances, out of which 9,560 are from the Jeopardy dataset and 10,000 are from the Kaggle dataset, for more details refer to Section 2.3.

Q: Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
A: The dataset consists of all instances derived from the raw data that we gathered and processed.

Q: Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information but might include, e.g., redacted text.
A: No.

Q: Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.)
A: No

Q: Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
A: Yes. Refer to Appendix Section B for an explanation. The split information is presented in Tables 9 and 10.

Q: Are there any errors, sources of noise, or redundancies in the dataset?
A: We perform rule-based filtration to remove noisy samples present in the dataset; however, it is not feasible to manually inspect all data instances.

Q: Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? (If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.)
A: The dataset is self-contained and can be downloaded, used, adapted, and redistributed without restrictions.

Q: Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
A: No, as all samples in the dataset are publicly available.

Q: Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
A: No.

Q: Does the dataset relate to people? (If not, you may skip the remaining questions in this section.)
A: No.

Q: Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
A: No.

Q: Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
A: No.

Q: Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
A: No.

Q: Any other comments?
A: No.

L.3 Collection process

Q: How was the data associated with each instance acquired? (Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.)
A: The data is obtained from Jeopardy (Jeopardy) and various Kaggle websites.

Q: What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? (How were these mechanisms or procedures validated?)
A: We manually downloaded the data.

Q: If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
A: The dataset is not sampled from a larger corpus.

Q: Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?
A: The dataset was collected from open-source websites, and the dataset is publicly available at https://huggingface.co/datasets/ashokurlana/HalluCounterEval.

Q: Over what timeframe was the data collected? (Does this timeframe match the creation timeframe of the data associated with the instances (e.g., a recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.)
A: The data was collected in late 2024.

Q: Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
A: No.

Q: Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
A: The dataset was obtained by downloading it from open-source websites. See Section 2 for more details.

Q: Were the individuals in question notified about the data collection? (If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.)
A: No. All datasets used to create HaluCounterEval are open source.

Q: Did the individuals in question consent to the collection and use of their data? (If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.)
A: No. All the datasets present in the HaluCounterEval are open-source.

Q: If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? (If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).)
A: N/A.

Q: Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? (If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.)
A: No.

Q: Any other comments?
A: No.

L.4 Preprocessing, cleaning, labeling

Q: Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? (If so, please provide a description. If not, you may skip the remainder of the questions in this section.)
A: Yes, detailed in Section 2.

Q: Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? (If so, please provide a link or other access point to the "raw" data.)
A: The "raw" data is saved and we plan to release it shortly.

Q: Is the software used to preprocess/clean/label the instances available? (If so, please provide a link or other access point.)
A: Yes, in the GitHub repository footnoted in the main content.

Q: Any other comments?
A: No.

L.5 Uses

Q: Has the dataset been used for any tasks already? (If so, please provide a description.)
A: We have used the dataset for training and testing purposes to perform reference-free hallucination detection. For more details, please refer to Section 4.

Q: Is there a repository that links to any or all papers or systems that use the dataset? (If so, please provide a link or other access point.)
A: No.

Q: What (other) tasks could the dataset be used for?
A: The dataset can be utilized for a wide range of NLP tasks concerning factual question-answering, and hallucination mitigation.

Q: Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? (For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?)
A: Yes, we applied rule-based filtration to remove noisy samples from the raw dataset, as detailed in Appendix Section A.

Q: Are there tasks for which the dataset should not be used? (If so, please provide a description.)
A: Our dataset may include misleading responses,

as the sample responses are sourced from various large language models (LLMs). Therefore, it should not be used for any purposes that could result in discrimination or harm.

Q: Any other comments?
A: No.

L.6 Distribution

Q: Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? (If so, please provide a description.)
A: Yes, the data will be free to the public to download, use, modify, and re-distribute.

Q: How will the dataset be distributed (e.g., tarball on the website, API, GitHub)? (Does the dataset have a digital object identifier (DOI)?)
A: The dataset is hosted at https://huggingface.co/datasets/ashokurlana/HalluCounterEval.

Q: When will the dataset be distributed?
A: The dataset is publicly hosted at https://huggingface.co/datasets/ashokurlana/HalluCounterEval.

Q: Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? (If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.)
A: Yes, the dataset is distributed under the CC BY 4.0 license.

Q: Have any third parties imposed IP-based or other restrictions on the data associated with the instances? (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.).
A: The datasets used in this paper are open-sourced, such that there are no restrictions associated with the data.

Q: Do any export controls or other regulatory restrictions apply to the dataset or individual instances? (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.)
A: No.

Q: Any other comments?
A: No.

L.7 Maintenance

Q: Who is supporting/hosting/maintaining the dataset?
A: Authors of this paper.

Q: How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
A: Via email or issues in the Hugging Face or GitHub repositories.

Q: Is there an erratum? (If so, please provide a link or other access point.)
A: No.

Q: Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? (If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?)
A: Currently there is no plan to update the dataset.

Q: If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? (If so, please describe these limits and explain how they will be enforced.)
A: No.

Q: Will older versions of the dataset continue to be supported/hosted/maintained? (If so, please describe how. If not, please describe how its obsolescence will be communicated to users.)
A: There is no older version of the dataset.

Q: If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? (If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.)

A: Yes, they can freely extend this dataset by downloading it from GitHub.

Q: Any other comments?
A: No.