The Illusion of Progress: Re-evaluating Hallucination Detection in LLMs

Denis Janiak[1] **Jakub Binkowski**[1] **Albert Sawczyn**[1] **Bogdan Gabrys**[2]
Ravid Shwartz-Ziv[3] **Tomasz Kajdanowicz**[1]
[1]Wroclaw University of Science and Technology
[2]University of Technology Sydney
[3]New York University

Abstract

Large language models (LLMs) have revolutionized natural language processing, yet their tendency to hallucinate poses serious challenges for reliable deployment. Despite numerous hallucination detection methods, their evaluations often rely on ROUGE, a metric based on lexical overlap that misaligns with human judgments. Through comprehensive human studies, we demonstrate that while ROUGE exhibits high recall, its extremely low precision leads to misleading performance estimates. In fact, several established detection methods show performance drops of up to 45.9% when assessed using human-aligned metrics like LLM-as-Judge. Moreover, our analysis reveals that simple heuristics based on response length can rival complex detection techniques, exposing a fundamental flaw in current evaluation practices. We argue that adopting semantically aware and robust evaluation frameworks is essential to accurately gauge the true performance of hallucination detection methods, ultimately ensuring the trustworthiness of LLM outputs.

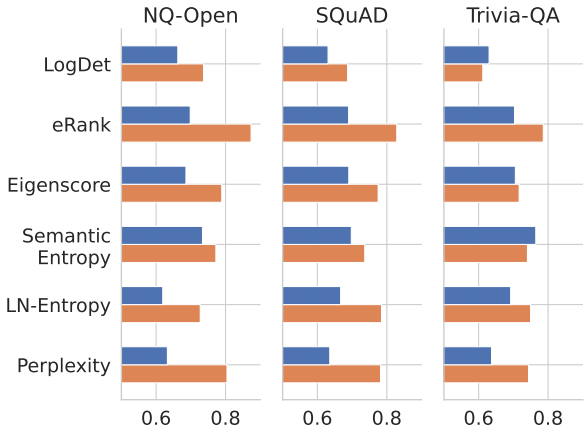

Figure 1: **ROUGE-based evaluation fails to reliably capture true hallucination detection capabilities.** Hallucination detection performance (AUROC) comparison of **ROUGE-L** and **LLM-as-Judge** evaluation across three datasets. Many methods show significant evaluation discrepancies.

1 Introduction

Large language models (LLMs) have transformed natural language processing, but their tendency to hallucinate—generating fluent yet factually incorrect outputs—poses a **critical challenge** for real-world applications (Huang et al., 2025). As LLMs are increasingly deployed in high-stakes scenarios, unsupervised hallucination detection has emerged as a promising solution, offering scalable evaluation without the generalization limitations of a supervised approach and costly annotation process (Su et al., 2024). A growing body of work has explored this direction (Chen et al., 2024; Farquhar et al., 2024; Du et al., 2024; Nikitin et al., 2024; Qiu and Miikkulainen, 2024; Duan et al., 2024; Nguyen et al., 2025), often relying on ROUGE as the primary correctness metric. ROUGE, originally developed to assess summary quality based on lexical overlap (Lin, 2004), is used to approximate factual consistency by applying threshold-based heuristics: responses with low ROUGE overlap to reference answers are often labeled as hallucinated. However, the suitability of ROUGE for assessing the factual accuracy of Question Answering (QA) responses, specifically in identifying hallucinations, has been largely assumed rather than rigorously validated. This assumption is especially critical in QA, where short, entity-centric answers are thought to make ROUGE suitable. Our findings show that even in these contexts, ROUGE can be misleading.

While prior critiques of ROUGE focus on its limitations in capturing fluency or adequacy in long-form summarization or dialogues (Honovich et al., 2022; Dziri et al., 2022; Zhong et al., 2022). In contrast, this paper presents a **systematic, large-scale empirical investigation** specifically evaluating ROUGE's efficacy in the context of QA hallucination detection. Our analysis goes beyond

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34717–34734
November 4-9, 2025

general critiques by quantitatively demonstrating ROUGE's key shortcomings—such as its susceptibility to response length—and how these issues can inflate the reported performance of hallucination detection methods. Furthermore, while ROUGE serves as our primary case study due to its ubiquity, we also demonstrate that other commonly used metrics share similar vulnerabilities, highlighting a broader deficiency in current evaluation practices.

To establish a human-aligned benchmark, we collect human judgments of factual correctness and compare metric outputs against these gold labels. We find that ROUGE exhibits **alarmingly low precision** in identifying actual factual errors. In contrast, an LLM-as-Judge approach (Zheng et al., 2023a) aligns far more closely with human assessments. Based on these insights, we re-evaluate existing detection methods under both ROUGE and human-aligned criteria, revealing dramatic performance drops (up to 45.9% for Perplexity and 30.4% for Eigenscore) when moving from ROUGE to LLM-as-Judge evaluation (see Figure 1).

Finally, we uncover a surprising baseline: simple length-based heuristics can match or exceed the performance of sophisticated detectors like Semantic Entropy. Through controlled causal experiments manipulating verbosity and input ambiguity, we demonstrate that longer or more ambiguous responses are more prone to hallucination and that metrics like ROUGE can be easily manipulated through trivial repetition, even when factual content remains unchanged. Our findings expose a **widespread overestimation** of current methods and underscore the urgent need for more reliable, human-aligned evaluation metrics in QA hallucination detection. Our contributions are as follows:

1. A human evaluation study validating LLM-as-Judge as a reliable metric for factual correctness, while showing that ROUGE and other n-gram or semantic metrics are poorly aligned with human judgments.
2. A systematic re-evaluation of existing hallucination detection methods, revealing that their reported effectiveness is often overstated when measured with ROUGE or similar metrics, which can hide critical flaws.
3. Identification of response length as a strong hallucination indicator, with simple length-based heuristics often matching or surpassing the performance of more sophisticated methods.

2 Related Work

Hallucination Detection Methods Recent research has shown that hallucinations in LLMs are inevitable (Xu et al., 2024), spurring work on two main detection paradigms: supervised and unsupervised. *Supervised* methods usually employ probing classifiers trained on labeled hidden states to detect hallucinations (Azaria and Mitchell, 2023; Orgad et al., 2024; Arteaga et al., 2024). While effective, they depend on costly human annotations and often fail to generalize across domains. *Unsupervised* methods detect hallucinations by estimating uncertainty directly—token-level confidence from single generations (Ren et al., 2023), sequence-level variance across multiple samples (Malinin and Gales, 2021; Farquhar et al., 2024), or hidden-state pattern analysis (Chen et al., 2024; Sriramanan et al., 2024a). While these methods show strong performance on standard benchmarks, our analysis reveals that simpler length-based baselines can achieve comparable results—echoing prior findings that simple baselines remain surprisingly competitive and underscoring the need for rigorous head-to-head comparisons (Fadeeva et al., 2023).

Evaluation Metrics and Their Limitations Traditional n-gram overlap measures such as ROUGE (Lin, 2004) remain popular for detecting hallucinations, despite their inability to reliably assess factual consistency (Honovich et al., 2022). Recent studies have further highlighted these limitations, particularly in multilingual settings where lexical overlap proves unreliable compared to NLI-based approaches (Kang et al., 2024). Even ROUGE-L, which tracks the longest common subsequence, often misses errors that leave surface overlap intact. To overcome these shortcomings, a family of embedding-based metrics — BERTScore (Zhang et al., 2020), UniEval (Zhong et al., 2022), AlignScore (Zha et al., 2023), and related approaches — has been proposed to capture deeper semantic similarity. However, embedding-based similarity does not always align with human assessments of factual correctness. By contrast, LLM-as-Judge methods (Zheng et al., 2023a) have shown strong agreement with human judgments in QA tasks (Thakur et al., 2025), offering a more reliable alternative. Our study builds on these insights by exposing the blind spots of ROUGE and other metrics, and validating LLM-as-Judge as a more faithful framework for factual evaluation.

3 Experimental Setup

3.1 Overview

Our experimental design aims to investigate both the shortcomings of current evaluation methods and the effectiveness of simpler alternatives.

3.2 Datasets and Models

For our experiments, we use three established QA datasets, each with distinct characteristics:

- **NQ-Open** (Kwiatkowski et al., 2019): Contains 3,610 question-answer pairs drawn from real Google search queries, representing natural information-seeking behavior
- **TriviaQA** (Joshi et al., 2017): A subset of 3,842 examples from the validation set, featuring trivia questions that often require specific factual knowledge
- **SQuAD** (Rajpurkar et al., 2018): 4,150 examples from the validation set (rc.nocontext), characterized by longer, more complex questions and answers

NQ-Open and TriviaQA primarily feature shorter questions and answers, whereas SQuADv2 contains longer inputs, making it suitable for evaluating our method in more complex contexts.

We generated answers using two open-source LLMs: LLAMA3.1-8B-INSTRUCT[1] (Grattafiori, 2024) and MISTRAL-7B-INSTRUCT-V0.3[2] (Jiang et al., 2023). For simplicity, we refer to these models as LLAMA and MISTRAL in our plots and tables.

3.3 Hallucination Detection Baselines

We compare our approach against established baselines that fall into two categories. **Uncertainty-based methods** estimate model confidence, including `Perplexity` (Ren et al., 2023), Length-Normalized Entropy (`LN-Entropy`) (Malinin and Gales, 2021), and Semantic Entropy (`SemEntropy`) (Farquhar et al., 2024), which use multiple generations to capture sequence-level uncertainty. **Consistency-based methods** analyze internal representations. `EigenScore` (Chen et al., 2024) computes generation consistency via eigenvalue spectra, while `LogDet` (Sriramanan et al., 2024a) measures covariance structure from single generations. We also evaluate Effective Rank (`eRank`) (Roy and Vetterli, 2007; Garrido et al., 2023), an intrinsic dimensionality measure we adapt as a novel hallucination indicator (see Appendix F.1).

[1] hf.co/meta-llama/Llama-3.1-8B-Instruct
[2] hf.co/mistralai/Mistral-7B-Instruct-v0.3

3.4 Ground Truth Labels

To obtain reliable ground truth labels for evaluating the correctness of generated responses, we utilize two complementary approaches:

LLM-as-Judge leverages GPT-4o-Mini (et al., 2024) for semantic assessment, following the methodology outlined in (Zheng et al., 2023b) and using a prompt adapted from (Orgad et al., 2025). This approach classifies generated responses into three categories: "correct," "incorrect," or "refuse" (with "refuse" being treated as a hallucination). By focusing on semantic equivalence and factual accuracy, this method goes beyond surface-level comparisons and exhibits strong alignment with human judgments (Thakur et al., 2025).

ROUGE-L F1 Score (Lin, 2004) measures the longest common subsequence between the generated response and the ground truth. Consistent with prior work (Farquhar et al., 2024), we apply a threshold of 0.3 for this metric. Including ROUGE-L allows us to compare our findings with existing literature and highlight the limitations of relying solely on lexical overlap for evaluating factual correctness. It helps to quantify the discrepancy between semantic understanding (assessed by the LLM judge) and simple word matching.

3.5 Evaluation Metrics

We employ Area Under the Receiver Operating Characteristic curve (AUROC) and Area Under the Precision-Recall curve (PR-AUC) as our primary evaluation metrics. AUROC assesses the ability of a hallucination detection method to correctly rank positive and negative instances (hallucinations vs. non-hallucinations). PR-AUC is particularly valuable when dealing with imbalanced datasets, which is often the case in hallucination detection, where non-hallucinated responses might be more frequent. Both metrics offer a threshold-independent evaluation of the ranking performance (Lin et al., 2023).

3.6 Implementation Details

We utilize pretrained model weights from the Hugging Face Transformers (Wolf et al., 2020) without any additional fine-tuning. Following (Farquhar et al., 2024), we generate 10 samples ($n = 10$) using temperature 1.0 for uncertainty estimation. Additionally, we generate one "best answer" sam-

ple with a temperature of 0.1 to serve as the best-generation estimate for performance evaluation.

The models are evaluated in both zero-shot and few-shot ($k = 5$) settings:

- **Zero-shot**: Models rely solely on their pre-existing knowledge, testing base capabilities
- **Few-shot**: Models receive five carefully selected examples demonstrating the expected answer formats

Both settings use a standardized prompt designed to elicit concise answers. The specific prompt, adapted from (Kossen et al., 2024), can be found in Appendix D. We report results for a single run unless specified otherwise.

4 Human Evaluation: The Gold Standard

Before analyzing the technical problems of hallucination detection methods, we first establish that commonly used evaluation metrics—specifically ROUGE—are poorly aligned with human judgments of factual correctness (Honovich et al., 2022; Kang et al., 2024). In contrast, an evaluation method based on LLM-as-Judge demonstrates much closer agreement with human assessments (Thakur et al., 2025). To illustrate this, we conducted a comprehensive human evaluation study.

Study Design We randomly selected 200 question–answer pairs from the Mistral answers on the NQ-Open dataset, ensuring a balanced representation of cases where ROUGE and LLM-as-Judge yield conflicting hallucination assessments. Each answer was independently assessed by three annotators using standardized guidelines from (Thakur et al., 2025), classifying responses as *correct*, *incorrect*, or *refuse* (we then classify model refusal as incorrect). The high inter-annotator agreement (Cohen's Kappa = 0.799) confirms the reliability of human judgments.

Key Findings Our results reveal a significant performance gap between LLM-as-Judge and ROUGE when benchmarked against human consensus. While ROUGE demonstrates high recall, it suffers from low precision, flagging many non-hallucinated content as errors. LLM-as-Judge achieves significantly higher precision, aligning more closely with human assessments, as shown in Table 1.

Implications Our findings underscore that ROUGE is a poor proxy for human judgment in evaluating hallucination detection. Despite its high precision, ROUGE fails to capture many critical errors, resulting in a significant misalignment with human assessments of factual correctness. In contrast, LLM-as-Judge exhibits strong agreement with human evaluations—achieving both high precision and recall—which motivates its adoption as a more robust, semantically aware evaluation method throughout this work.

Table 1: **LLM-as-Judge provides superior alignment with human judgment.** Comparison of ROUGE (with standard 0.3 threshold) and LLM-as-Judge against human labels.

Method	Precision	Recall	F1-Score	Agreement
LLM-as-Judge	0.736	0.957	0.832	0.723
ROUGE	0.401	0.957	0.565	0.142

5 Re-evaluating Hallucination Detection Methods

5.1 Limitations of ROUGE for Factual Accuracy Assessment in QA

The predominant reliance on ROUGE for evaluating QA hallucination detection methods warrants careful scrutiny, as its core design for lexical overlap does not inherently capture factual correctness. Our in-depth analysis, presented in Appendix G, reveals several critical failure modes that systematically undermine ROUGE's utility for this task. Key limitations include: sensitivity to response length, inability to handle semantic equivalence and susceptibility to false lexical matches.

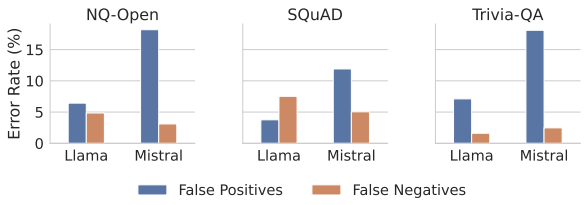

Figure 2: **ROUGE produces systematic errors across all evaluation settings.** Distribution of *False Negatives* and *False Positives* across different datasets and models highlights the inconsistency in ROUGE's evaluation.

These failure modes, illustrated with concrete examples and error distributions in Figure 2, highlight the potential for ROUGE to provide a misleading assessment of both LLM responses and the efficacy of hallucination detection techniques. This underscores the need for evaluation against more human-aligned metrics.

Table 2: **Detection methods show dramatic performance drops when evaluated against human-aligned metrics instead of ROUGE.** Performance comparison using AUROC scores for LLAMA and MISTRAL models across three datasets in zero-shot setting, where negative $\Delta\%$ values reveal ROUGE's overestimation of method effectiveness.

Model	Metric	NQ-Open			SQuAD			Trivia-QA			Mean		
		ROUGE	LLM	$\Delta\%$	ROUGE	LLM	$\Delta\%$	ROUGE	LLM	$\Delta\%$	ROUGE	LLM	$\Delta\%$
LLAMA	Perplexity	0.709	0.700	-1.2	0.703	0.687	-2.4	0.733	0.789	7.2	0.715	0.725	1.2
	LN-Entropy	0.521	0.605	13.9	0.558	0.611	8.7	0.563	0.636	11.5	0.547	0.617	11.4
	SE	0.778	0.742	-4.8	0.707	0.705	-0.2	0.769	0.832	7.6	0.751	0.760	0.9
	Eigenscore	0.816	0.686	-19.0	0.720	0.638	-12.7	0.752	0.734	-2.5	0.763	0.686	-11.4
	eRank	0.825	0.632	-30.6	0.754	0.621	-21.4	0.717	0.660	-8.6	0.765	0.638	-20.2
	LogDet	0.511	0.515	0.7	0.521	0.536	2.7	0.604	0.509	-18.6	0.545	0.520	-5.1
MISTRAL	Perplexity	0.852	0.584	-45.9	0.516	0.500	-3.2	0.843	0.627	-34.4	0.737	0.570	-27.8
	LN-Entropy	0.718	0.645	-11.3	0.734	0.657	-11.7	0.586	0.596	1.8	0.679	0.633	-7.1
	SE	0.836	0.729	-14.7	0.784	0.701	-11.9	0.726	0.707	-2.6	0.782	0.712	-9.7
	Eigenscore	0.873	0.669	-30.4	0.803	0.648	-24.0	0.775	0.652	-18.9	0.817	0.656	-24.4
	eRank	0.925	0.678	-36.4	0.518	0.511	-1.3	0.851	0.645	-31.9	0.765	0.611	-23.2
	LogDet	0.628	0.508	-23.6	0.562	0.518	-8.5	0.843	0.606	-39.2	0.678	0.544	-23.8

5.2 Quantifying the Evaluation Gap: ROUGE vs. LLM-as-Judge

Given the outlined limitations of ROUGE, we re-evaluated existing unsupervised hallucination detection methods using LLM-as-Judge, which, as validated by our human study, offers a closer alignment with human judgments of factual correctness.

Main results As detailed in Table 2, hallucination detection methods that show promise under ROUGE often suffer a substantial performance drop when re-evaluated with LLM-as-Judge. For instance, `Perplexity` sees its AUROC score plummet by as much as **45.9%** for the MISTRAL model on NQ-Open. Similarly, `Eigenscore` performance erodes by **19.0%** and **30.4%** for LLAMA and MISTRAL, respectively, on the same dataset. Even `eRank`, which posts impressive ROUGE-based scores, experiences a sharp decline of **30.6%** and **36.4%** under the LLM-as-Judge paradigm. Moreover, when evaluated using PR-AUC, we observe even larger performance discrepancies across all methods (see Tables 12 and 16 in the Appendix H.2); this amplifies the impact of class imbalance in the QA setup, as further evidenced by the low QA accuracies reported in Table 13.

Correlation This systematic discrepancy, visually underscored by the scatter plot in Figure 3, points to a fundamental inadequacy in ROUGE's ability to reflect true hallucination detection performance. The moderate Pearson correlation coefficient ($r = 0.55$) between the AUROC scores derived from these two evaluation approaches further suggests that methods may be inadvertently optimized for ROUGE's lexical overlap criteria rather than genuine factual correctness. Notably, among the evaluated detection techniques, only `Semantic Entropy` maintains a degree of relative stability, exhibiting more modest performance variations between the two evaluation frameworks.

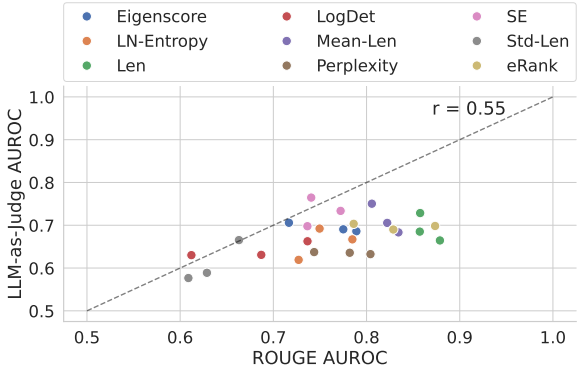

Figure 3: **ROUGE and human-aligned evaluations show weak correlation across detection methods.** Correlation between ROUGE and LLM-as-Judge AUROC scores for the MISTRAL model, with each point representing a metric's performance on specific dataset.

5.3 Impact of Few-Shot Examples on Evaluation Reliability

Our analysis of few-shot versus zero-shot settings reveals three key patterns in how examples affect evaluation stability (Table 3).

Improved Metric Stability Few-shot settings consistently yield more reliable evaluations across metrics. For LLAMA, the discrepancy between ROUGE and LLM-as-Judge narrows significantly with few-shot examples. For instance, `eRank` performance drop (for LLAMA) reduces from -16.7% in zero-shot to just -4.2% in few-shot settings.

This suggests that few-shot examples help standardize response formats with more consistent evaluation.

Table 3: **Few-shot examples reduce but don't eliminate evaluation biases.** Performance comparison showing relative differences between ROUGE and LLM-as-Judge in both settings.

Model	Metric	Few-Shot			Zero-Shot		
		ROUGE	LLM	Δ(%)	ROUGE	LLM	Δ(%)
LLAMA	Perplexity	0.783	0.784	0.0	0.715	0.725	1.5
	LN-Entropy	0.738	0.759	2.8	0.547	0.617	12.8
	SE	0.742	0.773	4.2	0.751	0.760	1.1
	Eigenscore	0.761	0.747	-1.9	0.763	0.686	-10.0
	eRank	0.707	0.678	-4.2	0.765	0.638	-16.7
MISTRAL	Perplexity	0.806	0.645	-20.0	0.747	0.579	-22.4
	LN-Entropy	0.754	0.659	-12.5	0.679	0.633	-6.8
	SE	0.750	0.732	-2.4	0.782	0.712	-8.9
	Eigenscore	0.760	0.694	-8.7	0.817	0.656	-19.7
	eRank	0.829	0.697	-15.9	0.773	0.612	-20.8

Model-Specific Effects The impact of few-shot examples varies notably between models. MISTRAL shows pronounced degradation in zero-shot settings, with performance drops up to 45.9% (Perplexity), while LLAMA maintains more consistent performance, with some metrics showing minimal degradation. This variation suggests that the architecture and pre-training may influence the effectiveness of few-shot calibration.

Metric Robustness Different metrics show varying levels of stability across settings. Semantic Entropy maintains the most consistent performance in both settings, while traditional metrics like Perplexity or LN-Entropy show higher sensitivity to setting changes.

Implications While few-shot examples generally improve evaluation reliability, the degree of improvement varies significantly across models and metrics. This suggests that robust hallucination detection systems should be validated under both conditions to ensure consistent performance across deployment scenarios. Of particular note is that few-shot examples reduce evaluation discrepancies by providing answer formats that more closely align with gold-standard responses. This indicates that some of the apparent improvements in few-shot settings may come from better format matching rather than enhanced factual assessment.

5.4 Evaluating beyond ROUGE

While ROUGE remains a widely adopted metric, its limitations underscore broader concerns about the reliability of lexical evaluation methods. To assess whether alternative metrics fare better, we extended our analysis to several others frequently used or proposed for text evaluation, including lexical metrics such as BLEU (Papineni et al., 2002) and semantic metrics such as BERTScore (Zhang et al., 2020), SummaC (Laban et al., 2022), and UniEval-fact (Zhong et al., 2022). We evaluated these metrics in both few-shot and zero-shot settings, benchmarking their outputs against our LLM-as-Judge labels, which show strong alignment with human judgments (see Table 1).

Table 4: **All metrics show limited alignment with human-like judgment, underscoring their shortcomings in capturing factual correctness.** Agreement of different correctness metrics with LLM-as-Judge labels in zero-shot settings. The results averaged across three QA datasets: NQ-Open, SQuAD, and TriviaQA.

Model	Metric	PRAUC	AUROC	F1	Precision	Recall
LLAMA	BERTScore	0.735	0.769	0.723	0.609	0.934
	BLEU	0.758	0.624	0.673	0.539	0.982
	ROUGE	0.891	0.878	0.812	0.728	0.926
	SummaC	0.826	0.782	0.725	0.616	0.944
	UniEval	0.828	0.830	0.762	0.739	0.804
MISTRAL	BERTScore	0.736	0.730	0.725	0.586	0.990
	BLEU	0.799	0.682	0.712	0.573	0.996
	ROUGE	0.865	0.825	0.757	0.629	0.971
	SummaC	0.836	0.778	0.758	0.648	0.950
	UniEval	0.720	0.706	0.693	0.674	0.746

Performance of Alternative Metrics As shown in Table 4, these alternative metrics also exhibit substantial shortcomings in reliably detecting hallucinations in QA tasks, particularly under zero-shot conditions. For example, BERTScore—despite leveraging contextual embeddings—often fails to outperform simpler lexical metrics in aligning with our LLM-as-Judge labels. BLEU and UniEval-fact similarly demonstrated limited effectiveness.

Implications These results suggest that the inadequacies of ROUGE are not isolated, but indicative of a broader challenge: current lexical and semantic metrics struggle to capture factual consistency, often favouring surface-level similarity or structural features such as length. Even when employing few-shot prompting (see Table 14 in the Appendix I), which can help with answer formatting, these metrics remain fundamentally constrained in their ability to assess factual correctness.

6 The Length Factor: A Hidden Signal in Hallucination Detection

Our analysis reveals a surprising and significant finding: response length alone serves as a powerful

signal for detecting hallucinations. This discovery challenges conventional wisdom about hallucination detection and raises fundamental questions about the complexity needed in detection methods. Our investigation demonstrates that: **(1)** Simple length statistics can serve as surprisingly effective hallucination detectors, often matching or exceeding more sophisticated methods; **(2)** The strong influence of length on current evaluation methods raises concerns about their ability to assess factual correctness independently of response verbosity; **(3)** This relationship may provide insights into the underlying mechanisms of how LLMs generate incorrect information.

6.1 Length Patterns in Hallucinated Responses

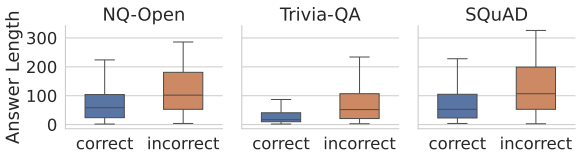

Figure 4: **Hallucinations have a distinct length signature in model outputs.** Distribution of answer lengths for MISTRAL in a few-shot settings with LLM-as-Judge labels, showing incorrect answers tend to be longer.

Analysis of response distributions using LLM-as-Judge labels reveals a striking pattern: hallucinated responses tend to be consistently longer and show greater length variance (Figure 4). While our primary experiments focus on short-form QA, this pattern also holds across other tasks. In particular, analysis of the HaluEval dataset—which includes summarization and dialogue tasks—confirms that length-based hallucination patterns generalize beyond QA (Figure 6 in Appendix J.1), suggesting a fundamental relationship between verbosity and hallucination.

This tendency toward longer responses likely reflects two key mechanisms. First, models attempt to maintain coherence while generating incorrect information, leading to additional context and elaboration. Second, initial errors often cascade into further mistakes, creating a "snowball effect" of increasing verbosity (Zhang et al., 2023)

6.2 Length Correlations with Existing Methods

To quantify this relationship, we examined correlations between response length and various hallucination detection metrics. Our analysis reveals

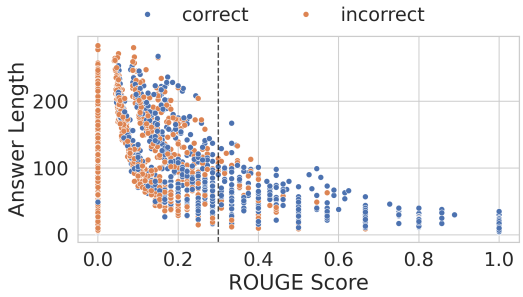

Figure 5: **ROUGE's bias against long responses undermines its reliability.** Distribution of answer length versus ROUGE score for MISTRAL in few-shot settings, revealing a strong correlation between length and ROUGE scores.

two critical findings. First, established methods show unexpectedly strong length correlations (see Table 5): `Eigenscore` and `eRank` exhibit particularly high correlations, suggesting these supposedly sophisticated methods may be primarily detecting length variations rather than semantic features. Second, ROUGE scores demonstrate a systematic length bias: As shown in Figure 5, responses exceeding 100 tokens consistently receive scores below the 0.3 threshold, regardless of factual accuracy. This aligns with prior observations of hallucination snowballing (Zhang et al., 2023), where LLMs compound initial errors with additional mistakes.

Table 5: **Sophisticated detection methods primarily capture length effects.** Pearson correlation coefficients between metrics and length, showing unexpectedly high values.

Method	Llama	Mistral
`LogDet`	-0.185	0.311
`Perplexity`	0.841	-0.423
`eRank`	0.763	0.803
`Eigenscore`	0.826	0.894
`LN-Entropy`	0.305	-0.753
`Semantic Entropy`	0.436	0.631

These correlations raise fundamental questions about whether current hallucination detection methods are truly capturing semantic features or simply leveraging length-based patterns.

6.3 Length as a Competitive Baseline

Given these strong correlations, we developed three simple length-based metrics: the raw length of a single generation (`Len`), the average length across multiple generations (`Mean-Len`), and the standard deviation of lengths across generations (`Std-Len`).

Evaluation results (Table 6) demonstrate that

these straightforward metrics achieve surprisingly competitive performance. The Mean-Len metric matches or outperforms sophisticated approaches like Eigenscore and LN-Entropy across multiple datasets. Response length variability proves to be a key indicator, with Std-Len showing particular effectiveness in identifying hallucinations. Perhaps most surprisingly, even the simple Len metric achieves competitive performance, challenging the fundamental need for complex detection methods.

Table 6: **Simple length-based metrics achieve competitive performance with sophisticated detection methods.** Hallucination detection performance (AUROC) compared across datasets and models using LLM-as-Judge since it shows better alignment with human judgements.

Model	Metric	NQ-Open	SQuAD	Trivia-QA	Mean
LLAMA	Perplexity	0.767	0.758	0.826	0.784
	LN-Entropy	0.732	0.717	0.829	0.759
	SE	0.730	0.741	0.849	0.773
	Eigenscore	0.744	0.733	0.762	0.747
	eRank	0.714	0.681	0.638	0.678
	Len	0.686	0.687	0.640	0.671
	Mean-Len	0.730	0.716	0.716	0.721
	Std-Len	0.727	0.721	0.806	0.751
MISTRAL	Perplexity	0.632	0.636	0.637	0.635
	LN-Entropy	0.619	0.667	0.692	0.659
	SE	0.734	0.698	0.765	0.732
	Eigenscore	0.686	0.691	0.706	0.694
	eRank	0.698	0.690	0.703	0.697
	Len	0.664	0.685	0.729	0.693
	Mean-Len	0.683	0.705	0.750	0.713
	Std-Len	0.577	0.589	0.665	0.610

6.4 Towards Causal Understanding of Length and Hallucination

To probe the relationship between response length and hallucination, we designed controlled experiments that manipulate factors such as verbosity, ambiguity, and complexity, with an additional analysis of adversarial versus non-adversarial questions reported for TriviaQA in Appendix 15. These follow-up experiments serve as valuable extensions, complementing the paper's primary contribution of exposing and challenging critical shortcomings in current hallucination evaluation, rather than aiming to definitively identify causal mechanisms (see Section N for further discussion).

The Repetition Experiment To demonstrate how ROUGE can be trivially misled by superficial changes in response length, even when factual content remains unchanged, we conducted a controlled experiment using systematic repetition. We modified model outputs by iteratively duplicating sentences while maintaining the same factual content. Results in Table 7 reveal a concerning trend: AUROC scores consistently improve with increased repetition, even though the information content remains unchanged. This experiment highlights a critical distinction: while verbose or repetitive responses may be inefficient, they aren't necessarily hallucinations if the core information is correct. However, current evaluation approaches, including both ROUGE and length-based metrics, fail to make this distinction.

Table 7: **ROUGE scores can be manipulated through simple repetition.** AUROC measurements for MISTRAL when repeating the same content multiple times.

Dataset	0	1	2	4
NQ-Open	0.852	0.935 (+9.7)	0.955 (+12.1)	0.964 (+13.1)
SQuAD	0.842	0.894 (+6.2)	0.909 (+8.0)	0.948 (+12.6)
Trivia-QA	0.843	0.901 (+6.9)	0.907 (+7.6)	0.919 (+9.0)

The Controlled Intervention (Isolating Length) To address the "correlation vs. causation" concern, we designed a controlled intervention experiment that isolates response length while holding the core factual content constant. Our hypothesis is that if response length is a causal factor in hallucination, then prompting the model to generate longer answers—even when the content is correct—should increase the likelihood of subtle hallucinations or factual drift. Prompts were designed to elicit answers that preserve the same underlying factual content, differing primarily in verbosity. For each question in the test set, we prompted the model in a few-shot setting to generate answers under four different prompt conditions: **Concise** (original), **Short** (to test sensitivity), **Regular** , and **Verbose** (see Appendix L.1 for full prompts). Table 8 reports the mean answer length, quartiles (Q1, Q2 [median], Q3), and accuracy (1 – hallucination rate).

Table 8: **Effect of response length on hallucination:** longer answers are more prone to factual drift.

Prompt	**Label**	**Mean**	**Q1**	**Q2**	**Q3**	**Accuracy**
Concise	correct	29.5	9	16	36.5	0.697
Concise	incorrect	65.7	21	46	89	0.697
Short	correct	28.7	10	16	37	0.698
Short	incorrect	60.3	20	43	84	0.698
Regular	correct	104.2	53	94	162	0.634
Regular	incorrect	140.7	82	160	193	0.634
Long	correct	196.9	181	196	212	0.604
Long	incorrect	197.1	182	197	213	0.604

Concise prompts yield much shorter and more accurate answers, and instructing brevity consistently lowers hallucination rates, supporting the view that longer outputs are more prone to subtle factual drift. The model follows brevity instructions more reliably when confident in its answer, which may explain the higher accuracy of short responses. These results indicate that longer responses may inadvertently introduce irrelevant or incorrect information, even when the underlying answer is known to the model. Thus, response length appears to be a contributing causal factor to hallucination—not merely correlated.

Deconstructing Hallucination Triggers To investigate which input factors most strongly induce hallucination, we introduce two types of controlled perturbations to the TriviaQA dataset while keeping the original questions intact. The **Ambiguous Input** variant rewrites questions to be indirect or under-specified, and the **Distractor Context** variant prepends a 2–3 sentence paragraph containing the correct answer alongside plausible but misleading details (full procedure in Appendix L.2). In Table 9, we report the mean answer length, quartiles (Q1, Q2 [median], Q3), and accuracy, defined as 1–hallucination rate for each condition.

Table 9: **Effect of input perturbations on hallucination:** Ambiguous questions substantially increase response length and hallucination, whereas distractor context has a smaller effect.

Modification	Label	Mean	Q1	Q2	Q3	Accuracy
Ambiguous	correct	48.4	13	28.5	66.0	0.564
Ambiguous	incorrect	83.6	25	62.5	133.2	0.564
Distractor	correct	33.2	10	16.0	41.0	0.671
Distractor	incorrect	76.9	18	51.0	126.0	0.671
Regular	correct	29.0	9	16.0	35.2	0.664
Regular	incorrect	66.3	18	44.0	90.5	0.664

When comparing across conditions, we observe that regular questions produce the shortest responses, while adding distractor context modestly increases the average answer length—especially in the upper tail—without harming accuracy. In contrast, ambiguous inputs trigger substantially longer answers and a pronounced drop in accuracy, indicating a higher hallucination rate. These findings suggest that input ambiguity is a more potent trigger for hallucination than misleading context: the model can effectively filter out distractors when the context is noisy but struggles when the question itself is underspecified.

7 Discussion

Our results reveal a clear misalignment between ROUGE and human judgments in identifying hallucinations. Despite the short, focused nature of QA answers—where n-gram overlap might seem sufficient—these metrics consistently reward fluent yet factually incorrect responses. While careful prompt engineering or dataset-specific post-processing may offer marginal improvements, these approaches often lack scalability and generalizability. As our experiments show, models frequently disregarded explicit brevity instructions (see Appendix D), making universally reliable prompts difficult to achieve.

Beyond ROUGE, evaluation with more sophisticated semantic metrics—BERTScore, BLEU, and UniEval-fact—against a strong LLM-based evaluator similarly revealed substantial disagreement, highlighting their limitations in capturing factual consistency. This is further underscored by our finding that simple response length can often be a more effective indicator of hallucinations than some sophisticated detection methods, questioning the current trajectory of detector development. Controlled interventions further show that longer responses, even when factually correct, are more prone to subtle factual drift, while input ambiguity exerts an even stronger effect, increasing both response length and hallucination rates. Together, these findings indicate that although verbosity can exacerbate hallucinations, the underlying input and reasoning dynamics are the primary determinants. Overall, our observations call for more robust, semantically aware evaluation paradigms that move beyond surface-level overlap metrics.

8 Conclusions

We demonstrate that prevailing overlap-based metrics systematically overestimate hallucination detection performance in QA, leading to illusory progress. LLM-as-Judge evaluation, validated against human judgments, exposes steep performance drops across all methods when judged for factual accuracy. Moreover, because simple signals like answer length can match complex detectors, we caution against over-engineering; effective baselines are essential for meaningful advancement.

Limitations

While our study provides valuable insights into the limitations of ROUGE for hallucination detection, several constraints should be acknowledged. First, our analysis primarily focuses on a subset of LLMs and datasets, which may not fully capture the diversity of models and tasks in the field. Consequently, the generalizability of our findings to other contexts remains to be validated. Second, although we propose response length as a simple yet effective heuristic for detecting hallucinations, this approach may not account for nuanced cases where longer responses are factually accurate. Additionally, our reliance on LLM-as-Judge as a benchmark for human-aligned evaluation, while more robust than ROUGE, is not without its biases and limitations. Future work should expand the scope of the analysis to include a broader range of models and datasets, beyond short question answering task. Finally, while our controlled experiments highlight the potential for the manipulation of ROUGE scores, further research is needed to develop metrics that are both robust against such manipulations and aligned with human judgment. The primary risk is that over-reliance on length-based heuristics and potentially biased human-aligned metrics could lead to inaccurate assessments of hallucination detection methods, resulting in the deployment of LLMs that may not reliably ensure factual accuracy in high-stakes applications.

Acknowledgments

This work was funded by the European Union under the Horizon Europe grant OMINO – Overcoming Multilevel INformation Overload (grant number 101086321, https://ominoproject.eu/). Views and opinions expressed are those of the authors alone and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the European Research Executive Agency can be held responsible for them. It was also co-financed with funds from the Polish Ministry of Education and Science under the programme entitled International Co-Financed Projects, grant no. 573977. This work was co-funded by the National Science Centre, Poland under CHIST-ERA Open & Reusable Research Data & Software (grant number 2022/04/Y/ST6/00183). Additionally, this work received funding from the Foundation for Polish Science under agreement FENG.02.02-IP.05-0314/23.

References

Gabriel Y. Arteaga, Thomas B. Schön, and Nicolas Pielawski. 2024. Hallucination Detection in LLMs: Fast and Memory-Efficient Finetuned Models. ArXiv:2409.02976 [cs].

Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. ArXiv:2304.13734 [cs].

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*.

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

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

OpenAI et al. 2024. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. LM-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, Singapore. Association for Computational Linguistics.

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

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. 2023. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. ArXiv:2210.02885 [cs].

Aaron et al. Grattafiori. 2024. The Llama 3 Herd of Models. ArXiv:2407.21783 [cs].

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating Factual Consistency Evaluation. ArXiv:2204.04991 [cs].

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. ArXiv:2310.06825 [cs].

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. ArXiv:1705.03551 [cs].

Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. 2024. Comparing hallucination detection metrics for multilingual generation.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs. ArXiv:2406.15927 [cs].

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:452–466. Place: Cambridge, MA Publisher: MIT Press.

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

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. ArXiv:2305.11747 [cs].

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. Publisher: arXiv Version Number: 3.

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

Hieu Nguyen, Zihao He, Shoumik Atul Gandre, Ujjwal Pasupulety, Sharanya Kumari Shivakumar, and Kristina Lerman. 2025. Smoothing out hallucinations: Mitigating llm hallucination with smoothed knowledge distillation.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. In *Advances in Neural Information Processing Systems*, volume 37, pages 8901–8929. Curran Associates, Inc.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. ArXiv:2410.02707.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Xin Qiu and Risto Miikkulainen. 2024. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *Advances in Neural Information Processing Systems*, volume 37, pages 134507–134533. Curran Associates, Inc.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. ArXiv:1806.03822 [cs].

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2023. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.

Olivier Roy and Martin Vetterli. 2007. The Effective Rank: a Measure of Effective Dimensionality.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024a. Llm-check: Investigating detection of hallucinations in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 34188–34216. Curran Associates, Inc.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024b. LLM-Check: Investigating Detection of Hallucinations in Large Language Models.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran HU, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. ArXiv:2403.06448 [cs].

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is Inevitable: An Innate Limitation of Large Language Models. ArXiv:2401.11817.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. How Language Model Hallucinations Can Snowball. ArXiv:2305.13534 [cs].

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, E. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023a. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *ArXiv*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Appendix

A Licenses and Computational Resources

A.1 Datasets, models license

The datasets and models used in this study are subject to specific licenses. NQ-Open, TriviaQA, and SQuAD are available under licenses that permit academic use. The LLAMA3.1-8B-INSTRUCT and MISTRAL-7B-INSTRUCT-V0.3 models are open-source and can be accessed under their respective licenses, which allow for research and non-commercial use.[3]

A.2 Hardware Specifications

We generated data using Nvidia A40 with 40GB VRAM. For the remaining computations, we used CPU.

B Human Involvement and Ethics

B.1 Annotator Recruitment and Consent

Participants were recruited through personal networks (friends and acquaintances) and participated voluntarily without financial compensation. They were informed of the study's purpose and data usage beforehand. Verbal consent was obtained, and no personally identifiable information was collected. Participants had the right to withdraw at any time.

B.2 Demographics

All annotators were residents of Poland. No systematic collection of age, gender, or other demographic information was conducted.

[3]For detailed license information, please refer to the respective dataset and model documentation.

C Use of AI Assistance

AI assistants, such as ChatGPT, were utilized in various aspects of the research, including coding, data analysis, and writing tasks. These tools helped automate repetitive tasks, generate initial drafts, and assist in exploring potential solutions. However, all AI-generated outputs were reviewed and refined by researchers to ensure accuracy and coherence.

D Main Prompts

We used the following prompt formats to elicit responses from the models:

- **QA (Zero-shot)**: Minimal prompt with no examples (Listing 1)
- **QA (Few-shot)**: Adapted from (Kossen et al., 2024), includes multiple QA examples (Listing 2)
- **LLM-as-Judge**: Evaluation prompt with correctness labels, adapted from (Orgad et al., 2024) (Listing 3)

Listing 1: Zero-shot prompt template

```
Answer the following question as briefly
    as possible.

Question: {question}
Answer:
```

Listing 2: QA (Few-shot) prompt template

```
Answer the following question as briefly
    as possible.

Here are several examples:

Question: What is the capital of France?
Answer: Paris

Question: Who wrote Romeo and Juliet?
Answer: William Shakespeare

Question: What is the boiling point of
    water in Celsius?
Answer: 100

Question: How many continents are there
    on Earth?
Answer: Seven

Question: What is the fastest land
    animal?
Answer: Cheetah

Question: {question}
Answer:
```

Listing 3: LLM-as-Judge prompt template

```
Answer the following question as briefly
    as possible.

Here are several examples:

Question: who is the young guitarist who
    played with Buddy Guy?
Ground Truth: Quinn Sullivan, Eric Gales
Model Answer: Ronnie Earl
Correctness: incorrect

Question: What is the name of the actor
    who plays Iron Man in the Marvel
    movies?
Ground Truth: Robert Downey Jr.
Model Answer: Robert Downey Jr. played
    the role of Tony Stark/Iron Man in
    the Marvel Cinematic Universe films.
Correctness: correct

Question: What is the capital of France?
Ground Truth: Paris
Model Answer: I don't have enough
    information to answer this question.
Correctness: refuse

Question: Who was the first person to
    walk on the moon?
Ground Truth: Neil Armstrong
Model Answer: I apologize, but I cannot
    provide an answer without verifying
    the historical facts.
Correctness: refuse

Question: {question}
Ground Truth: {gold}
Model Answer: {prediction}
Correctness:
```

E Additional Analysis of Human Evaluation

For the human evaluation component of our study (Section 4), we intentionally curated a dataset of instances where ROUGE and our LLM-as-Judge metric provided conflicting assessments regarding the presence of hallucinations. This targeted selection strategy was employed to enable a focused examination of ROUGE's specific failure modes. By concentrating on these points of disagreement, we aimed to gain deeper insights into the scenarios where ROUGE's reliance on lexical overlap demonstrably misaligns with human judgments of factual accuracy and overall response quality.

F Evaluation Metrics and Hallucination Detection

F.1 eRank

eRank leverages eigenvalue-based entropy estimation in hidden states:

$$\text{eRank} = \exp\left(-\sum_{k=1}^{m} p_k \log p_k\right) \quad (1)$$

where $p_k = \frac{\lambda_k}{\sum_{j=1}^{m} \lambda_j}$, and λ_k are the eigenvalues of the covariance matrix $\Sigma = Z^T Z$ computed on the hidden states Z.

We use Effective Rank (eRank) as a proxy for how "spread out" or "diverse" the final-layer hidden representations are; however, we can also use the hidden representation from the middle-layer. Intuitively, if the model's representation space collapses to fewer dimensions (i.e., low eRank), it may indicate that the model is relying on less context or ignoring crucial input signals—often manifesting as hallucinations. Conversely, a higher eRank suggests a richer, more nuanced encoding of the input, which typically correlates with more grounded and accurate responses. This approach builds on prior work (Sriramanan et al., 2024b) (LogDet), which computes the log-determinant of the covariance matrix.

While initial evaluations under ROUGE suggested some promise, we found that eRank did not consistently correlate with hallucination rates across all datasets and settings when assessed using human-aligned metrics. These 'negative results' illustrate how ROUGE's limitations can mislead method development.

G Understanding ROUGE's Failure Modes

Through detailed error analysis, we identify three critical limitations in ROUGE's evaluation approach: (1) sensitivity to response length, (2) inability to handle semantic equivalence, and (3) over-reliance on exact lexical matches. Our analysis reveals that these limitations lead to both false negatives—factually correct responses marked as incorrect—and false positives—incorrect responses receiving high scores. As shown in Figure 2, these errors occur frequently across different datasets and models.

G.1 Length-Based Penalties

Question: When was *Pride and Prejudice* written?
Prediction: "Pride and Prejudice was written by Jane Austen and published in 1813."
Gold Answer: "1813'

ROUGE systematically penalizes factually correct but verbose answers. In this example, despite providing accurate information with helpful context, the response receives a low score purely due to length mismatch. As shown in Figure 5, this bias affects longer responses regardless of their factual accuracy, with responses exceeding 100 tokens consistently scoring below our 0.3 threshold. Notably, this is **the most frequent type of error** ROUGE makes.

G.2 Semantic Equivalence Failures

Question: What is one element a topographic map shows?
Prediction: "Elevation"
Gold Answer: "Relief"

ROUGE fails to recognize semantic equivalence between different phrasings. Here, despite "elevation" and "relief" being contextually equivalent terms in topography, ROUGE assigns a lower score due to lexical mismatch. This limitation systematically undervalues responses that use valid alternative terminology.

G.3 False Lexical Matches

Question: "How many episodes of *Grey's Anatomy* season 14?"
Prediction: "23 episodes."
Gold Answer: "24 episodes."

ROUGE can assign high scores to factually incorrect answers that share a surface structure with the reference. Despite the critical numerical error, the response receives a relatively high score due to its match with surrounding words. This creates a dangerous bias toward structurally similar but factually incorrect answers.

H Quantitative Results

H.1 Metric Evaluation: AUROC

Tables 10 and 11 present comprehensive results comparing LLM-based and ROUGE-based evaluation metrics across three datasets: NQ-Open, SQuAD, and Trivia-QA. We evaluate nine different metrics using **AUROC** evaluation metric for both Llama and Mistral models under zero-shot and few-shot settings.

H.2 Metric Evaluation: PR-AUC

Tables 12 and 16 provide PR-AUC scores under the same conditions.

Table 10: Full comparison of LLM-based and ROUGE-based evaluation metrics across different datasets (NQ-Open, SQuAD, and Trivia-QA) for Llama and Mistral models in **zero-shot** setting using **AUROC** evaluation metric. The $\Delta\%$ columns show the relative percentage difference between LLM and ROUGE scores. Mean columns present the averaged scores across all datasets.

Model	Metric	NQ-Open			SQuAD			Trivia-QA			Mean		
		ROUGE	LLM	$\Delta\%$	ROUGE	LLM	$\Delta\%$	ROUGE	LLM	$\Delta\%$	ROUGE	LLM	$\Delta\%$
Llama	Perplexity	0.709	0.700	-1.2	0.703	0.687	-2.4	0.733	0.789	7.2	0.715	0.725	1.2
Llama	LN-Entropy	0.521	0.605	13.9	0.558	0.611	8.7	0.563	0.636	11.5	0.547	0.617	11.4
Llama	SE	0.778	0.742	-4.8	0.707	0.705	-0.2	0.769	0.832	7.6	0.751	0.760	0.9
Llama	Eigenscore	0.816	0.686	-19.0	0.720	0.638	-12.7	0.752	0.734	-2.5	0.763	0.686	-11.4
Llama	eRank	0.825	0.632	-30.6	0.754	0.621	-21.4	0.717	0.660	-8.6	0.765	0.638	-20.2
Llama	Len	0.834	0.616	-35.3	0.777	0.622	-24.9	0.760	0.691	-10.0	0.790	0.643	-23.4
Llama	LogDet	0.511	0.515	0.7	0.521	0.536	2.7	0.604	0.509	-18.6	0.545	0.520	-5.1
Llama	Mean-Len	0.825	0.654	-26.1	0.743	0.643	-15.7	0.771	0.743	-3.8	0.780	0.680	-15.2
Llama	Std-Len	0.711	0.644	-10.5	0.664	0.627	-6.0	0.759	0.754	-0.7	0.711	0.675	-5.7
Mistral	Perplexity	0.852	0.584	-45.9	0.516	0.500	-3.2	0.843	0.627	-34.4	0.737	0.570	-27.8
Mistral	LN-Entropy	0.718	0.645	-11.3	0.734	0.657	-11.7	0.586	0.596	1.8	0.679	0.633	-7.1
Mistral	SE	0.836	0.729	-14.7	0.784	0.701	-11.9	0.726	0.707	-2.6	0.782	0.712	-9.7
Mistral	Eigenscore	0.873	0.669	-30.4	0.803	0.648	-24.0	0.775	0.652	-18.9	0.817	0.656	-24.4
Mistral	eRank	0.925	0.678	-36.4	0.518	0.511	-1.3	0.851	0.645	-31.9	0.765	0.611	-23.2
Mistral	Len	0.934	0.634	-47.2	0.860	0.624	-37.8	0.929	0.673	-37.9	0.908	0.644	-41.0
Mistral	LogDet	0.628	0.508	-23.6	0.562	0.518	-8.5	0.843	0.606	-39.2	0.678	0.544	-23.8
Mistral	Mean-Len	0.890	0.643	-38.4	0.828	0.626	-32.2	0.875	0.667	-31.3	0.864	0.645	-34.0
Mistral	Std-Len	0.516	0.512	-0.7	0.540	0.505	-6.9	0.613	0.572	-7.2	0.556	0.530	-4.9

H.3 QA Accuracy Across Settings

Table 13 presents the accuracies on the QA datasets. These accuracies are computed by selecting the most likely answer at a low temperature setting and comparing it to labels derived from either ROUGE or LLM-as-Judge evaluations.

Table 13: Accuracies of different models, datasets, and prompts for the QA task.

				Accuracy	
Dataset	Model	Prompt	# Refused	ROUGE	LLM
NQ-Open	Llama	Few-Shot	692	28.1%	29.2%
NQ-Open	Llama	Zero-Shot	139	24.2%	43.2%
NQ-Open	Mistral	Few-Shot	117	20.9%	35.8%
NQ-Open	Mistral	Zero-Shot	72	7.8%	39.0%
SQuAD	Llama	Few-Shot	924	22.0%	18.3%
SQuAD	Llama	Zero-Shot	447	20.2%	25.0%
SQuAD	Mistral	Few-Shot	230	16.0%	22.6%
SQuAD	Mistral	Zero-Shot	116	5.8%	25.3%
Trivia-QA	Llama	Few-Shot	95	63.7%	69.4%
Trivia-QA	Llama	Zero-Shot	39	58.8%	71.1%
Trivia-QA	Mistral	Few-Shot	11	53.8%	69.7%
Trivia-QA	Mistral	Zero-Shot	2	29.0%	64.8%

I Ground Truth Labeling Metrics

To evaluate and compare automatic labeling strategies, we examined the agreement between various evaluation metrics and the LLM-as-Judge annotations (Table 14). This analysis provides insight into the reliability of proxy labeling methods for hallucination detection.

Table 14: **Few-shot evaluation metrics vs. LLM-as-Judge labels.** Average agreement across NQ-Open, SQuAD, and TriviaQA, showing how standard QA metrics align with LLM-based judgments.

Model	Metric	PRAUC	AUROC	F1	Precision	Recall
LLAMA	BERTScore	0.810	0.848	0.776	0.742	0.859
	BLEU	0.775	0.536	0.699	0.576	0.976
	ROUGE	0.935	0.921	0.883	0.866	0.906
	SummaC	0.850	0.776	0.760	0.653	0.977
	UniEval	0.943	0.933	0.862	0.868	0.868
MISTRAL	BERTScore	0.764	0.770	0.749	0.637	0.958
	BLEU	0.784	0.627	0.707	0.581	0.987
	ROUGE	0.903	0.878	0.820	0.738	0.932
	SummaC	0.855	0.795	0.758	0.657	0.957
	UniEval	0.813	0.801	0.754	0.751	0.778

J Answer Length Distribution

J.1 HaluEval

Figure 6 illustrates answer lengths across the HaluEval dataset (Li et al., 2023).

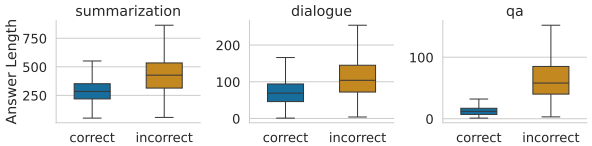

Figure 6: **Length-based hallucination patterns generalize across datasets.** Answer length distribution for HaluEval tasks, showing consistent patterns.

Table 11: Full comparison of LLM-based and ROUGE-based evaluation metrics across different datasets (NQ-Open, SQuAD, and Trivia-QA) for Llama and Mistral models in **few-shot** setting using **AUROC** evaluation metric. The $\Delta\%$ columns show the relative percentage difference between LLM and ROUGE scores. Mean columns present the averaged scores across all datasets.

Model	Metric	NQ-Open			SQuAD			Trivia-QA			Mean		
		ROUGE	LLM	$\Delta\%$	ROUGE	LLM	$\Delta\%$	ROUGE	LLM	$\Delta\%$	ROUGE	LLM	$\Delta\%$
Llama	Perplexity	0.814	0.767	-6.1	0.736	0.758	2.9	0.800	0.826	3.1	0.783	0.784	-0.0
Llama	LN-Entropy	0.753	0.732	-2.9	0.663	0.717	7.5	0.799	0.829	3.6	0.738	0.759	2.7
Llama	SE	0.738	0.730	-1.1	0.688	0.741	7.1	0.800	0.849	5.7	0.742	0.773	3.9
Llama	Eigenscore	0.813	0.744	-9.3	0.725	0.733	1.2	0.745	0.762	2.3	0.761	0.746	-1.9
Llama	eRank	0.794	0.714	-11.2	0.708	0.681	-4.0	0.620	0.638	2.8	0.707	0.678	-4.1
Llama	Len	0.761	0.686	-10.9	0.694	0.687	-1.0	0.620	0.640	3.1	0.692	0.671	-2.9
Llama	LogDet	0.729	0.690	-5.6	0.659	0.636	-3.7	0.590	0.618	4.5	0.659	0.648	-1.6
Llama	Mean-Len	0.799	0.730	-9.4	0.713	0.716	0.4	0.681	0.716	4.8	0.731	0.721	-1.4
Llama	Std-Len	0.777	0.727	-7.0	0.705	0.721	2.2	0.783	0.806	2.9	0.755	0.751	-0.6
Mistral	Perplexity	0.804	0.632	-27.1	0.782	0.636	-23.0	0.744	0.637	-16.7	0.777	0.635	-22.3
Mistral	LN-Entropy	0.727	0.619	-17.4	0.785	0.667	-17.7	0.750	0.692	-8.3	0.754	0.659	-14.5
Mistral	SE	0.772	0.734	-5.3	0.737	0.698	-5.6	0.741	0.765	3.1	0.750	0.732	-2.6
Mistral	Eigenscore	0.789	0.686	-15.0	0.775	0.691	-12.2	0.717	0.706	-1.5	0.760	0.694	-9.6
Mistral	eRank	0.874	0.698	-25.1	0.829	0.690	-20.1	0.786	0.703	-11.8	0.830	0.697	-19.0
Mistral	Len	0.879	0.664	-32.2	0.857	0.685	-25.1	0.858	0.729	-17.7	0.865	0.693	-25.0
Mistral	LogDet	0.737	0.663	-11.2	0.687	0.631	-8.9	0.612	0.630	2.9	0.679	0.641	-5.7
Mistral	Mean-Len	0.834	0.683	-22.1	0.822	0.705	-16.5	0.806	0.750	-7.4	0.821	0.713	-15.3
Mistral	Std-Len	0.609	0.577	-5.6	0.629	0.589	-6.8	0.663	0.665	0.3	0.634	0.610	-4.0

K Answer-Length Prompts

To investigate the effect of response length on hallucination, we prompted the model to generate answers at different verbosity levels. Four prompt variants were used: Concise, Short, Regular, and Verbose.

- **Concise** (as in the main paper):

```
Answer the following question as
briefly as possible.
```

- **Short** (to test sensitivity):

```
Answer the following question in few
 words.
```

- **Regular**:

```
Answer the following question.
```

- **Verbose**:

```
Answer the following question in a
detailed and comprehensive manner.
```

L Input Perturbation Prompts

To study hallucination triggers, we created two input perturbation variants: **Ambiguous Input** and **Distractor Context**. We subsampled 1,000 examples from the TriviaQA dataset and evaluated the few-shot Mistral model under three conditions: Ambiguous, Distractor, and the original (Regular) questions. The prompts used for automated question rewriting are provided below.

L.1 Ambiguous Input

Each question was automatically rewritten to be intentionally indirect or under-specified. For example:
Original: "What city is the capital of France?"
Rewritten: "What is the main administrative center of the French nation?"

Listing 4: Prompt used for rewriting to make more ambigous

```
You are given a factual question.
Your task is to rephrase it to make it
sound more ambiguous, indirect, or open-
ended, while still allowing someone
knowledgeable to infer and provide the
correct specific answer. Avoid adding or
 removing factual content - focus on
phrasing that introduces uncertainty or
generality. Output only the rephrased
question without any additional
explanation or commentary.

System prompt: "You are an expert at
rephrasing questions to make them more
challenging."
```

L.2 Distractor Context

Each question was prepended with a 2–3 sentence paragraph embedding the correct answer alongside plausible but incorrect details. For example:
Original: "Which country won the 2007 FIFA Women's World Cup?"
Rewritten: "The 2007 FIFA Women's World Cup

Table 12: Full comparison of LLM-based and ROUGE-based evaluation metrics across different datasets (NQ-Open, SQuAD, and Trivia-QA) for Llama and Mistral models in **zero-shot** setting using **PR-AUC** evaluation metric. The $\Delta\%$ columns show the relative percentage difference between LLM and ROUGE scores. Mean columns present the averaged scores across all datasets.

Model	Metric	NQ-Open			SQuAD			Trivia-QA			Mean		
		ROUGE	LLM	$\Delta\%$	ROUGE	LLM	$\Delta\%$	ROUGE	LLM	$\Delta\%$	ROUGE	LLM	$\Delta\%$
Llama	Perplexity	0.833	0.680	-22.4	0.863	0.823	-4.8	0.594	0.514	-15.6	0.763	0.672	-14.3
Llama	LN-Entropy	0.717	0.611	-17.4	0.793	0.773	-2.6	0.570	0.652	12.5	0.693	0.679	-2.5
Llama	SE	0.845	0.695	-21.5	0.864	0.829	-4.1	0.575	0.533	-7.9	0.761	0.686	-11.2
Llama	Eigenscore	0.850	0.670	-26.8	0.866	0.809	-7.1	0.565	0.574	1.6	0.760	0.684	-10.8
Llama	eRank	0.782	0.607	-28.9	0.820	0.783	-4.6	0.674	0.760	11.3	0.759	0.717	-7.4
Llama	Len	0.865	0.681	-27.2	0.885	0.820	-8.0	0.605	0.548	-10.4	0.785	0.683	-15.2
Llama	LogDet	0.852	0.659	-29.2	0.873	0.810	-7.8	0.602	0.562	-7.1	0.776	0.677	-14.7
Llama	Mean-Len	0.851	0.658	-29.3	0.870	0.808	-7.7	0.573	0.568	-0.9	0.765	0.678	-12.6
Llama	Std-Len	0.825	0.647	-27.6	0.846	0.802	-5.5	0.562	0.570	1.4	0.744	0.673	-10.6
Mistral	Perplexity	0.664	0.536	-23.8	0.951	0.754	-26.0	0.690	0.752	8.3	0.768	0.681	-13.8
Mistral	LN-Entropy	0.882	0.664	-32.8	0.920	0.790	-16.4	0.625	0.633	1.3	0.809	0.696	-16.0
Mistral	SE	0.956	0.725	-31.8	0.964	0.819	-17.7	0.808	0.510	-58.3	0.909	0.685	-35.9
Mistral	Eigenscore	0.957	0.698	-37.1	0.965	0.804	-20.0	0.818	0.544	-50.3	0.913	0.682	-35.8
Mistral	eRank	0.658	0.506	-30.0	0.955	0.755	-26.4	0.534	0.704	24.2	0.716	0.655	-10.7
Mistral	Len	0.964	0.682	-41.4	0.973	0.803	-21.0	0.849	0.536	-58.4	0.929	0.674	-40.3
Mistral	LogDet	0.964	0.699	-37.9	0.950	0.753	-26.1	0.847	0.550	-54.1	0.920	0.667	-39.4
Mistral	Mean-Len	0.958	0.671	-42.8	0.966	0.786	-22.9	0.833	0.548	-52.1	0.919	0.668	-39.3
Mistral	Std-Len	0.891	0.583	-52.8	0.889	0.724	-22.7	0.755	0.605	-24.8	0.845	0.637	-33.4

was marked by intense competition and surprising outcomes. Many expected Brazil, with their star player Marta, to dominate the tournament, while others believed Japan or Norway might prevail based on their strong qualifying performances. However, the tournament concluded with a different team emerging victorious. Which country won the 2007 FIFA Women's World Cup?"

Listing 5: Prompt used for generating distractor context

```
You are given a factual question.
Your task is to prepend a context
paragraph (2-3 sentences) that contains
plausible but incorrect information (
distractors) related to the topic of the
 question. Then, present the original
question below the context. Ensure the
distractor context is believable and
topically related but factually
incorrect or misleading. Output only the
 final result without any additional
explanation.

System prompt: "You are an expert at
rephrasing questions to make them more
challenging."
```

M Adversarial examples with TruthfulQA

We investigated the performance difference between adversarial and non-adversarial questions in the TriviaQA dataset. Surprisingly, our initial results indicate that adversarial examples are not more difficult for the model to answer; in fact, the model performs slightly better on adversarial questions.

We did, however, observe notable differences in answer lengths between hallucinated and correct responses. This is shown in Table 15, which presents statistics for the few-shot Mistral model used in Figure 4. Notably, across all three datasets examined in this study, the distribution of answer lengths was consistent between the Mistral and LLaMA models, as well as between zero-shot and few-shot settings.

Table 15: **Answer-length and accuracy statistics for adversarial vs. non-adversarial TriviaQA questions.** Mean answer length, quartiles (Q1, Q2 [median], Q3), and accuracy (1 – hallucination rate) for the few-shot Mistral model show that adversarial questions are not harder than regular ones and yield similar length distributions.

Prompt	Label	Mean	Q1	Q2	Q3	Accuracy
Adversarial	incorrect	143	87	141	197	0.47
Adversarial	correct	133.2	79	121	183	0.47
Non-Adversarial	incorrect	143.5	88	146	195	0.439
Non-Adversarial	correct	131	77	128	183.5	0.439

Additionally, we found several issues with the TriviaQA dataset itself. After manually reviewing a number of cases, we discovered that the LLM-as-Judge annotations were occasionally inaccurate. To

Table 16: Full comparison of LLM-based and ROUGE-based evaluation metrics across different datasets (NQ-Open, SQuAD, and Trivia-QA) for Llama and Mistral models in **few-shot** setting using **PR-AUC** evaluation metric. The $\Delta\%$ columns show the relative percentage difference between LLM and ROUGE scores. Mean columns present the averaged scores across all datasets.

Model	Metric	NQ-Open			SQuAD			Trivia-QA			Mean		
		ROUGE	LLM	$\Delta\%$	ROUGE	LLM	$\Delta\%$	ROUGE	LLM	$\Delta\%$	ROUGE	LLM	$\Delta\%$
Llama	Perplexity	0.844	0.824	-2.4	0.861	0.891	3.4	0.551	0.502	-9.8	0.752	0.739	-2.9
Llama	LN-Entropy	0.810	0.796	-1.8	0.828	0.874	5.3	0.525	0.522	-0.5	0.721	0.731	1.0
Llama	SE	0.814	0.802	-1.5	0.842	0.879	4.3	0.536	0.506	-6.1	0.731	0.729	-1.1
Llama	Eigenscore	0.829	0.802	-3.4	0.852	0.876	2.7	0.511	0.542	5.7	0.731	0.740	1.7
Llama	eRank	0.746	0.726	-2.8	0.711	0.762	6.8	0.679	0.737	7.9	0.712	0.742	4.0
Llama	Len	0.834	0.806	-3.5	0.856	0.884	3.1	0.522	0.571	8.7	0.737	0.754	2.8
Llama	LogDet	0.817	0.800	-2.1	0.859	0.882	2.6	0.526	0.582	9.6	0.734	0.755	3.4
Llama	Mean-Len	0.825	0.798	-3.4	0.852	0.878	2.9	0.509	0.553	7.9	0.729	0.743	2.5
Llama	Std-Len	0.820	0.794	-3.2	0.846	0.873	3.1	0.526	0.524	-0.3	0.731	0.730	-0.1
Mistral	Perplexity	0.506	0.520	2.7	0.624	0.673	7.4	0.740	0.778	4.9	0.623	0.657	5.0
Mistral	LN-Entropy	0.508	0.505	-0.6	0.587	0.615	4.5	0.759	0.825	8.0	0.618	0.648	4.0
Mistral	SE	0.872	0.754	-15.7	0.898	0.843	-6.5	0.609	0.538	-13.3	0.793	0.712	-11.8
Mistral	Eigenscore	0.873	0.738	-18.4	0.902	0.842	-7.2	0.598	0.567	-5.5	0.791	0.716	-10.4
Mistral	eRank	0.515	0.526	2.0	0.855	0.789	-8.4	0.606	0.736	17.8	0.659	0.684	3.8
Mistral	Len	0.897	0.735	-22.1	0.918	0.848	-8.3	0.687	0.530	-29.7	0.834	0.704	-20.0
Mistral	LogDet	0.895	0.734	-21.9	0.869	0.793	-9.5	0.673	0.561	-19.8	0.812	0.696	-17.1
Mistral	Mean-Len	0.879	0.734	-19.7	0.907	0.844	-7.5	0.629	0.548	-14.7	0.805	0.709	-14.0
Mistral	Std-Len	0.827	0.683	-20.9	0.873	0.808	-8.0	0.546	0.608	10.1	0.749	0.700	-6.3

further explore this, we conducted a study comparing the judgment of GPT-4o-mini with that of the larger GPT-4.1 model. Specifically, we calculated Cohen's Kappa score to measure the agreement between the two models in labeling answers as correct or incorrect based on a known reference answer.

The results from Table 17 show strong agreement across most datasets, except for TruthfulQA, which exhibited a notably lower Kappa score. Due to the unreliability of the results on this dataset, we decided to exclude TruthfulQA from further analysis.

Table 17: **Agreement between GPT-4.1 and GPT-4o-mini as judges.** Cohen's Kappa scores measuring inter-model agreement on answer correctness across four datasets, revealing strong alignment except for TruthfulQA, which shows notably lower reliability.

Dataset	NQOpen	SQuAD	TriviaQA	TruthfulQA
Cohen's Kappa	0.883	0.854	0.939	0.714

N Causality Discussion

We argue that while longer responses may correlate with the presence of hallucinations, response length itself is not a direct causal factor, but rather a consequence of underlying reasoning processes - longer answers can co-occur with various confounding factors. To illustrate that increased length does not inherently cause hallucination, we can consider Chain-of-Thought (CoT) reasoning. LLMs often generate longer sequences through CoT, effectively utilizing intermediate steps to enhance their computational expressiveness and reasoning depth. Far from inducing hallucination, this process has been widely demonstrated to reduce factual errors and improve the accuracy of question-answering abilities, even as it increases response length.