Rethinking Hallucinations: Correctness, Consistency, and Prompt Multiplicity

Prakhar Ganesh [1 2] Reza Shokri [3] Golnoosh Farnadi [1 2]

[1]McGill University [2]Mila [3]National University of Singapore

Hallucinations: Incorrect Knowledge or Randomness?

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 ✗

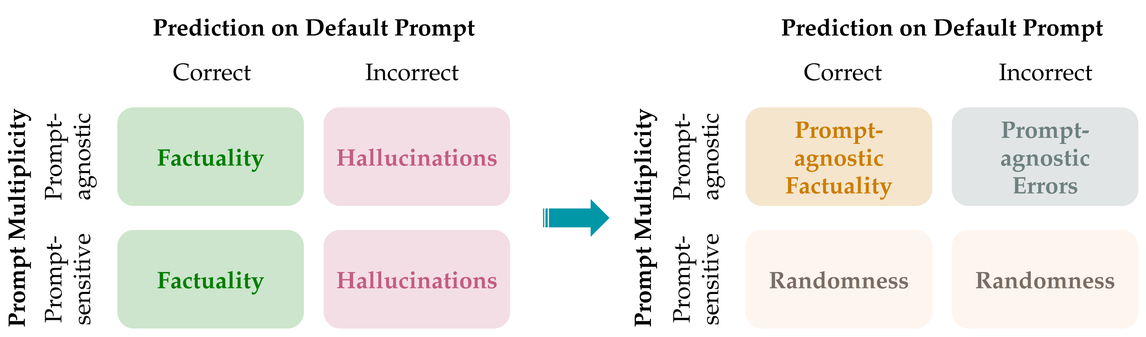
Consistent errors can contribute to a wide spread of misinformation.

Different harms are treated the same in existing hallucination evaluation!

Mapping Evaluations

Factually correct generations that are prompt-sensitive, despite being correct for the default prompt, should be categorized as *randomness*.

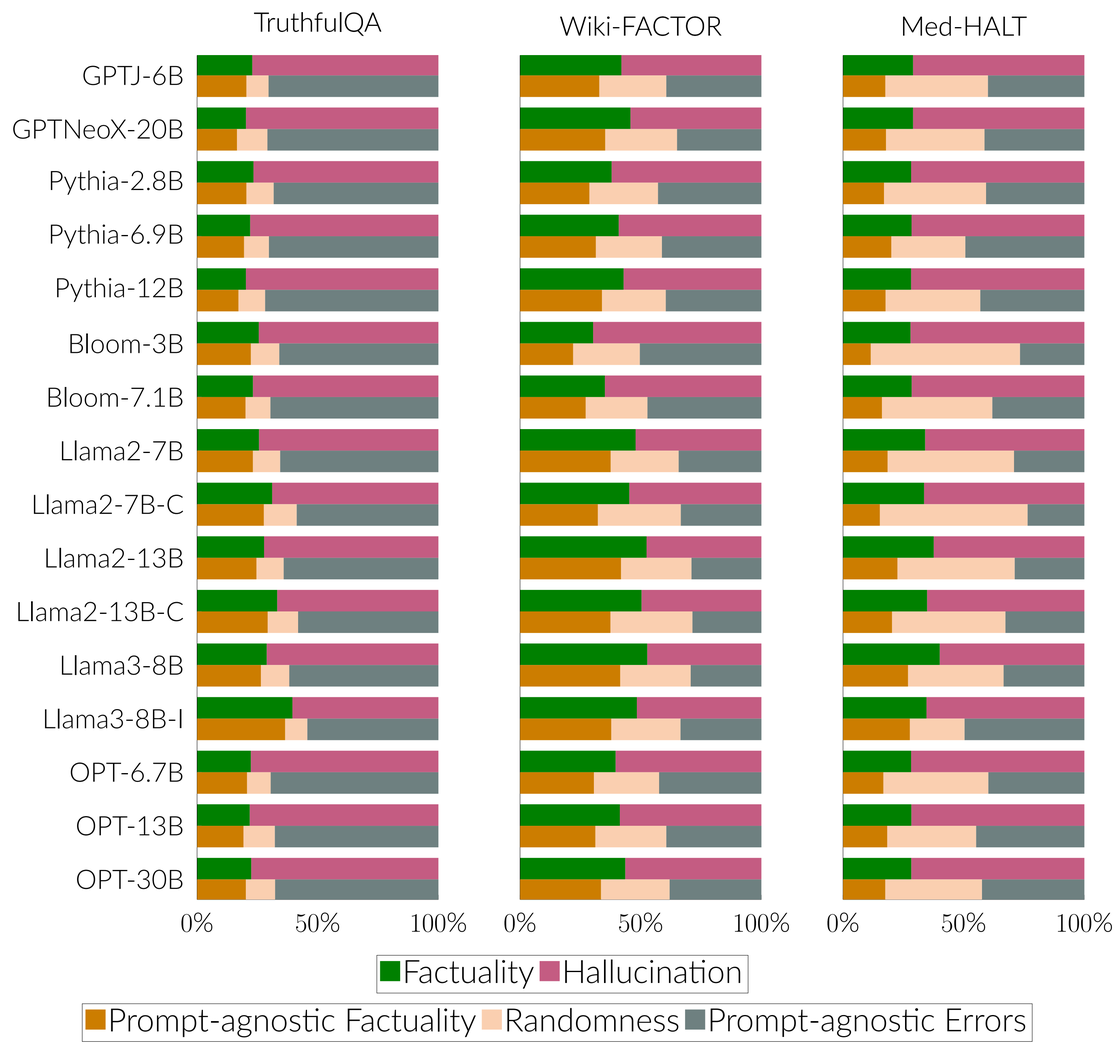
We also use the term *prompt-agnostic factuality* and *prompt-agnostic errors* to describe prompt-agnostic generations.



Empirical Results

Mapping the evaluations to our framework. Answers that were originally "factual" overstate correct facts that a model can generate consistently, i.e., prompt-agnostic factuality.

Thus, the true extent of potential harm is greater than what is captured by "hallucination" in existing benchmarks.



Detecting Consistency not Correctness

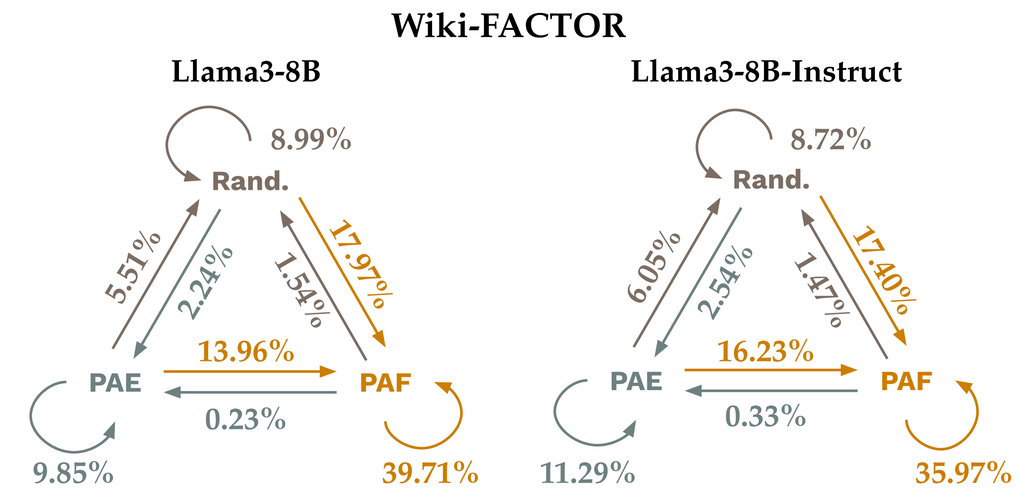
Detection techniques primarily capture consistency not correctness, i.e., they are not detecting hallucinations, but instead randomness!

	Detecting Correctness (p-values)				Detecting Consistency (p-values)			
	Perp.	Ent.	Surp.	SelfC.	Perp.	Ent.	Surp.	SelfC.
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

Datasets

Inconsistencies in Knowledge Retrieval

Beyond overall improvements, we find a redistribution of errors during mitigation using knowledge-retrieval. Questions that exhibited prompt-agnostic errors instead showed randomness, while a smaller portion followed the opposite trend. The retrieval itself is highly sensitive to prompt changes, thus introducing randomness.



PAF: Prompt-agnostic Factuality;
Rand.: Randomness; **PAE:** Prompt-agnostic Errors

Rethinking Hallucination Evaluation

- 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.
- A key challenge remains: extending our framework beyond the MCQ setting. The freedom of unconstrained generation introduces new complexities—such as inconsistencies in evaluation setups that rely on LLM judges.

prakhar.ganesh@mila.quebec