Rethinking Hallucinations: Correctness, Consistency, and Prompt Multiplicity

Prakhar Ganesh
McGill University & Mila
prakhar.ganesh@mila.quebec

Reza Shokri
National University of Singapore
reza@comp.nus.edu.sg

Golnoosh Farnadi
McGill University & Mila
farnadig@mila.quebec

Abstract

Large language models (LLMs) are known to ”hallucinate” by generating false or misleading outputs. Hallucinations pose various harms, from erosion of trust to widespread misinformation. Existing hallucination evaluation, however, focuses only on ”correctness” and often overlooks ”consistency”, necessary to distinguish and address these harms. To bridge this gap, we introduce *prompt multiplicity*, a framework for quantifying consistency through prompt sensitivity. Our analysis reveals significant multiplicity (over 50% inconsistency in benchmarks like Med-HALT), suggesting that hallucination-related harms have been severely underestimated. Furthermore, we study the role of consistency in hallucination detection and mitigation. We find that: (a) detection techniques capture consistency, not correctness, and (b) mitigation techniques like RAG can introduce additional inconsistencies. By integrating prompt multiplicity into hallucination evaluation, we provide an improved framework of potential harms and uncover critical limitations in current detection and mitigation strategies.

1 Introduction

Large language models (LLMs) have been widely adopted, excelling in numerous tasks across diverse domains (Guo et al., 2023; Kasneci et al., 2023; Etsenake & Nagappan, 2024). Despite their growing use, LLMs suffer from a critical limitation: generation of false, nonsensical or misleading outputs, studied under the umbrella of hallucinations. The term ”hallucinations” has evolved over the years, shifting from positive use in computer vision (Baker & Kanade, 2000; Hsu et al., 2010) to a 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).

With growing interest in this field, several benchmarks have been developed to assess hallucination risks in LLMs (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 still remains largely overlooked—consistency across prompt variations. While prompt sensitivity has been extensively studied in LLM benchmarking (Sclar et al., 2023; Shi et al., 2023; Pezeshkpour & Hruschka, 2024; Alzahrani et al., 2024; Mizrahi et al., 2024), it has not received the same attention in hallucination. We argue that this oversight exists because prompt sensitivity literature focuses solely on accuracy variance, and previous works have found that overall accuracies on hallucination

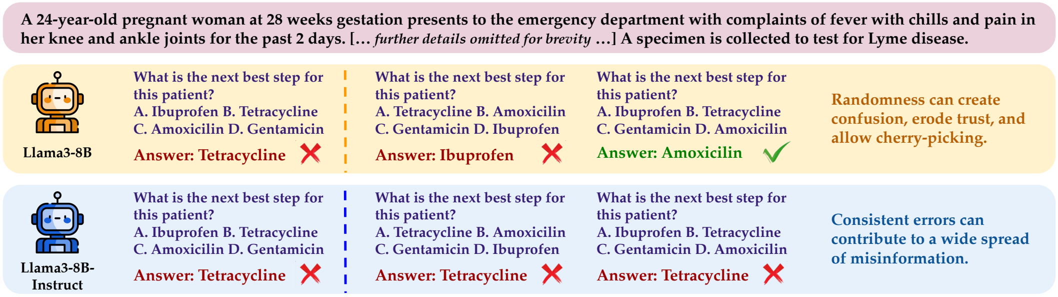

A 24-year-old pregnant woman at 28 weeks gestation presents to the emergency department with complaints of fever with chills and pain in her knee and ankle joints for the past 2 days. [... further details omitted for brevity ...] A specimen is collected to test for Lyme disease.
Llama3-8B
What is the next best step for this patient? A. Ibuprofen B. Tetracycline C. Amoxicilin D. Gentamicin
Answer: Tetracycline ✗
What is the next best step for this patient? A. Tetracycline B. Amoxicilin C. Gentamicin D. Ibuprofen
Answer: Ibuprofen ✗
What is the next best step for this patient? A. Ibuprofen B. Tetracycline C. Gentamicin D. Amoxicilin
Answer: Amoxicilin ✓
Randomness can create confusion, erode trust, and allow cherry-picking.
Llama3-8B-Instruct
What is the next best step for this patient? A. Ibuprofen B. Tetracycline C. Amoxicilin D. Gentamicin
Answer: Tetracycline ✗
What is the next best step for this patient? A. Tetracycline B. Amoxicilin C. Gentamicin D. Ibuprofen
Answer: Tetracycline ✗
What is the next best step for this patient? A. Ibuprofen B. Tetracycline C. Gentamicin D. Amoxicilin
Answer: Tetracycline ✗
Consistent errors can contribute to a wide spread of misinformation.

Figure 1: Different harms that are treated the same in existing evaluation. Prompt sensitivity under MCQ options shuffling. Example from the Med-HALT dataset (Pal et al., 2023).

benchmarks remain stable even under paraphrasing (Lin et al., 2022; Hong et al., 2024; Pal et al., 2023).

However, overall accuracy stability can hide the lack of consistency in individual generations. In our work, we formalize consistency in hallucination evaluation through the lens of multiplicity (Marx et al., 2020; Black et al., 2022a), and show that these benchmarks exhibit *high multiplicity*, i.e., the model's response to individual questions changes frequently based on the prompt structure. For instance, while the Med-HALT dataset (Pal et al., 2023) has an accuracy variance of less than 0.5% under changing prompt structure, it showed more than 50% multiplicity (Table 1), i.e., for more than 50% questions the model generates different facts based on the prompt structure, despite stable average accuracy of correct generations.

Leveraging this axis of consistency quantified using multiplicity, we provide a more nuanced decomposition of hallucination errors. We find that existing evaluations solely based on correctness can hide differences and underestimate the real risks of hallucinations. Consider two well-known datasets, TruthfulQA (Lin et al., 2022) and Med-HALT (Pal et al., 2023), with similar accuracies ($25 - 30\%$). However, we show that models make very distinct errors on these datasets, with TruthfulQA dominated by consistent yet factually incorrect generations, and Med-HALT dominated by randomness and inconsistency (Figure 3). Moreover, we find consistently correct generations on these datasets are far lower than their accuracies ($15 - 20\%$), highlighting the overestimation of model capabilities.

The distinction between various errors helps us classify the real harms and plays a pivotal role in shaping discussions on addressing hallucinations. We position existing detection and mitigation techniques within our framework, identifying limitations and improving our understanding of their effectiveness. We connect detection techniques with multiplicity, demonstrating that they detect consistency, not correctness (Figure 4). Thus, there is a misalignment between detection, which aims to detect consistency, and the benchmarks, which are instead designed to evaluate correctness. Finally, we move to mitigation, showing that while the introduction of components like retrieval-augmented generation (RAG) (Ram et al., 2023) reduce overall hallucination rates, these improvements can hide a new inconsistency due to prompt sensitivity of the retrieval itself (Figure 5).

Our key contributions are:

- **Prompt multiplicity in LLM hallucination evaluation:** We formalize consistency in hallucination evaluation by defining *prompt multiplicity*, leveraging existing tools from the multiplicity literature (§3.2). We highlight severe prompt multiplicity across six different benchmarks and 16 different models (from six model families), undermining the reliability of existing evaluation frameworks in quantifying the true harms of hallucinations (§4.2).
- **An improved taxonomy for benchmarking:** We propose a refined taxonomy for hallucination benchmarking by quantifying *'prompt-agnostic vs prompt-sensitive'* (Yin et al., 2024) and *'randomness'* (Venkit et al., 2024), through the lens of prompt multiplicity (§3.3). We also highlight the advantages of our framework in assessing real-world risks and illustrate several dataset-specific trends to map future progress in various domains (§4.3).
- **Hallucination detection and mitigation under prompt multiplicity:** We establish that existing detection techniques do not detect correctness, but instead detect a different axis of hallucination

evaluation, i.e., consistency (§5.1), highlighting the disconnect between methods and benchmarks. Finally, we show that mitigation techniques like RAG are also affected by prompt sensitivity, and thus introduce additional inconsistencies (§5.2).

2 Related Work

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

LLM Hallucination Benchmarks. Hallucinations in LLMs have garnered significant interest, with extensive work on categorization, evaluation, detection, and mitigation (Huang et al., 2023; Ji et al., 2023; Wang et al., 2023; Zhang et al., 2023; Tonmoy et al., 2024). Various hallucination benchmarks have been developed, with 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 for a holistic evaluation of hallucinations. We propose a new evaluation framework that incorporates consistency, and can be extended to any existing benchmark.

Prompt Sensitivity in LLMs. Prompt sensitivity studies in LLMs have revealed that even minor changes to the input or the prompt structure can impact model behaviour (Lu et al., 2022; Shi et al., 2023; Sclar et al., 2023; Voronov et al., 2024). Recent research has also heavily focused on the MCQ format, widely used in LLM evaluations, finding that changes to the order or representation of choices can also affect model accuracy (Zheng et al., 2023; Pezeshkpour & Hruschka, 2024; Alzahrani et al., 2024; Polo et al., 2024; Mizrahi et al., 2024). However, literature on prompt sensitivity in hallucinations remains limited. While Lin et al. (2022); Pal et al. (2023); Hong et al. (2024) have performed small-scale ablation studies in their work to study the impact of prompt paraphrasing on hallucination benchmarks, they found stable overall accuracy trends and thus did not explore question-level behaviour of hallucinations. 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, predictive multiplicity (Marx et al., 2020), refers to the existence of multiple models with similar overall accuracy but different individual-level predictions. We extend the notion of multiplicity to what we call *prompt multiplicity* in LLMs. Specifically, we study how competing prompt structures can yield similar benchmark accuracy while generating different individual-level answers. We use the multiplicity framing to take advantage of the existing literature.

3 Hallucinations: Incorrect Knowledge or Randomness?

In the 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). This covers a wide range of model behaviour, from ”incorrect knowledge” to ”randomness” (Venkit et al., 2024). However, hallucinations as factually incorrect knowledge embedded in the model[1] due to outdated information, flawed data sources, biases, or myths present in the training data scraped from the web (Huang et al., 2023; Lin et al., 2022) form a distinct category from hallucinations as random but plausible-sounding generations, sometimes referred to as ’confabulations’ (Millidge, 2023; Farquhar et al., 2024).

In this section, we begin by discussing the two broad categories of harm from hallucinations, emphasizing the key distinction between them, i.e., ”consistency”. As existing benchmarks do not measure consistency, to address this gap, we draw from the multiplicity literature and formalize

[1] We use *knowledge embedded in the model* to refer to prompt-agnostic knowledge (Yin et al., 2024). Despite the tension between ’knowledge in LLMs can be difficult to extract’ (Gekhman et al., 2025; Yin et al., 2024) and ’LLMs can be forced to generate any correct or incorrect fact’ (Yao et al., 2023), for the sake of understanding the potential harms of LLM hallucinations, we argue that any information consistently repeated by the model can be considered knowledge embedded in the model.

prompt multiplicity. Finally, we use this new axis of consistency to provide a more refined taxonomy for hallucination evaluation, directly mapping to several aspects of hallucinations that have been previously discussed but never explicitly quantified.

3.1 Harms from LLM Hallucinations and the Role of Consistency

The harm caused by hallucinations in language models depends on several factors, including the model's use case, the user's level of trust in the model, and their expertise, among others (Venkit et al., 2024; Elsayed, 2024; Bender et al., 2021). One way to broadly categorize these harms is by evaluating the consistency of hallucinations. To understand this distinction, consider the example in Figure 1. Two models, Llama3-8B and Llama3-8B-Instruct (Dubey et al., 2024), make the same error on the Med-HALT dataset (Pal et al., 2023). Without accounting for consistency, both errors appear identical and are labelled as ”hallucinations” in existing benchmarks. However, by testing the models multiple times with various equivalent prompts (here, shuffling the order of MCQ options), we uncover a key difference.

Llama3-8B exhibits inconsistency, i.e., it selects different answers depending on the prompt variation. This unpredictability can erode user trust, confuse even an expert working alongside the model, and introduce the risk of cherry-picking certain responses. In contrast, Llama3-8B-Instruct consistently provides the same incorrect answer. It repeatedly identifies Tetracycline as its choice, which is unfortunately the wrong antibiotic in this situation, as unlike Amoxicilin, it has known risks for pregnant women. This consistency in hallucination creates a different harm: rather than hiding uncertainty with confident generations, the model is propagating misinformation. The two categories of harm can be defined as follows.

Harms due to randomness. Hallucinations can arise when the model is uncertain about the correct answer or is confidently guessing. Such hallucinations would be likely *prompt-sensitive* (Yin et al., 2024), i.e., 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. Detecting these errors requires quantifying the uncertainty of LLM generations (Vashurin et al., 2024; Savage et al., 2024).

Harms due to incorrect knowledge embedded in the model. Hallucinations can also occur when LLMs encode incorrect or partial knowledge, misconceptions, or myths, from the training data. These can mislead users in critical contexts or contribute to a wider spread of misinformation (Venkit et al., 2024). Such hallucinations are likely *prompt-agnostic* (Yin et al., 2024), i.e., the model consistently generates the same incorrect response. These errors cannot be addressed by simply measuring uncertainty, and might require filtering unreliable training data or fact-checking the generated sentences using external knowledge.

Consistency thus plays an important role in understanding the causes, impact, and effective strategies to address hallucinations. An inconsistent hallucination stems from randomness, while a consistent one may reveal flawed data sources. Thus, by incorporating consistency into hallucination evaluation, we can develop a more nuanced understanding of these risks.

3.2 Defining Prompt Multiplicity

Literature on prompt sensitivity focuses primarily on accuracy stability (Sclar et al., 2023; Voronov et al., 2024; Mizrahi et al., 2024), which unfortunately does not capture the question-level consistency concerns discussed above. Interestingly, this exact problem lies at the heart of the field of predictive multiplicity (Marx et al., 2020; Black et al., 2022a). Predictive multiplicity refers to the existence of multiple models achieving similar accuracy, yet exhibiting distinct behaviour in individual predictions. Drawing from this, we propose *prompt multiplicity*, the idea that different prompt structures can achieve similar average hallucination accuracy, yet produce distinct behaviour for individual questions. This framework allows us to capture ”consistency” in LLM hallucination evaluation.

In our paper, we use MCQ-style benchmarks, where the goal is to select the factually correct continuation from a set of options (more details in §4.1). Each question in the benchmark $\mathbf{x}_k \in \mathbf{X}$ is first formatted using a prompt structure $\mathbf{p}^i$, which may involve modifications like prefixing demonstrations to the question, adding instructions, etc., before it is fed to a model $\mathbf{G}$. We use the notation

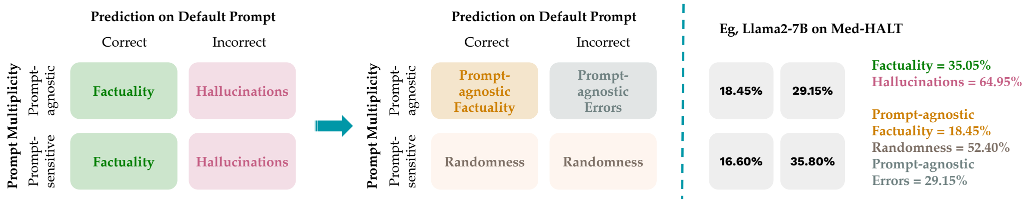


Figure 2: The mapping from existing terms like *"hallucinations"* and *"factuality"* to a more nuanced taxonomy of *"prompt-agnostic factuality"*, *"prompt-agnostic errors"*, and *"randomness"*.

$\mathbf{G}(\mathbf{p}^i(\mathbf{x}_k))$ to indicate the final MCQ choice of the model $\mathbf{G}$ for a question $\mathbf{x}_k$ with prompt structure $\mathbf{p}^i$. Building on Marx et al. (2020), we define prompt multiplicity and ambiguity under a set of prompt structures $\mathbb{P} = [\mathbf{p}^1, \mathbf{p}^2, ..., \mathbf{p}^r]$ as follows:

Definition 1 (Prompt Multiplicity)**.** *Given a model* $\mathbf{G}$*, a benchmark* $\mathbf{X}$*, and a set of prompt structures,* $\mathbb{P} = [\mathbf{p}^1, \mathbf{p}^2, ..., \mathbf{p}^r]$*, the benchmark is said to exhibit prompt multiplicity if* $\exists\ \mathbf{p}^i, \mathbf{p}^j \in \mathbb{P}$ *such that* $\mathbf{G}(\mathbf{p}^i(\mathbf{x}_k)) \neq \mathbf{G}(\mathbf{p}^j(\mathbf{x}_k))$ *for some question* $\mathbf{x}_k \in \mathbf{X}$*.*

Definition 2 (Ambiguity)**.** *Given a model* $\mathbf{G}$*, a benchmark* $\mathbf{X}$*, and a set of prompt structures,* $\mathbb{P} = [\mathbf{p}^1, \mathbf{p}^2, ..., \mathbf{p}^r]$*, ambiguity is the proportion of questions in the benchmark that can output different choices depending on the prompt structure,*

$$Ambiguity = \frac{1}{n}\sum_{k=1}^{n} \max_{\mathbf{p}^i, \mathbf{p}^j \in \mathbb{P}} \mathbb{1}[\mathbf{G}(\mathbf{p}^i(\mathbf{x}_k)) \neq \mathbf{G}(\mathbf{p}^j(\mathbf{x}_k))] \tag{1}$$

While we can use ambiguity to quantify the severity of multiplicity in a benchmark, we also need a metric to define "consistency" for each question separately. For this, we turn to Cooper et al. (2024), and define self-consistency as,

Definition 3 (Self-consistency)**.** *Given a model* $\mathbf{G}$*, a question* $\mathbf{x}_k \in \mathbf{X}$*, and a set of prompt structures,* $\mathbb{P} = [\mathbf{p}^1, \mathbf{p}^2, ..., \mathbf{p}^r]$*, self-consistency is the probability of getting the same output choice from two randomly chosen prompt structures* $\mathbf{p}^i, \mathbf{p}^j \sim \mathbb{P}$*,*

$$SC_{x_k} = 1 - Prob_{\mathbf{p}^i, \mathbf{p}^j \sim \mathbb{P}}[\mathbf{G}(\mathbf{p}^i(\mathbf{x}_k)) \neq \mathbf{G}(\mathbf{p}^j(\mathbf{x}_k))] \tag{2}$$

3.3 Mapping Prompt Multiplicity to Hallucination Evaluation

Building on the definitions above, we now introduce a new axis of evaluation in our framework, "consistency". While existing benchmarks only divide the evaluations along the axis of correctness, we argue that incorporating consistency can provide more nuance to the discussion and quantify various forms of harm. We use the self-consistency metric (Definition 3) to categorize questions along the consistency axis into prompt-sensitive and prompt-agnostic, adopted from Yin et al. (2024) and defined as follows:

Definition 4 (Prompt-sensitive)**.** *A question* $\mathbf{x}_k \in \mathbf{X}$ *is considered prompt-sensitive if its self-consistency score* SC_{x_k} *is below a given threshold* τ*,*

$$Prompt\text{-}sensitive \Leftarrow \mathbb{1}[SC_{x_k} < \tau] \tag{3}$$

Definition 5 (Prompt-agnostic)**.** *A question* $\mathbf{x}_k \in \mathbf{X}$ *is considered prompt-agnostic if its self-consistency score* SC_{x_k} *is equal to or above a given threshold* τ*,*

$$Prompt\text{-}agnostic \Leftarrow \mathbb{1}[SC_{x_k} \geq \tau] \tag{4}$$

We use $\tau = 0.8$ throughout the paper, unless otherwise specified.

A refined evaluation terminology: We argue that factually correct generations that are prompt-sensitive, despite being accurate for the default benchmark prompt structure, should be treated with the same level of caution as factually incorrect prompt-sensitive generations. In other words, if the generation of factually incorrect information is highly dependent on the prompt structure, it should be categorized as **randomness**, irrespective of whether this randomness happens to produce the correct output for the default prompt structure of the benchmark, as it possesses the same risk of generating a factually incorrect sentence for a different prompt structure. Moreover, 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 shown in Figure 2.

4 Prompt Multiplicity in LLM Hallucination Benchmarks

We now turn to the empirical results. We first highlight severe multiplicity in hallucination benchmarks, despite stable accuracy. Next, we map the evaluation to our framework, revealing how existing benchmarks underestimate potential real-world harm caused by hallucinations. We conclude with dataset-specific trends and takeaways.

4.1 Experiment Setup

Datasets. We use the following factuality hallucination benchmarks: Wiki-FACTOR (Muhlgay et al., 2024), Med-HALT (Pal et al., 2023), TruthfulQA (Lin et al., 2022), TrueFalse (Azaria & Mitchell, 2023), CommonsenseQA (Talmor et al., 2019), and FEVER (Thorne et al., 2018). Details on each dataset are provided in the appendix (§A.1). We primarily focus on TruthfulQA, Wiki-FACTOR, and Med-HALT in the main paper, while other results are delegated to the appendix (§C). We use the perplexity-based evaluation by Muhlgay et al. (2024), where the LLM chooses the best MCQ option based on the length-normalized perplexity.

We decided to stick with only MCQ-style benchmarks for our study, since freeform generation requires additional automated methods to evaluate generated outputs, such as an LLM judge (Lin et al., 2022; Li et al., 2023; Dong et al., 2024)–which can introduce its own errors, biases, and multiplicity (Li et al., 2024; Ye et al.; Panickssery et al., 2024).

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. We simulate prompt variations in a structured manner, wherever possible, as they can be applied uniformly across the dataset. This includes shuffling the order of demonstrations (Lu et al., 2022) (TruthfulQA, FEVER, TrueFalse) or shuffling the order of MCQ options (Zheng et al., 2023; Pezeshkpour & Hruschka, 2024) (Med-HALT, CommonsenseQA). However, the Wiki-FACTOR benchmark does not provide any opportunity for structured variations. Instead, we turn to automated paraphrasing and use a fine-tuned T5 model (Raffel et al., 2020) trained on a paraphrase dataset from ChatGPT (Vorobev & Kuznetsov, 2023a;b). More details on prompt variations are in the appendix (§A.1).

4.2 Hallucination Benchmarks show High Multiplicity and Underestimate Risks

Despite low accuracy variance, LLM hallucination benchmarks exhibit severe prompt multiplicity. To illustrate this, we collect the average accuracy, standard deviation, and ambiguity, across different variations, in Table 1 (only the biggest models from each family are shown, the rest are in Table 2). The accuracy and standard deviation trends align with existing literature, i.e., low variance in accuracy. This explains why previous research has largely overlooked prompt sensitivity. However, the ambiguity scores tell a more compelling story, revealing significant prompt multiplicity within these benchmarks. For instance, LLama2-13B-Chat on Med-HALT achieves $\sim 35\%$ accuracy with a standard deviation of only 0.23%, potentially signalling stability. Yet, its ambiguity score is $\sim 60\%$,

	TruthfulQA		Wiki-FACTOR		Med-HALT	
	Accuracy (%)	Ambiguity (%)	Accuracy (%)	Ambiguity (%)	Accuracy (%)	Ambiguity (%)
GPTNeoX-20B	20.09$_{\pm 1.26}$	22.89	45.74$_{\pm 1.38}$	41.95	28.98$_{\pm 0.43}$	52.26
Pythia-12B	20.31$_{\pm 0.89}$	19.58	42.90$_{\pm 0.96}$	38.61	28.18$_{\pm 0.39}$	50.07
Bloom-7.1B	23.18$_{\pm 1.32}$	18.36	35.14$_{\pm 0.78}$	37.27	28.51$_{\pm 0.62}$	56.40
Llama2-13B-C	32.97$_{\pm 1.03}$	21.79	50.32$_{\pm 1.06}$	47.09	34.84$_{\pm 0.42}$	60.54
Llama3-8B-I	39.34$_{\pm 0.75}$	17.14	48.39$_{\pm 1.19}$	42.32	34.55$_{\pm 0.23}$	31.03
OPT-30B	22.55$_{\pm 0.90}$	23.38	43.58$_{\pm 0.91}$	41.35	28.32$_{\pm 0.42}$	50.00

Table 1: High ambiguity across a wide range of model families and benchmarks.

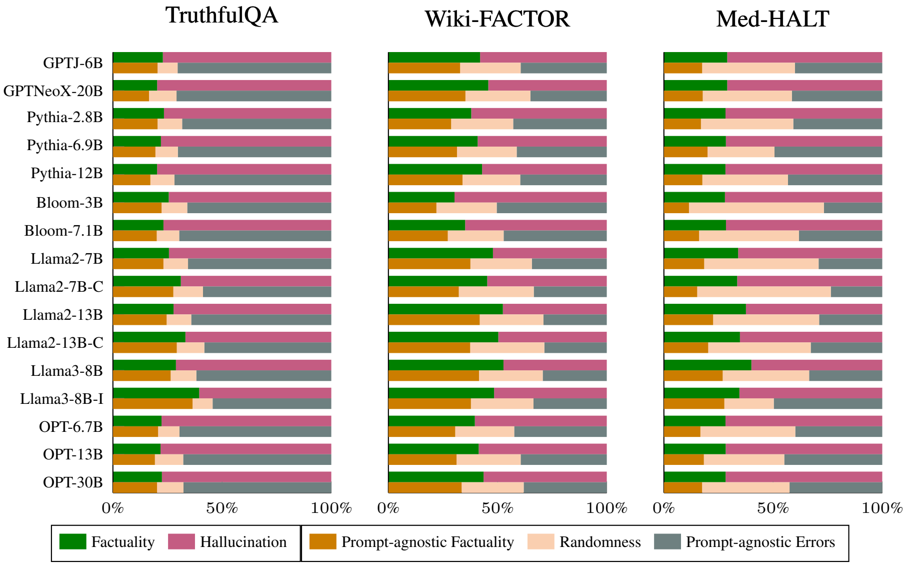


Figure 3: Existing LLM hallucination evaluation terminology vs our framework.

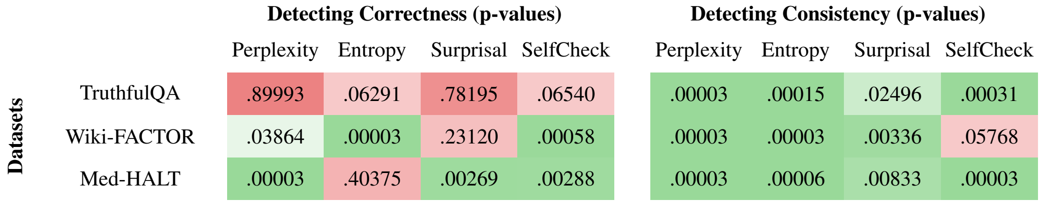
i.e., the model changes the generated fact for $\sim 60\%$ of the dataset simply based on the prompt structure.

Another intriguing result in Table 1 is the comparison between Llama2-13B-Chat and Llama3-8B-Instruct on the Med-HALT dataset. Even though both models have similar accuracies, the ambiguities signal vast differences in the types of errors and underlying behaviour, i.e., significantly high ”randomness” for the former. To understand this difference, we map the evaluations to our framework and provide more nuanced results in Figure 3. 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. Thus, the true extent of potential harms—both prompt-agnostic errors and randomness together—is greater than what is captured by ”hallucination” in existing benchmarks.

4.3 Dataset-specific Trends

We next turn to some dataset-specific trends to highlight insights for future research.

TruthfulQA. The TruthfulQA dataset was designed to capture various misconceptions and myths, with carefully crafted adversarial prompts (Lin et al., 2022). Given this construction, it is no surprise that most errors in TruthfulQA are prompt-agnostic, while only a small fraction are attributed to randomness. This makes TruthfulQA an excellent example of a benchmark that can highlight flawed data sources used to train a model. One particularly noteworthy result is the randomness rate of ~ 10–12% across all models, despite their varying accuracy levels. We believe this could be



Datasets	Detecting Correctness (p-values)				Detecting Consistency (p-values)			
	Perplexity	Entropy	Surprisal	SelfCheck	Perplexity	Entropy	Surprisal	SelfCheck
TruthfulQA	.89993	.06291	.78195	.06540	.00003	.00015	.02496	.00031
Wiki-FACTOR	.03864	.00003	.23120	.00058	.00003	.00003	.00336	.05768
Med-HALT	.00003	.40375	.00269	.00288	.00003	.00006	.00833	.00003

Figure 4: Ease of differentiating based on correctness vs consistency, using detection scores.

due to ambiguous questions which require the model to generate 'I have no comment', leading to consistency issues for TruthfulQA.

Wiki-FACTOR. The Wiki-FACTOR dataset is constructed using Wikipedia articles, with automatically generated semantically close but adversarial multiple-choice options, thereby increasing the percentage of data points showing randomness (Muhlgay et al., 2024). 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 entrance exams around the world (Pal et al., 2023). We see that it exhibits a significantly higher percentage of randomness compared to other benchmarks. While TruthfulQA demonstrates concerns of unreliable data in training and would require leveraging an external knowledge source to help mitigate errors, Med-HALT represents the other extreme where analyzing the model's uncertainty can be an effective way to detect potential hallucinations.

5 Hallucination Detection and Mitigation

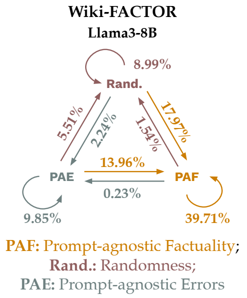
We extend our discussion to existing hallucination detection and mitigation techniques. We provide two key observations: (a) detection techniques primarily differentiate between prompt-agnostic and prompt-sensitive generations rather than identifying factual or hallucinated outputs, and (b) mitigation techniques that rely on knowledge retrieval are themselves influenced by prompt sensitivity, thus introducing additional inconsistencies.

5.1 Detecting Consistency not Correctness

We start by studying several hallucination detection techniques under our framework. Specifically, we test: (a) **Perplexity**, a simple baseline for hallucination detection (Ren et al., 2022; Chen et al.); (b) **Entropy**, which addresses some of Perplexity's shortcomings (Vashurin et al., 2024); (c) **Surprisal**, based on claims of surprisal using embedding similarity by Duan et al. (2024); and (d) **SelfCheck**, which adapts the intuition behind SelfCheckGPT (Manakul et al., 2023). Each technique produces a final score that can be used to classify the output as a hallucination or not. All detection scores are calculated for the default prompt structure. More details on detection techniques are provided in the appendix (§A.2).

Once we compute the detection scores, we average them separately over all answers labelled as 'correct' and 'incorrect' (for the default prompt structure). This aims to capture the distinction along the axis of correctness. We repeat this across all 16 models in our setup, thus creating a set of 16 average scores for correct answers and the same for incorrect answers. We then test the following hypothesis using the Wilcoxon test: *Assuming the differences in average detection scores for correct and incorrect answers are symmetric around a central value, this central value is zero.* The test aims to determine whether these techniques can easily separate the benchmark into correct and incorrect answers. Finally, we repeat the analysis along the axis of consistency instead of correctness, i.e., averaging detection scores separately for answers labelled as 'prompt-agnostic' and 'prompt-sensitive'.

Note that the detection scores could have more predictive power than measured here, as we're simplifying the distribution into an average. The objective of this test is only to highlight the inherent alignment of detection with consistency, instead of correctness.



	Ambiguity over Retrieved Docs					
	Wiki-FACTOR			FEVER		
	PAF	PAE	Rand.	PAF	PAE	Rand.
GPTNeoX-20B	.27	.45	.70	.87	.91	.94
Pythia-12B	.26	.42	.75	.81	.89	.93
Bloom-7.1B	.25	.44	.73	.87	.90	.95
Llama2-13B-C	.28	.45	.69	.80	.85	.91
Llama3-8B-I	.27	.45	.73	.88	.90	.93
OPT-30B	.26	.43	.74	.87	.87	.95

Figure 5: **(Left)** Change in category for questions in the benchmark after the addition of RAG. **(Right)** Randomness under RAG is dominated by the inconsistency of retrieval.

The p-values for the hypothesis tests, corresponding to each combination of benchmark and detection technique, are presented in Figure 4. The results reveal a striking pattern: the p-values are higher when distinguishing based on correctness, compared to consistency, where the p-values are notably low. This shows that the detection techniques are primarily capturing consistency and not correctness, i.e., they are not detecting hallucinations, but instead randomness. This isn't entirely surprising, as many of them are fundamentally designed to detect uncertainty. However, it emphasizes the disconnect between the benchmarks, based on correctness, and the detection techniques, based instead on consistency.

5.2 Prompt Sensitivity and Knowledge-Retrieval

Most existing hallucination mitigation techniques rely on incorporating external knowledge, typically guided by a retrieval mechanism to find relevant information (Ji et al., 2023). Beyond overall improvements, we find several intriguing shifts in model behaviour during mitigation. Many questions that originally exhibited prompt-agnostic errors instead show randomness, while a smaller portion also shows the opposite trend. Upon deeper investigation, we find that the new component in the pipeline, retrieval, is itself sensitive to prompt variations, introducing an additional layer of inconsistency into the system.

Retrieval-Augmented Generation (RAG) Setup. Before jumping into the results, we clarify key details of our RAG setup. We use the in-context retrieval augmentation technique proposed by Ram et al. (2023), using BM25 (Robertson et al., 2009), a sparse word-based retriever. We also leverage the same Wikipedia corpus (Ram et al., 2023) and focus on two datasets: Wiki-FACTOR and FEVER, both originally constructed using facts extracted from Wikipedia. To study the sensitivity of RAG, we introduce paraphrasing variations in the FEVER dataset similar to Wiki-FACTOR, unlike the rest of the paper where we only shuffled demonstrations. Further details about the RAG setup can be found in the appendix (§A.3).

Results. We first study the impact of RAG on our evaluations. Figure 5 (Left) captures the movement of all questions in the dataset, as they shift from their original category without RAG to a new category with RAG. The self-loops, thus, indicate questions that remain in the same category. Unsurprisingly, we observe a significant shift towards prompt-agnostic factuality (PAF), i.e., fewer hallucinations. However, a more intriguing result is the redistribution of errors: questions transitioning between prompt-agnostic errors (PAE) and randomness, with an overall flow toward randomness. We argue that this stems from the RAG itself being highly sensitive to prompt changes, thus introducing randomness.

To validate this, we conduct the following test. We extend the idea of 'ambiguity' (Definition 2) to retrieved documents, defining *ambiguity over retrieved documents* as the proportion of questions in the benchmark that can retrieve a different document depending on the prompt structure. Figure 5 (Right) presents these scores across different categories with RAG. The results show that questions exhibiting randomness have significantly higher ambiguity over retrieved documents than others, i.e., the retrieval of different documents for different paraphrasing of the same prompts creates inconsistency. Thus, while RAG can help mitigate factually incorrect generations, it also introduces its own instability into the pipeline. These results further emphasize the value of evaluating hallucinations within our framework.

6 Conclusion and Future Work

In this paper, we proposed an improved framework for evaluating hallucinations, emphasizing the role of consistency in distinguishing different hallucination harms and informing appropriate detection and mitigation strategies. While our work establishes a strong foundation, several open questions remain. A key challenge is extending to benchmarks that allow freeform generation. Although our fundamental arguments will generalize, the freedom of unconstrained generation introduces new complexities—such as inconsistencies in evaluation setups that rely on LLM judges and redefining self-consistency in the context of language rather than discrete MCQ options. Additionally, future work on different types of prompt variations is also needed. In conclusion, our framework provides a more nuanced approach to hallucination evaluation, allowing the exploration of more effective solutions.

Acknowledgments

We thank Florian Carichon and Khaoula Chehbouni for their valuable feedback and comments on earlier drafts of the paper. Funding support for project activities has been partially provided by the Canada CIFAR AI Chair, FRQNT scholarship, and NSERC discovery award. We also express our gratitude to Compute Canada and Mila clusters for their support in providing facilities for our evaluations.

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.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

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.

Hanyu Duan, Yi Yang, and Kar Yan Tam. Do llms know about hallucination? an empirical investigation of llm's hidden states. *arXiv preprint arXiv:2402.09733*, 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.

Hassan Elsayed. The impact of hallucinated information in large language models on student learning outcomes: A critical examination of misinformation risks in ai-assisted education. *Northern Reviews on Algorithmic Research, Theoretical Computation, and Complexity*, 9(8):11–23, 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.

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

Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpector, Jonathan Herzig, and Roi Reichart. Inside-out: Hidden factual knowledge in llms. *arXiv preprint arXiv:2503.15299*, 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.

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.

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*, 2024.

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.

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.

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

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.

Beren Millidge. Llms confabulate not hallucinate. *Online verfügbar unter https://www. beren. io/2023-03-19-LLMs-confabulate-not-hallucinate*, 2023.

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.

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.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023.

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.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H Chen. Large language model uncertainty measurement and calibration for medical diagnosis and treatment. *medRxiv*, pp. 2024–06, 2024.

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.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Akim Tsvigun, Daniil Vasilev, Rui Xing, Abdelrahman Boda Sadallah, Kirill Grishchenkov, Sergey Petrakov, et al. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *arXiv preprint arXiv:2406.15627*, 2024.

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.

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.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*, 2023.

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.

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 Additional Details on Experiment Setup

A.1 Datasets and Prompt Variations

TruthfulQA. We use the MCQ task from TruthfulQA, and 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 50 different prompt variations, i.e., 50 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 paraphraser 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 Reasoning Hallucination Test (RHT) of the Med-HALT dataset, and the original instruction prompt used by the authors (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 20 different prompt variations, i.e., 20 different shufflings of the MCQ options.

CommonsenseQA. We use the development set of the dataset and 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 50 different prompt variations, i.e., 50 different random choices of the 16-shot demonstrations.

FEVER. We use the shared task development set of the dataset and 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 50 different prompt variations, i.e., 50 different random choices of the 16-shot demonstrations.

TrueFalse. We use all topics combined from the TrueFalse dataset and perform a 16-shot evaluation of the 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 no separate train set to sample from and hence the demonstrations are sampled from the evaluation dataset itself. Thus, the sampled demonstration in certain cases might even contain the final question. We measure all metrics across 50 different prompt variations, i.e., 50 different random choices of the 16-shot demonstrations.

A.2 Detection Techniques

We provide details on the scores calculated for all four detection techniques in our paper.

Perplexity. We simply use the length-normalized perplexity score of the best option.

Entropy. We treat the length-normalized perplexity scores of all options as scores of a classification problem, and measure the entropy of the prediction. In other words, we first normalize the perplexity scores across all options to turn them into probabilities, and then measure the entropy of the probabilities across all options.

Surprisal. We measure the cosine similarity between the representation of the question and the representation of the chosen option appended to the question, as suggested by Duan et al. (2024). Here, the representation of a sentence is the output of the final transformer layer of the model, i.e., the final hidden state, for the last token.

SelfCheck. Adopting from Manakul et al. (2023), we simply append the chosen option to the question, followed by a follow-up question 'Is the above statement correct?', and check for the probability of the next token being 'Yes'.

A.3 Mitigation Setup Details

RAG Setup. We use the open-source code provided by Ram et al. (2023)[2].

[2] `https://github.com/AI21Labs/in-context-ralm`

	TruthfulQA		**Wiki-FACTOR**		**Med-HALT**	
	Accuracy (%)	Ambiguity (%)	Accuracy (%)	Ambiguity (%)	Accuracy (%)	Ambiguity (%)
GPTJ-6B	$22.86_{\pm0.71}$	13.83	$41.98_{\pm0.90}$	39.65	$28.99_{\pm0.77}$	50.17
GPTNeoX-20B	$20.09_{\pm1.26}$	22.89	$45.74_{\pm1.38}$	41.95	$28.98_{\pm0.43}$	52.26
Pythia-2.8B	$23.37_{\pm1.20}$	16.65	$37.93_{\pm0.84}$	40.21	$28.21_{\pm0.70}$	49.78
Pythia-6.9B	$21.99_{\pm1.19}$	13.95	$40.87_{\pm0.89}$	39.08	$28.39_{\pm0.42}$	37.63
Pythia-12B	$20.31_{\pm0.89}$	19.58	$42.90_{\pm0.96}$	38.61	$28.18_{\pm0.39}$	50.07
Bloom-3B	$25.56_{\pm1.18}$	16.16	$30.27_{\pm0.83}$	38.58	$27.95_{\pm1.42}$	70.07
Bloom-7.1B	$23.18_{\pm1.32}$	18.36	$35.14_{\pm0.78}$	37.27	$28.51_{\pm0.62}$	56.40
Llama2-7B	$25.65_{\pm0.73}$	16.16	$47.87_{\pm1.32}$	38.81	$34.00_{\pm0.65}$	61.79
Llama2-7B-C	$31.11_{\pm0.79}$	19.34	$45.25_{\pm1.05}$	47.70	$33.56_{\pm1.15}$	70.14
Llama2-13B	$27.76_{\pm0.66}$	17.26	$52.41_{\pm1.52}$	41.08	$37.57_{\pm0.21}$	58.00
Llama2-13B-C	$32.97_{\pm1.03}$	21.79	$50.32_{\pm1.06}$	47.09	$34.84_{\pm0.42}$	60.54
Llama3-8B	$28.85_{\pm1.16}$	18.48	$52.69_{\pm1.54}$	40.25	$40.06_{\pm0.61}$	48.28
Llama3-8B-I	$39.34_{\pm0.75}$	17.14	$48.39_{\pm1.19}$	42.32	$34.55_{\pm0.23}$	31.03
OPT-6.7B	$22.26_{\pm0.92}$	15.42	$39.58_{\pm1.04}$	38.81	$28.20_{\pm0.78}$	51.22
OPT-13B	$21.81_{\pm1.11}$	19.71	$41.34_{\pm1.04}$	42.59	$28.30_{\pm0.53}$	43.86
OPT-30B	$22.55_{\pm0.90}$	23.38	$43.58_{\pm0.91}$	41.35	$28.32_{\pm0.42}$	50.00

Table 2: Extended results across all models of Table 1.

FEVER Dataset. The FEVER dataset contains several demonstrations that are shuffled to create prompt variations, as mentioned above. However, for the mitigation portion of the paper, we wanted to highlight the impact of prompt paraphrasing on the retrieval component, and hence, we changed the way prompt variations were created for FEVER. While we still provide the demonstrations, we do not shuffle them and instead paraphrase the question the same way as we did for the Wiki-FACTOR dataset, using the same T5-based paraphrase. Similar to Wiki-FACTOR, we create 10 variations, i.e., 10 different paraphrases of each prompt.

B Extended Results for Tables in the Main Paper

Several results in the main text were reported only for a few models, and we extend the rest of the results here. Extended results for Table 1 are present in Table 2 and extended results for Figure 5 (Right) are present in Table 3. The trends in these models are still similar to the trends in the main paper.

C Results on CommonsenseQA, FEVER, and TrueFalse

Additional results on CommonsenseQA, FEVER, and TrueFalse datasets are in Table 4 and Figure 6. 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.

	Ambiguity over Retrieved Docs					
	Wiki-FACTOR			**FEVER**		
	PAF	PAE	Rand.	PAF	PAE	Rand.
GPT-J-6B	.26	.43	.74	.86	.87	.95
GPTNeoX-20B	.27	.45	.70	.87	.91	.94
Pythia-2.8B	.25	.45	.75	.87	.88	.95
Pythia-6.9B	.26	.45	.74	.87	.89	.95
Pythia-12B	.26	.42	.75	.81	.89	.93
Bloom-3B	.24	.45	.72	.92	.91	.88
Bloom-7.1B	.25	.44	.73	.87	.90	.95
Llama-2-7B	.28	.44	.74	.90	.89	.92
Llama-2-7B-Chat	.27	.44	.70	.90	.89	.92
Llama-2-13B	.29	.44	.73	.90	.90	.91
Llama-2-13B-C	.28	.45	.69	.80	.85	.91
Llama-3-8B	.29	.46	.70	.89	.91	.94
Llama-3-8B-I	.27	.45	.73	.88	.90	.93
OPT-6.7B	.26	.44	.74	.89	.89	.93
OPT-13B	.27	.42	.72	.92	.90	.89
OPT-30B	.26	.43	.74	.87	.87	.95

Table 3: Extended results across all models of Figure 5 (Right).

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

Table 4: Additional results for ambiguity scores.

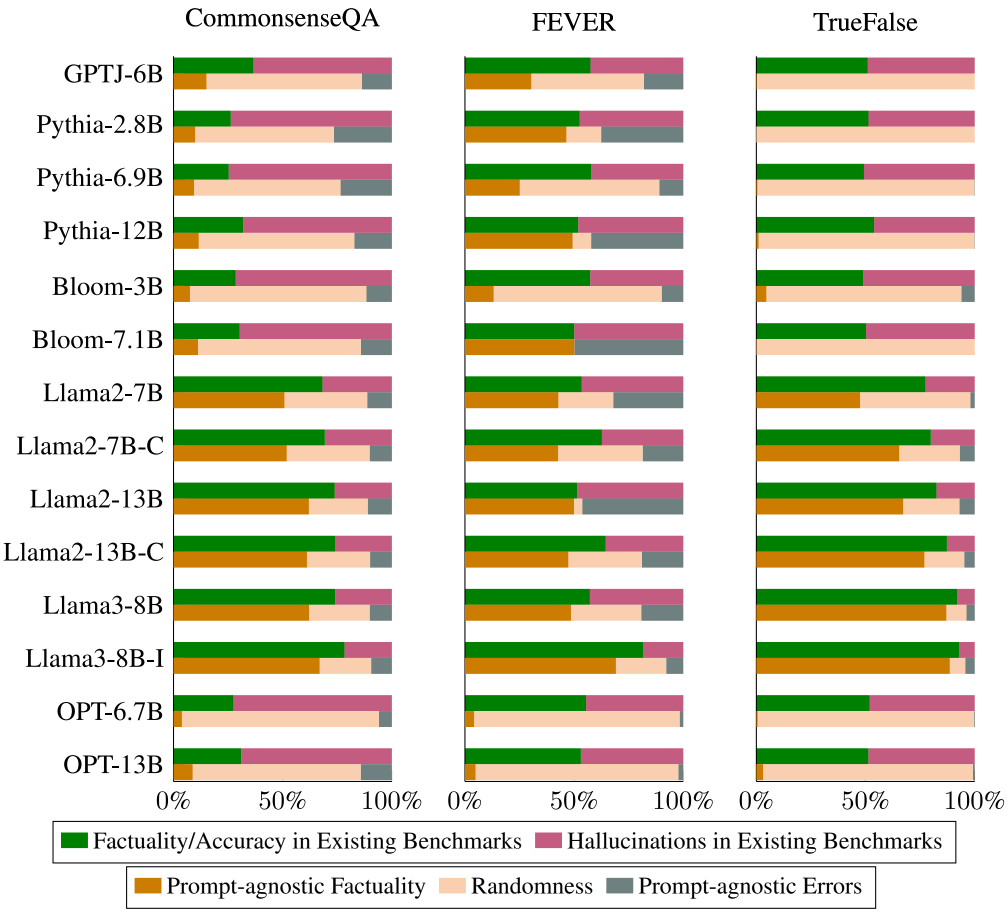


Figure 6: Additional LLM hallucination benchmark results under our new framework.