ON THE ROLE OF PROMPT MULTIPLICITY IN LLM HALLUCINATION EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are known to "hallucinate" by generating false or misleading outputs. Existing hallucination benchmarks often overlook prompt sensitivity, due to stable accuracy scores despite prompt variations. However, such stability can be misleading. In this work, we introduce *prompt multiplicity*—the multiplicity of individual hallucinations depending on the input prompt—and study its role in LLM hallucination benchmarks. We find severe multiplicity, with even more than 50% of responses changing between correct and incorrect answers simply based on the prompt for certain benchmarks, like Med-HALT. Prompt multiplicity also gives us the lens to distinguish between randomness in generation and consistent factual inaccuracies, providing a more nuanced understanding of LLM hallucinations and their real-world harms. By situating our discussion within existing hallucination taxonomies–supporting their quantification–and exploring its relationship with uncertainty in generation, we highlight how prompt multiplicity fills a critical gap in the literature on LLM hallucinations.

1 INTRODUCTION

Large language models (LLMs) have become widely adopted, excelling in numerous tasks across diverse domains (Guo et al., 2023; Kasneci et al., 2023; Naveed et al., 2023; Etsenake & Nagappan, 2024; Wang et al., 2024). Despite their growing use, LLMs suffer from a critical limitation: generation of false, factually incorrect, nonsensical or misleading outputs, studied under the umbrella of "hallucinations" (Ji et al., 2023; Zhang et al., 2023; Huang et al., 2023; Tonmoy et al., 2024).

The term "hallucinations" has evolved over the years, shifting from its positive use in computer vision (Baker & Kanade, 2000; Hsu et al., 2010) to its predominantly negative association in natural language processing (NLP) (Karpathy, 2015; Huang et al., 2023; Ji et al., 2023; Zhang et al., 2023). It is commonly defined as *'generated content that is nonsensical or unfaithful to the provided source content'* (Filippova, 2020; Maynez et al., 2020; Zhou et al., 2021; Ji et al., 2023).

With the growing interest in this field, several benchmarks have been developed to assess the risk of hallucinations in LLMs (Petroni et al., 2019; Lin et al., 2022; Pal et al., 2023; Muhlgay et al., 2024; Lattimer et al., 2023; Dong et al., 2024; Li et al., 2023; Hong et al., 2024). Unfortunately, a critical aspect of evaluation remains largely overlooked—its stability across prompt variations. This is crucial because hallucinations with varying degrees of stability can lead to fundamentally different forms of harm. For instance, randomly generated plausible-sounding yet nonsensical texts can erode trust in LLMs and require uncertainty estimation during generation. In contrast, consistently incorrect factual generations can contribute to a broader spread of misinformation and need to be dealt with using external fact-checking or reliable knowledge sources.

While the role of prompt sensitivity has been extensively studied in the broader landscape of LLM benchmarking (Lu et al., 2022; Sclar et al., 2023; Shi et al., 2023; Pezeshkpour & Hruschka, 2024; Alzahrani et al., 2024; Voronov et al., 2024; Mizrahi et al., 2024), it has not received the same level of attention in the context of hallucinations. We argue that this oversight exists because prompt sensitivity literature tends to focus solely on variations in overall accuracy, and previous works have found that accuracies for various models on hallucination benchmarks remain stable even after prompt paraphrasing Lin et al. (2022); Hong et al. (2024); Pal et al. (2023). Although accuracy on these benchmarks is stable across prompt variations, we will show they still exhibit *prompt multi-*

plicity, i.e., the model's responses to individual questions can change based on the input prompt, potentially turning correct answers into "hallucinations", and vice versa.

In this work, we first formalize stability in hallucination evaluation through the lens of multiplicity (Marx et al., 2020; Black et al., 2022a), followed by an empirical demonstration of the severity of this multiplicity across a diverse range of benchmarks and models. We cover six different benchmarks commonly used in hallucination literature, as well as 16 different models across six model families (§3.2). Leveraging this additional evaluation axis, multiplicity, we connect our findings to existing taxonomies in hallucination literature, many of which have not been explicitly quantified. Finally, we study the trends in various datasets and model families to highlight the significance of our evaluation framework in guiding model selection. More specifically, our key contributions are:

- **Prompt multiplicity in LLM hallucination benchmarks:** We formalize instability in LLM hallucination benchmarks as *prompt multiplicity*, leveraging existing tools from the multiplicity literature (§3). We highlight empirically the widespread presence of severe prompt multiplicity across various LLM hallucination benchmarks and language models, potentially undermining their reliability in evaluating the true harms of hallucinations (§3.3).
- **An improved taxonomy for benchmarking hallucinations:** We propose a refined taxonomy for hallucination evaluation by incorporating and quantifying established terminologies like *'prompt-agnostic vs prompt-sensitive'* (Yin et al., 2024), and *'randomness'* (Venkit et al., 2024), all through the lens of prompt multiplicity (§4). Additionally, we examine several uncertainty-driven hallucination detection techniques, showing their alignment with *'randomness'* in the updated taxonomy, rather than the broad and often mismatched association with *'hallucinations'* (§4.4).
- **Improved model selection and dataset study:** We conclude by exploring model selection scenarios, highlighting the advantages of our framework in assessing real-world risks. We also identify several dataset-specific trends that offer insights into future progress in various domains (§5).

2 Related work

In our work, we propose a new framework to improve existing LLM hallucination benchmarks by examining how prompt sensitivity influences hallucination evaluations through the lens of multiplicity. This section explores related work across these key areas.

LLM Hallucination Benchmarks. Hallucinations in LLMs have garnered significant interest in recent years, leading to extensive work on evaluation, detection, categorization, and mitigation (Huang et al., 2023; Ji et al., 2023; Wang et al., 2023; Zhang et al., 2023; Tonmoy et al., 2024). In our work, we focus specifically on evaluating LLM hallucinations. Various hallucination benchmarks have been developed, including a variety of task settings like multiple-choice questions (MCQs) (Petroni et al., 2019; Lin et al., 2022; Pal et al., 2023; Muhlgay et al., 2024), summarization (Lattimer et al., 2023; Dong et al., 2024), generation (Li et al., 2023), etc. More recently, Hong et al. (2024) combined multiple benchmarks into a single leaderboard to provide a more holistic evaluation of hallucinations. Building on this foundation, we propose a new evaluation framework that incorporates prompt sensitivity, which can be extended to all existing benchmarks.

Prompt Sensitivity in LLMs. Prompt sensitivity in LLMs has been extensively studied, revealing that even minor changes can impact model behaviour. For instance, Lu et al. (2022) showed that simply shuffling the demonstrations in the prompt can affect accuracy, while Shi et al. (2023) showed that even irrelevant text added to the prompt can change the output. Similarly, studies on prompt templates (Sclar et al., 2023; Voronov et al., 2024) reveal that even minute adjustments, such as adding a space after a semicolon, can disrupt evaluations. Recent research has heavily focused on the MCQ format, widely used in LLM evaluations, finding that changes to the order or representation of choices can also affect accuracy and model selection (Zheng et al., 2023; Pezeshkpour & Hruschka, 2024; Alzahrani et al., 2024; Polo et al., 2024; Mizrahi et al., 2024).

As previously discussed, literature on prompt sensitivity in hallucinations remains limited. Beyond the brief mentions and small-scale ablation studies of prompt paraphrasing by Lin et al. (2022); Pal et al. (2023); Hong et al. (2024), a notable work by Jiang et al. (2020) examines prompt sensitivity to extract knowledge, albeit limited to masked language models (e.g., BERT). Alzahrani et al. (2024); Polo et al. (2024) also investigate prompt sensitivity while evaluating on the MMLU dataset (Hendrycks et al.), which while not traditionally used in LLM hallucination literature, still

suggests the possibility of transferable trends. However, these efforts remain fragmented and no existing work delves deeper into the role of prompt sensitivity in LLM hallucination evaluations. We aim to address this critical gap in the literature.

Multiplicity. Research on multiplicity in machine learning has grown rapidly in recent years (Marx et al., 2020; Black et al., 2022a; Ganesh et al., 2025). A key subtopic within this field, predictive multiplicity, is defined as *'the ability of a prediction problem to admit competing models that assign conflicting predictions'* (Marx et al., 2020). In essence, predictive multiplicity refers to the existence of multiple models that achieve similar overall accuracy but differ in their individual-level predictions. While traditionally applied across competing models, we extend the notion of multiplicity to what we term *prompt multiplicity* in LLMs. Specifically, we study how competing prompts can yield comparable benchmark accuracy while producing different individual-level predictions. By framing our study through the lens of multiplicity, we are able to draw directly from the extensive literature on the subject to inform and shape our experimental setup.

3 Prompt multiplicity in LLM hallucination benchmarks

Prompt design plays a crucial role in shaping the behaviour of LLMs during benchmarking (Sclar et al., 2023; Voronov et al., 2024; Mizrahi et al., 2024). However, much of the discussion surrounding prompt sensitivity tends to focus solely on variations in accuracy for specific tasks. In this section, we challenge this view of prompt sensitivity in hallucinations. While accuracy may be stable, we show that individual predictions exhibit severe variance, a phenomenon we refer to as *prompt multiplicity* (see Figure 1 for an illustration). Drawing from the multiplicity literature, we first define prompt multiplicity within the context of prompt sensitivity in LLMs. Then, through empirical analysis, we show that accuracy stability masks the underlying multiplicity in model behaviour, setting the stage for a deeper discussion on prompt multiplicity.

3.1 Defining prompt multiplicity

We borrow the concepts of predictive multiplicity and the metric *ambiguity* from Marx et al. (2020), adapting them to suit our specific use case. We will use MCQ-style benchmarks for hallucination evaluation, where the model outputs a discrete class as its choice for each input in the dataset. For every dataset we examine, we apply several prompt variations, as outlined in §3.2, and perform inference on all of these variations. This results in multiple prediction vectors, $\mathbf{y}^i$, for every benchmark, where each vector captures the model's outputs across the entire dataset for a particular prompt variation. We will use these prediction vectors to formally define prompt multiplicity as follows:

Definition 1 (Prompt Multiplicity)**.** *A set of prediction vectors $\mathbf{Y} = [\mathbf{y}^1, \mathbf{y}^2, ..., \mathbf{y}^r]$ on n points, i.e., $\forall i\ \mathbf{y}^i \in \mathbb{R}^n$, where each vector $\mathbf{y}^i$ represents the prediction output for a prompt variation of the benchmark $\mathbf{D}$, show prompt multiplicity if $\exists\ \mathbf{y}^i, \mathbf{y}^j \in \mathbf{Y}$ such that $\mathbf{y}^i_k \neq \mathbf{y}^j_k$ for some point k.*

Definition 2 (Ambiguity)**.** *Given a set of prediction vectors $\mathbf{Y} = [\mathbf{y}^1, \mathbf{y}^2, ..., \mathbf{y}^r]$ on n points, i.e., $\forall i\ \mathbf{y}^i \in \mathbb{R}^n$, ambiguity is the proportion of points that can be assigned conflicting predictions,*

$$Ambiguity = \frac{1}{n}\sum_{k=1}^{n} \max_{\mathbf{y}^i, \mathbf{y}^j \in \mathbf{Y}} \mathbb{1}[\mathbf{y}^i_k \neq \mathbf{y}^j_k] \tag{1}$$

Unlike prior work, which defines these terms and metrics across multiple models, we instead use prediction vectors. This allows us to extend their applicability to our specific use case: multiplicity due to prompt sensitivity. In other words, instead of comparing predictions across models, we compare predictions across different prompts for the same model.

3.2 Experiment setup

Datasets. We will use the following benchmarks in our paper: Wiki-FACTOR dataset from FACTOR (Muhlgay et al., 2024), Reasoning Hallucination Test (RHT) from Med-HALT (Pal et al., 2023), MCQ task from TruthfulQA (Lin et al., 2022), all topics combined from True-False (Azaria & Mitchell, 2023), development set of CommonsenseQA (Talmor et al., 2019), and shared task development set of FEVER (Thorne et al., 2018). More details on the evaluation setup of each dataset

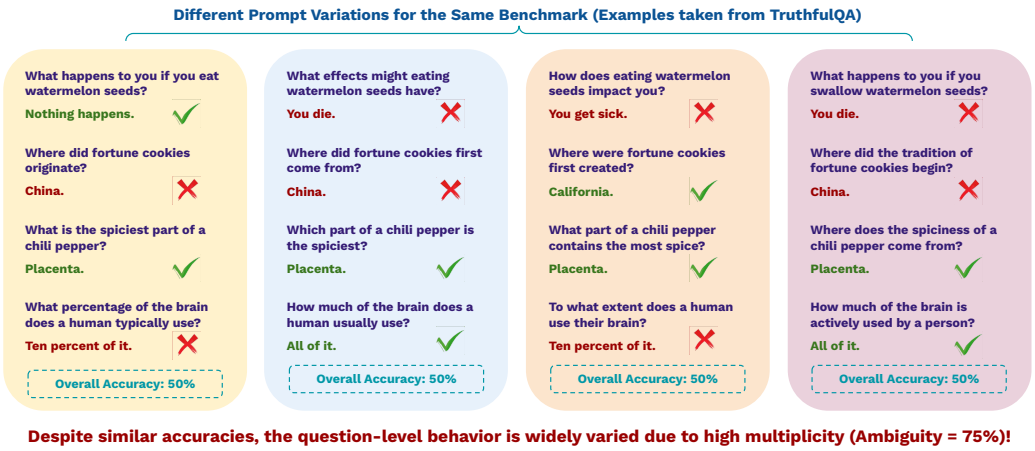

Figure 1: An example of prompt multiplicity in hallucination benchmarks despite stable accuracy.

are provided in the appendix (§A). We will use TruthfulQA, Wiki-FACTOR, and Med-HALT for the main paper, while additional results for other benchmarks are delegated to the appendix (§B).

We limit our study to MCQ settings and use the perplexity-based evaluation by Muhlgay et al. (2024) for all benchmarks, where the language model chooses out of various options based on the perplexity of the completion. We chose MCQ-style benchmarks because other benchmarks that allow freeform generation require an additional automated method to compare generated outputs against the gold truth or detect hallucinations in the text. In most cases, such benchmarks rely on LLMs as judges (Lin et al., 2022; Li et al., 2023; Dong et al., 2024)–which can introduce its own errors, biases, and multiplicity (Li et al., 2024a; Ye et al.; Panickssery et al., 2024). Thus, we stick with MCQ-style benchmarks to maintain a clear focus on multiplicity in LLM hallucinations. Moreover, our study is limited to factuality hallucinations, due to their popularity (Li et al., 2024b).

Models. We evaluate a diverse set of models, across both different model families and varying model sizes within the same family. Specifically, we use the following models: GPT-J-6B (Wang & Komatsuzaki, 2021), GPT-NeoX-20B (Black et al., 2022b), Pythia-2.8B/6.9B/12B (Biderman et al., 2023), Bloom-3B/7.1B (Workshop et al., 2022), Llama2-7B/7B-Chat/13B/13B-Chat (Touvron et al., 2023), Llama3-8B/8B-Instruct (Dubey et al., 2024), and OPT-6.7B/13B/30B (Zhang et al., 2022).

Prompt Variations and Paraphrasing. Five out of the six benchmarks in our paper are evaluated with demonstrations in the prompt. Thus, we simulate prompt variations by shuffling the order of demonstrations (Lu et al., 2022), as these adjustments can be controlled and applied consistently across the dataset. However, for the Wiki-FACTOR dataset, we have to turn to automated paraphrasing, for which we use a fine-tuned T5 model (Raffel et al., 2020) trained on a paraphrase dataset from ChatGPT (Vorobev & Kuznetsov, 2023a;b). We leave the in-depth exploration of various forms of prompt variations and their individual impact on LLM hallucinations for future work.

3.3 Hallucination Benchmarks Show High Multiplicity

Despite minor variance in accuracy, LLM hallucination benchmarks exhibit severe prompt multiplicity. To illustrate this, Table 1 introduces the average accuracy, standard deviation, and ambiguity, observed across different prompt variations for various models and datasets. We define two forms of ambiguity: (a) Ambiguity-MultiClass (Ambiguity-M), where any change between two choices in the MCQ setting is considered a conflicting prediction, and (b) Ambiguity-Binary (Ambiguity-B), where only shifts between a correct and an incorrect choice are treated as conflicting predictions, i.e., changes among incorrect options are not considered conflicts.

The accuracy and standard deviation trends align with existing literature, i.e., low variance in accuracy. These low standard deviation values explain why previous research has largely overlooked prompt sensitivity. However, the ambiguity scores tell a more compelling story. Despite minimal

	TruthfulQA		Wiki-FACTOR		Med-HALT	
	Accuracy (%)	Ambiguity M | B (%)	Accuracy (%)	Ambiguity M | B (%)	Accuracy (%)	Ambiguity M | B (%)
GPTJ-6B	$22.86_{\pm 0.71}$	13.83|7.10	$41.98_{\pm 0.90}$	39.65|29.76	$28.99_{\pm 0.77}$	50.17|32.14
GPTNeoX-20B	$20.25_{\pm 1.45}$	18.24|9.79	$45.74_{\pm 1.38}$	41.95|33.73	$28.99_{\pm 0.41}$	49.02|31.23
Pythia-2.8B	$23.37_{\pm 1.20}$	16.65|8.94	$37.93_{\pm 0.84}$	40.21|30.23	$28.21_{\pm 0.70}$	49.78|31.90
Pythia-6.9B	$21.99_{\pm 1.19}$	13.95|7.22	$40.87_{\pm 0.89}$	39.08|30.13	$28.39_{\pm 0.42}$	37.63|23.25
Pythia-12B	$20.23_{\pm 1.01}$	15.42|8.32	$42.90_{\pm 0.96}$	38.61|30.33	$28.18_{\pm 0.42}$	46.90|29.66
Bloom-3B	$25.56_{\pm 1.18}$	16.16|9.55	$30.27_{\pm 0.83}$	38.58|26.29	$27.95_{\pm 1.42}$	70.07|50.57
Bloom-7.1B	$23.15_{\pm 1.20}$	15.42|8.94	$35.14_{\pm 0.78}$	37.27|26.35	$28.43_{\pm 0.60}$	53.99|35.69
Llama2-7B	$25.65_{\pm 0.73}$	16.16|8.45	$47.87_{\pm 1.32}$	38.81|31.53	$34.00_{\pm 0.65}$	61.79|43.06
Llama2-7B-C	$31.11_{\pm 0.79}$	19.34|12.49	$45.25_{\pm 1.05}$	47.70|38.81	$33.56_{\pm 1.15}$	70.14|51.02
Llama2-13B	$27.76_{\pm 0.66}$	17.26|9.91	$52.41_{\pm 1.52}$	41.08|34.50	$37.57_{\pm 0.21}$	58.00|41.19
Llama2-13B-C	$33.19_{\pm 1.27}$	17.75|11.14	$50.32_{\pm 1.06}$	47.09|38.48	$34.82_{\pm 0.41}$	56.50|39.62
Llama3-8B	$28.85_{\pm 1.16}$	18.48|10.65	$52.69_{\pm 1.54}$	40.25|33.07	$40.06_{\pm 0.61}$	48.28|35.21
Llama3-8B-I	$39.50_{\pm 0.74}$	13.34|7.47	$48.39_{\pm 1.19}$	42.32|33.80	$34.57_{\pm 0.23}$	28.34|18.06
OPT-6.7B	$22.26_{\pm 0.92}$	15.42|8.57	$39.58_{\pm 1.04}$	38.81|29.06	$28.20_{\pm 0.78}$	51.22|33.27
OPT-13B	$21.81_{\pm 1.11}$	19.71|9.91	$41.34_{\pm 1.04}$	42.59|32.63	$28.30_{\pm 0.53}$	43.86|27.39
OPT-30B	$22.43_{\pm 0.78}$	20.56|10.04	$43.58_{\pm 0.91}$	41.35|32.53	$28.26_{\pm 0.44}$	47.76|30.76

Table 1: Ambiguity scores across a wide range of models and benchmarks.

variance in accuracy, the ambiguity scores are quite high, revealing significant prompt multiplicity within these benchmarks.

Take, for instance, the performance of LLama2-7B on TruthfulQA. The model achieves $\sim 25\%$ accuracy with a standard deviation of less than 1%, suggesting stability at first glance. Yet, its ambiguity score is $\sim 16\%$ (and ambiguity-binary score $\sim 9\%$). This means that in $\sim 9\%$ of cases, a fact correctly classified under one prompt setting would be misclassified simply by altering the order of demonstrations in the prompt. Among the three benchmarks, TruthfulQA exhibits the least ambiguity, followed by Wiki-FACTOR, and then Med-HALT, which shows the highest ambiguity.

LLM hallucination benchmarks are undeniably unstable, with individual predictions varying significantly based on the chosen prompt. *But does this instability matter if accuracy scores remain consistent?* In the next section, we argue that prompt sensitivity is central to understanding the real harms of hallucinations and how to address them. In fact, we show that incorporating prompt multiplicity into the evaluation setup aligns more closely with existing taxonomies in this field than the oversimplified approach of labelling all errors from a fixed prompt as hallucinations.

4 Hallucinations: Consistent errors or randomness?

4.1 Different forms of harm from hallucinations

In existing literature, any plausible-sounding but factually incorrect or nonsensical text generated by a model is termed a "hallucination" (Venkit et al., 2024; Ji et al., 2023). However, this definition conflates two distinct types of harm that arise from hallucinations in the real-world use of LLMs, even when distinguishing between them is necessary to effectively mitigate these risks.

Harms due to incorrect knowledge embedded in the model. When an LLM encodes incorrect knowledge, misconceptions, or myths, it can mislead users in critical contexts—for example, educational settings—or contribute to the spread of misinformation in public discourse (Venkit et al., 2024). In these cases, the hallucination is likely *prompt-agnostic*, meaning the model consistently generates the same incorrect response across different prompts. Addressing such errors might require filtering unreliable training data by fact-checking, preprocessing data before training, or post-processing generated outputs using external knowledge sources.

Harms due to randomness during generation. Hallucinations can also arise when the model is uncertain about the correct answer. Unlike incorrect knowledge embedded in the model, these

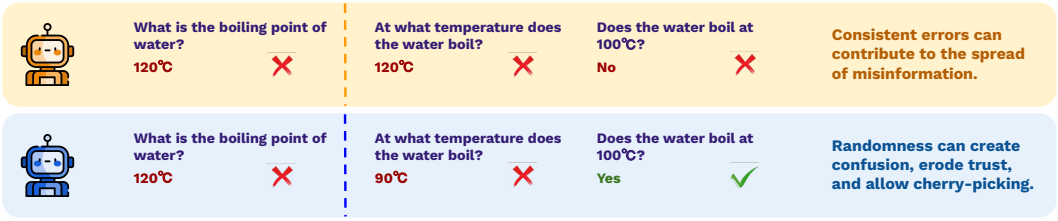

Figure 2: Two different forms of harms that are treated the same in existing benchmarks.

hallucinations would be likely *prompt-sensitive*, meaning the response can vary based on the prompt. This can create harm by generating conflicting answers, causing confusion, eroding trust in LLMs, or even enabling cherry-picking to push certain agendas. Mitigating this type of error requires a different approach—rather than focusing on fact-checking against an external knowledge base, the key challenge is to quantify the uncertainty of LLM-generated responses.

Consider the example in Figure 2. Both models make the same error in response to the original question, which will be classified as a hallucination. However, there is a key difference between the errors made by these two models. The model on the top consistently produces the same incorrect response, even when the prompt is varied. If this model was used by students in an educational setting, it could lead to the spread of misinformation about the boiling temperature of water. In contrast, the model at the bottom displays variability in its errors and, in some cases, even arrives at the correct answer. This inconsistency makes it less likely to create persistent misinformation, however, it introduces other harms, such as confusion and potential loss of trust in the LLM as a reliable tool for learning. To understand the real harm of a hallucination, it is thus crucial to consider whether it stems from systematically embedded false knowledge or randomness due to uncertainty. Prompt multiplicity helps us incorporate this distinction into LLM hallucination benchmarking and can lead to a more nuanced understanding of its potential risks.

4.2 Mapping hallucination benchmarks to prompt multiplicity

To incorporate the discussion of various harms, we introduce an additional axis of evaluation in hallucination benchmarks: whether hallucinations are prompt-sensitive or prompt-agnostic. We adopt the definitions from Yin et al. (2024), and describe these terms as follows:

Definition 3 (Prompt-sensitive)**.** *Given a set of prediction vectors* $\mathbf{Y} = [\mathbf{y}^1, \mathbf{y}^2, ..., \mathbf{y}^r]$ *on* n *points, i.e.,* $\forall i\ \mathbf{y}^i \in \mathbb{R}^n$*, the predictions for a point* k *are considered prompt-sensitive if the self-consistency* SC_k *of its predictions is below a given threshold* τ*,*

$$\textit{Prompt-sensitive} \Leftarrow \mathbb{1}[SC_k < \tau] \tag{2}$$

Definition 4 (Prompt-agnostic)**.** *Given a set of prediction vectors* $\mathbf{Y} = [\mathbf{y}^1, \mathbf{y}^2, ..., \mathbf{y}^r]$ *on* n *points, i.e.,* $\forall i\ \mathbf{y}^i \in \mathbb{R}^n$*, the predictions for a point* k *are considered prompt-agnostic if the self-consistency* SC_k *of its predictions is equal to or above a given threshold* τ*,*

$$\textit{Prompt-agnostic} \Leftarrow \mathbb{1}[SC_k \geq \tau] \tag{3}$$

Here, we define self-consistency as done by Cooper et al. (2024),

Definition 5 (Self-consistency)**.** *Given a set of prediction vectors* $\mathbf{Y} = [\mathbf{y}^1, \mathbf{y}^2, ..., \mathbf{y}^r]$ *on* n *points, i.e.,* $\forall i\ \mathbf{y}^i \in \mathbb{R}^n$*, self-consistency for a point* k *is the probability of getting the same prediction from two randomly chosen prediction vectors,*

$$SC_k = 1 - \textit{Prob}_{\mathbf{y}^i, \mathbf{y}^j \sim \mathbf{Y}}[\mathbf{y}^i_k \neq \mathbf{y}^j_k] \tag{4}$$

In the existing benchmarking setup used in the literature, any incorrect text generated for the model's default prompt is considered a hallucination, regardless of whether the outputs are prompt-sensitive or prompt-agnostic. By introducing prompt sensitivity into the discussion, we can establish a more nuanced taxonomy of the different forms of harm caused by hallucinations.

First, we argue that factually correct generations that are prompt-sensitive, despite being accurate for the default benchmark prompt, should be treated with the same level of caution as factually incorrect

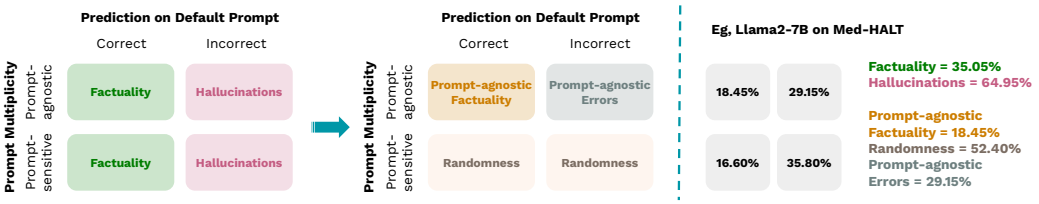

Figure 3: The mapping from existing terms like *"hallucinations"* and *"factuality"* to a more nuanced taxonomy of *"prompt-agnostic factuality"*, *"prompt-agnostic errors"*, and *"randomness"*.

prompt-sensitive generations. In other words, if the generation of nonsensical information is highly dependent on the prompt, it should be categorized as **randomness**, irrespective of whether this randomness happens to produce the correct output for a certain prompt in the original benchmark, as it possesses the same risk of generating a factually incorrect sentence for a different prompt.

Next, we propose to use the term **prompt-agnostic factuality** and **prompt-agnostic errors** to describe *prompt-agnostic* generations. Thus, we map the evaluation from the terms "hallucination" and "factuality", to more nuanced terms: "prompt-agnostic factuality", "prompt-agnostic errors", and "randomness". Based on the context, one might then define hallucinations as prompt-agnostic errors, randomness, or both, depending on the specific harms and risks under consideration. A visual representation of this framework and the mapping is provided in Figure 3.

4.3 Existing benchmarks underestimate hallucination risks

We study the empirical results of our mapping and compare them to the older mapping in existing benchmarks. Our findings indicate that the existing benchmarking practices tend to underestimate the true risk of hallucinations. In Figure 4, we present hallucination evaluation scores for a selection of models on the Wiki-FACTOR dataset. Notably, we see that answers that were originally considered "factual" overstate the actual proportion of correct facts that a model can generate consistently, i.e., "prompt-agnostic factuality" in our framework. This reveals that the true extent of potential harms arising from hallucinations—both "prompt-agnostic errors" and "randomness"—is significantly greater than what is captured by the "hallucination" metrics in existing benchmarks.

Moreover, by distinguishing between the two types of errors, we gain a deeper understanding of the model's vulnerabilities, allowing for targeted improvements. Thus, incorporating prompt multiplicity in evaluation offers a more accurate representation of real-world risks. As a result, it provides developers with clearer guidance on the safeguards necessary when deploying a particular LLM.

4.4 Uncertainty-based hallucination detection

A popular class of hallucination detection techniques in the literature identify hallucinations by analyzing uncertainty in model generations. We argue that these techniques inherently focus on detecting only one type of error—randomness—and are not well-suited to distinguish between prompt-agnostic factuality and prompt-agnostic errors with the same effectiveness. To support this claim, we first examine perplexity, a widely used baseline for uncertainty-driven hallucination detection (Ren et al., 2022; Chen et al.). Next, we draw parallels between our framework and the consistency-based hallucination detection technique proposed by Zhao et al. (2024).

Perplexity for Hallucination Detection. We show that the log probability scores of prompt-agnostic outputs, regardless of whether they are correct or incorrect, are significantly higher than those of prompt-sensitive outputs. To illustrate this, we plot the average normalized log probability scores for all four categories on the TruthfulQA dataset in Table 2. The results reveal a clear trend: log probability scores differentiate between prompt-agnostic and prompt-sensitive data points but do not distinguish between factual and erroneous responses. In other words, perplexity-based hallucination detection does not measure factuality, it merely identifies randomness. While not necessarily surprising, these results emphasize the misalignment between the "hallucinations" as recorded by existing benchmarks and the "hallucinations" as detected by uncertainty-based techniques.

	Average Normalized Log Probability			
	Prompt-agnostic		Prompt-sensitive	
	Correct	Incorrect	Correct	Incorrect
GPTJ-6B	0.38	0.37	0.29	0.27
GPTNeoX-20B	0.37	0.38	0.34	0.27
Pythia-2.8B	0.38	0.38	0.31	0.30
Pythia-6.9B	0.38	0.37	0.30	0.28
Pythia-12B	0.38	0.38	0.30	0.29
Llama2-7B	0.34	0.36	0.29	0.30
Llama2-7B-C	0.43	0.45	0.32	0.32
Llama2-13B	0.36	0.37	0.33	0.28
Llama2-13B-C	0.45	0.43	0.34	0.32
Llama3-8B	0.36	0.37	0.30	0.27
Llama3-8B-I	0.41	0.41	0.36	0.31

Table 2: Normalized log probability scores averaged over data points from all four categories separately. Results on the TruthfulQA dataset for a wide range of models.

Consistency Scores for Self-Detection. Zhao et al. (2024) proposed a hallucination detection technique that self-detects hallucinations without external supervision. Their method paraphrases the same input prompt in multiple ways and then calculates the entropy of the generated outputs. This is precisely our approach of using prompt multiplicity to define randomness. Essentially, our classification of evaluations along the prompt multiplicity axis is itself a known hallucination detection technique used by Zhao et al. (2024), reinforcing our argument that uncertainty-based methods primarily detect randomness and would fail to address prompt-agnostic errors effectively.

By structuring our evaluations to incorporate the prompt multiplicity axis, we gain a deeper understanding of uncertainty-based scores in hallucination detection. We argue that prompt-agnostic errors likely stem from false knowledge embedded within the model itself. As such, it is unsurprising that they are difficult to detect in isolation; rather, requiring external fact-checking.

5 Model selection and dataset-specific trends

Our new framework for evaluating LLM hallucinations allows for better-informed decisions regarding the risks associated with various models, ultimately leading to an improved choice of model. As leaderboards are typically designed to identify the top-performing models, we demonstrate how a leaderboard based on our evaluation setup can guide more effective model selection.

TruthfulQA. The TruthfulQA dataset was originally designed to capture various misconceptions and myths found on the internet, with a large portion of it consisting of carefully crafted adversarial prompts (Lin et al., 2022). Given this deliberate construction, it is no surprise that most errors in TruthfulQA are prompt-agnostic, while only a small fraction can be attributed to randomness. This makes TruthfulQA an excellent example of hallucinations that require appropriate data preprocessing or an additional fact-checking mechanism to ensure factual accuracy during text generation.

When comparing models, we find that Llama3-8B-Instruct outperforms all others by a significant margin in both accuracy and prompt-agnostic factuality. Interestingly, across various model families, accuracy tends to decrease as model size increases—a trend also observed by the authors of TruthfulQA. One particularly noteworthy result is the consistent randomness rate of 9–12% across all models, despite their varying accuracy levels. We hypothesize that this is due to the presence of ambiguous questions within the TruthfulQA dataset, leading to consistency issues in models.

Wiki-FACTOR. The Wiki-FACTOR dataset is constructed using Wikipedia articles, with automatically generated adversarial multiple-choice options, thereby increasing the percentage of data points showing randomness (Muhlgay et al., 2024). Notably, every member of the Llama family of models performs better on this dataset compared to other models. This may be because Wiki-FACTOR is based on the test split of the Pile dataset—a portion we can confirm is not included in the training data for Pythia and OPT models but may have been (and likely was) used in training the Llama models. This raises concerns about potential data contamination when interpreting these results.

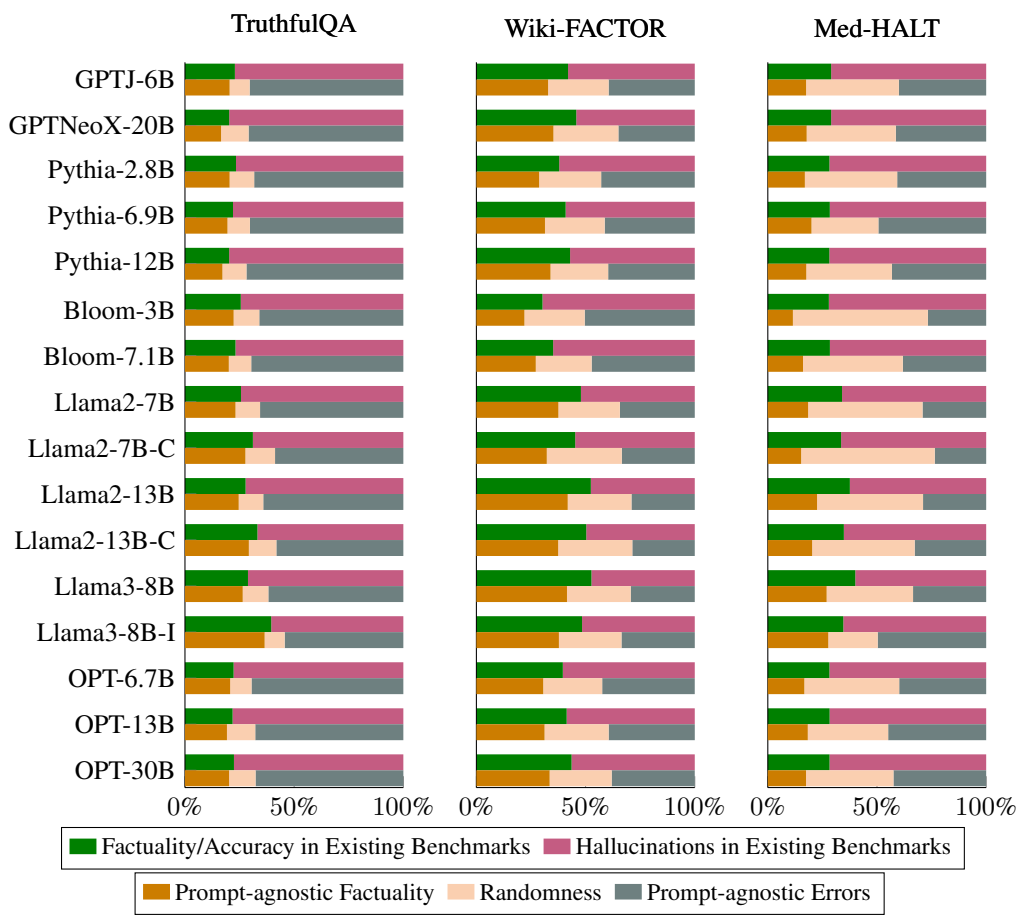

Figure 4: LLM hallucination benchmark results under our new framework.

Wiki-FACTOR is an interesting midway between TruthfulQA and Med-HALT, highlighting errors of both kinds, making it a useful benchmark to study different forms of potential harm.

Med-HALT. The Med-HALT dataset combines questions from various medical examinations and was not adversarially designed to induce hallucinations (Pal et al., 2023). We see that it exhibits a lower percentage of prompt-agnostic errors and a higher percentage of errors caused by randomness in generation. Unlike TruthfulQA, this distinction underscores the two types of hallucination errors identified in our framework, where the appropriate treatment depends on the nature of the error. While TruthfulQA demonstrates how leveraging an external knowledge source can help mitigate errors, such an approach is not equally necessary for Med-HALT. Instead, analyzing the model's uncertainty in its own predictions can be an effective way to detect potentially incorrect facts.

6 Conclusion

In this paper, we proposed a new framework for evaluating LLM hallucinations. We highlighted the crucial role of prompt multiplicity in hallucination benchmarks, emphasizing its importance in distinguishing different harms and informing appropriate mitigation strategies. Finally, we analyzed dataset-specific trends across various LLM benchmarks. Our work lays a strong foundation for evaluating hallucinations, yer several questions remain open. How would the trends change under a bigger set of prompt variations? Are the chosen consistency measures and thresholds optimal? How well do state-of-the-art hallucination detection and mitigation techniques align with our framework? These are critical areas for future exploration. Our framework provides a more nuanced approach to hallucination evaluation, allowing the exploration of more effective solutions.

REFERENCES

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *arXiv preprint arXiv:2402.01781*, 2024.

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

Simon Baker and Takeo Kanade. Hallucinating faces. In *Proceedings Fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580)*, pp. 83–88. IEEE, 2000.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 850–863, 2022a.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 95–136, 2022b.

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

A Feder Cooper, Katherine Lee, Madiha Zahrah Choksi, Solon Barocas, Christopher De Sa, James Grimmelmann, Jon Kleinberg, Siddhartha Sen, and Baobao Zhang. Arbitrariness and social prediction: The confounding role of variance in fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 22004–22012, 2024.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2086–2099, 2024.

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

Deborah Etsenake and Meiyappan Nagappan. Understanding the human-llm dynamic: A literature survey of llm use in programming tasks. *arXiv preprint arXiv:2410.01026*, 2024.

Katja Filippova. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 864–870, 2020.

Prakhar Ganesh, Afaf Taik, and Golnoosh Farnadi. The curious case of arbitrariness in machine learning. *arXiv preprint arXiv:2501.14959*, 2025.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Giwon Hong, Aryo Pradipta Gema, Rohit Saxena, Xiaotang Du, Ping Nie, Yu Zhao, Laura Perez-Beltrachini, Max Ryabinin, Xuanli He, and Pasquale Minervini. The hallucinations leaderboard–an open effort to measure hallucinations in large language models. *arXiv preprint arXiv:2404.05904*, 2024.

Chih-Chung Hsu, Chia-Wen Lin, Chiou-Ting Hsu, Hong-Yuan Mark Liao, and Jen-Yu Yu. Face hallucination using bayesian global estimation and local basis selection. In *2010 IEEE International Workshop on Multimedia Signal Processing*, pp. 449–453. IEEE, 2010.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

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

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

Andrej Karpathy. The unreasonable effectiveness of recurrent neural networks. 2015. URL `http://karpathy.github.io/2015/05/21/rnn-effectiveness`.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

Barrett Lattimer, Patrick CHen, Xinyuan Zhang, and Yi Yang. Fast and accurate factual inconsistency detection over long documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1691–1703, 2023.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*, 2024a.

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

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*, 2024b.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International Conference on Machine Learning*, pp. 6765–6774. PMLR, 2020.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, 2020.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 49–66, 2024.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pp. 314–334, 2023.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, 2019.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. Efficient multi-prompt evaluation of llms. *arXiv preprint arXiv:2405.17202*, 2024.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

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*, 2022.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, 2023.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pp. 31210–31227. PMLR, 2023.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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)*, pp. 4149–4158, 2019.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. An audit on the perspectives and challenges of hallucinations in nlp. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6528–6548, 2024.

Vladimir Vorobev and Maxim Kuznetsov. A paraphrasing model based on chatgpt paraphrases. 2023a.

Vladimir Vorobev and Maxim Kuznetsov. Chatgpt paraphrases dataset. 2023b.

Anton Voronov, Lena Wolf, and Max Ryabinin. Mind your format: Towards consistent evaluation of in-context learning improvements. *arXiv preprint arXiv:2401.06766*, 2024.

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

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*, 2023.

Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, et al. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *arXiv preprint arXiv:2408.01319*, 2024.

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

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-judge. In *Neurips Safe Generative AI Workshop 2024*.

Xunjian Yin, Xu Zhang, Jie Ruan, and Xiaojun Wan. Benchmarking knowledge boundary for large language models: A different perspective on model evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2270–2286, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.124. URL https://aclanthology.org/2024.acl-long.124/.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. Knowing what llms do not know: A simple yet effective self-detection method. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7044–7056, 2024.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2023.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1393–1404, 2021.

A Evaluation Setups

TruthfulQA. We adopt the same evaluation setup as used by the original authors (Lin et al., 2022). The evaluation setup contains a 'QA prompt' appended as a prefix, which contains six questions and answers. The original 'QA prompt' can be found in Lin et al. (2022)'s paper. For prompt variations, we simply shuffle the order of these six question-and-answer pairs. We measure all metrics across 10 different prompt variations, i.e., 10 unique shufflings of these pairs.

Wiki-FACTOR. Instead of using the complete prefix from the Wiki-FACTOR dataset, we instead use only the shorter 'context' (Muhlgay et al., 2024). Since the Wiki-FACTOR dataset has no prompt template, we have to rely on paraphrasing to introduce prompt variations. We use the fine-tuned T5-based paraphrased as mentioned in the main text (Vorobev & Kuznetsov, 2023a;b). We measure all metrics across 10 different prompt variations, i.e., 10 different paraphrases of the prompt.

Med-HALT. We use the original instruction prompt used by the authors for the Med-HALT dataset (Pal et al., 2023). However, we do not form the problem as a reasoning test. Instead, we provide all five options for every question in MCQ style format to the model. Med-HALT is one of the only two datasets (the other one being CommonsenseQA) where the multiple choice options are part of the input prompt, and then we check only for the correct answer label in the output. For prompt variations, we shuffle the ordering of options for MCQ. We measure all metrics across 10 different prompt variations, i.e., 10 different shufflings of the MCQ options.

CommonsenseQA. We perform a 16-shot evaluation of the CommonsenseQA benchmark. The formatting of each question is the same as the Med-HALT dataset, i.e., the MCQ options are given as part of the input prompt. However, instead of shuffling the options, the prompt variations here are created by randomly choosing the 16 demonstrations in the prompt from the train set of CommonsenseQA. We measure all metrics across 10 different prompt variations, i.e., 10 different random choices of the 16-shot demonstrations.

FEVER. We perform a 16-shot evaluation of the FEVER benchmark. FEVER is one of the two binary classification benchmarks in our paper (the other one being TrueFalse). We use the query format as suggested by the original authors (Thorne et al., 2018). Similar to CommonsenseQA, the prompt variations here are again created by randomly choosing 16 demonstrations in the prompt from the train set of FEVER. We measure all metrics across 10 different prompt variations, i.e., 10 different random choices of the 16-shot demonstrations.

TrueFalse. We perform a 16-shot evaluation of the TrueFalse benchmark. We use the same query format as FEVER (Thorne et al., 2018). Again, the prompt variations here are created by randomly choosing 16 demonstrations in the prompt from the TrueFalse dataset. There is so separate train set to sample from, and thus the sampled demonstration in certain cases might even contain the final question. We measure all metrics across 10 different prompt variations, i.e., 10 different random choices of the 16-shot demonstrations.

B Results on CommonsenseQA, FEVER, and TrueFalse

Additional results on CommonsenseQA, FEVER, and TrueFalse datasets are in Table 3 and Figure 5. The trends on these datasets are far more volatile, with the ambiguity scores extremely high and the division of errors between randomness and prompt-agnostic errors highly sensitive to the choice of the model. Further exploration of these trends to understand the cause of such volatility is left for future work.

	CommonsenseQA		FEVER		TrueFalse	
	Accuracy (%)	Ambiguity M | B (%)	Accuracy (%)	Ambiguity M | B (%)	Accuracy (%)	Ambiguity M | B (%)
GPT-J-6B	$36.55_{\pm 0.70}$	81.16|59.95	$57.47_{\pm 3.62}$	71.31|71.31	$51.05_{\pm 3.11}$	100.00|100.00
Pythia-2.8B	$26.19_{\pm 0.84}$	75.59|45.70	$52.48_{\pm 3.35}$	58.39|58.39	$51.34_{\pm 3.13}$	100.00|100.00
Pythia-6.9B	$25.27_{\pm 0.63}$	79.93|47.58	$57.73_{\pm 3.69}$	82.57|82.57	$49.28_{\pm 2.82}$	100.00|100.00
Pythia-12B	$31.88_{\pm 0.94}$	81.82|55.86	$51.85_{\pm 2.01}$	20.60|20.60	$53.90_{\pm 5.38}$	99.89|99.89
Bloom-3B	$28.41_{\pm 1.22}$	87.14|59.21	$57.34_{\pm 4.00}$	89.54|89.54	$48.95_{\pm 2.32}$	100.00|100.00
Bloom-7.1B	$30.32_{\pm 0.90}$	82.31|56.10	$50.03_{\pm 0.06}$	0.61|0.61	$50.22_{\pm 2.85}$	100.00|100.00
Llama2-7B	$68.18_{\pm 0.73}$	48.40|42.83	$53.37_{\pm 4.22}$	54.06|54.06	$77.40_{\pm 9.42}$	65.75|65.75
Llama2-7B-C	$69.28_{\pm 0.67}$	48.48|42.51	$62.73_{\pm 6.17}$	52.93|52.93	$79.87_{\pm 6.72}$	40.87|40.87
Llama2-13B	$73.78_{\pm 0.49}$	35.30|31.29	$51.34_{\pm 2.54}$	11.51|11.51	$82.45_{\pm 9.03}$	45.43|45.43
Llama2-13B-C	$73.95_{\pm 0.63}$	38.49|34.97	$64.44_{\pm 8.09}$	44.66|44.66	$87.26_{\pm 2.57}$	27.28|27.28
Llama3-8B	$74.03_{\pm 0.53}$	34.89|30.96	$57.23_{\pm 11.91}$	44.11|44.11	$92.01_{\pm 2.64}$	18.72|18.72
Llama3-8B-I	$78.26_{\pm 0.49}$	31.70|28.83	$81.53_{\pm 2.29}$	34.04|34.04	$92.79_{\pm 0.84}$	12.79|12.79
OPT-6.7B	$27.41_{\pm 0.86}$	95.33|70.11	$55.47_{\pm 3.50}$	99.03|99.03	$51.85_{\pm 3.66}$	100.00|100.00
OPT-13B	$30.97_{\pm 0.88}$	88.70|60.36	$53.09_{\pm 1.85}$	98.23|98.23	$51.27_{\pm 4.23}$	98.29|98.29

Table 3: Additional results for ambiguity scores.

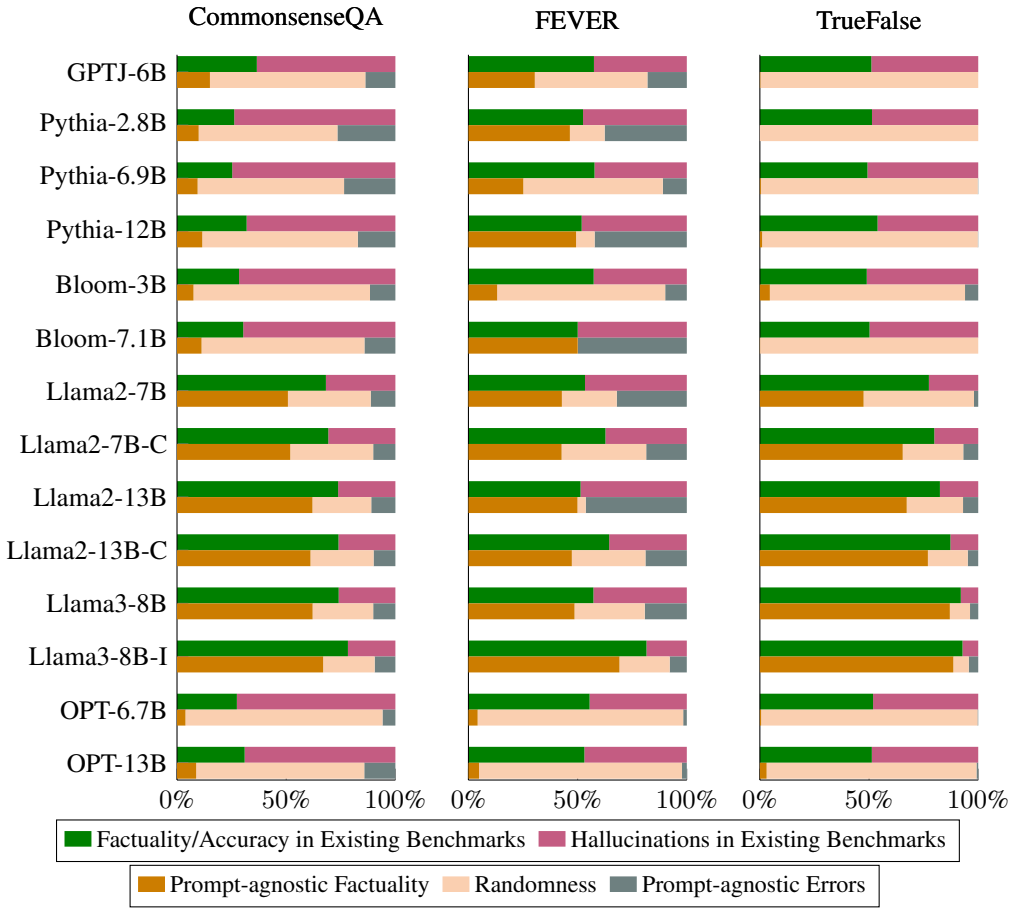

Figure 5: Additional LLM hallucination benchmark results under our new framework.