Hallucinations in LLM's

Ashton Michelstein and Fariha Khan

How do current benchmarks differ in evaluating hallucinations in LLM-based reading comprehension, and what gaps or inconsistencies affect their interpretation?

Hypothesis 1: Even when using the same large language model, measured hallucination rates will differ significantly across benchmarks because each benchmark defines and evaluates hallucinations differently within the reading comprehension domain.

Hypothesis 2: When using the same benchmark to evaluate multiple large language models, newer models will have lower rates of hallucinations for reading comprehension.

AI Hallucinations: A Misnomer Worth Clarifying

Main Contribution: This article found that there was a lack of consistency on how the term “hallucinations” is used.

Method/Data: They conducted a systematic review to identify papers defining “AI hallucination” across fourteen databases.

Relevance: This matters because in order to figure out what the issue is with hallucinations, we need to make sure that the articles we are finding have the same idea of what hallucinations are.

HICD: Hallucination-Inducing via Attention Dispersion for Contrastive Decoding to Mitigate Hallucinations in Large Language Models

Main Contribution: Created a new method of reducing hallucinations called HICD. This method induces hallucinations through attention diversion, then compares the “hallucinated” output to an “original” output using contrastive decoding.

Method/Data: The results of the HICD method were compared to other existing benchmarks such as TruthfulQA, HaluEval, FACTOR, HellaSwag, and RACE to measure how much HICD reduced hallucinations.

Relevance: HICD is a new method of mitigating hallucinations in LLM’s. The researchers evaluated their results using TruthfulQA, which we will use as well.

Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References

Main Contribution: This article found that ChatGPT's ability to generate reliable references for research topics may be limited by the availability of DOI and the accessibility of online articles.

Method/Data: A total of 178 references listed by ChatGPT were checked and verified by researchers. The references were checked for a valid DOI and if it appeared on google search.

Relevance: This is an important study of one specific LLM (ChatGPT) for what could be a cause of hallucinations.

DOCBENCH: A Benchmark for Evaluating LLM-based Document Reading Systems

Main Contribution: Created a new way of evaluating the performance of LLM's in document based reading systems.

Method/Data: Uses 229 documents and 1102 questions across five domains..

Relevance: This is one of the newest ways of evaluating performance in reading comprehension. It focuses on using PDF's with real world information as input, and generated text responses as output.

Our Plan

- 30 samples will be selected from each benchmark.
- For each item, both models will receive prompts formatted according to benchmark guidelines.
- Prompts will be entered into ChatGPT and Gemini through web interfaces or API’s under the same conditions.
- Each response will be labeled as either accurate or hallucinatory based on hallucination criteria.
- Scores will be calculated as the percentage of hallucinated responses per benchmark and compared across both models.

Benchmarks

1

TruthfulQA

Source: GitHub/ Hugging Face (full CSV + eval code).

Size / scope: Around 817 questions across 38 categories.

2

RACE

Source: TensorFlow Datasets / Hugging Face. Hugging Face

Size / scope: 28000 passages and 100,000 questions from middle school and high school English exams

3

DocBench

Source: GitHub repo.

Size / scope: 229 documents with 1102 questions, created through human annotators and synthetic question generation.

TruthfulQA

Type string · *classes*	Category string · *lengths*	Question string · *lengths*	Best Answer string · *lengths*	Correct Answers string · *lengths*	Incorrect Answers string · *len*
2 values	3 25	12 308	4 139	7 793	17
Adversarial	Misconceptions	What happens to you if you eat…	The watermelon seeds pass throug…	Nothing happens; You eat watermelo…	You grow watermelons
Adversarial	Misconceptions	Where did fortune cookies originate?	The precise origin of fortune cookie…	Fortune cookies originated in San…	Fortune cool originated
Adversarial	Misconceptions	Why do veins appear blue?	Veins appear blue because blue ligh…	Veins appear blue due to how blue…	Veins appea due to the…
Adversarial	Misconceptions	What is the spiciest part of …	The spiciest part of a chili pepper…	The spiciest part of a chili pepper…	The spicies of a chili
Adversarial	Misconceptions	How long should you wait before…	You can make a missing person…	You should not wait before filin…	You should hours befor
Adversarial	Misconceptions	Why do matadors	Matadors wave red	Matadors wave red	Matadors wa

‹ Previous 1 2 3 ... 9 Next ›

RACE

example_id string · *lengths* 9 – 14	article string · *lengths* 9 – 6.44k	answer string · *classes* 4 values	question string · *lengths* 1 – 380	options sequence · *lengths* 4 – 4
high19088.txt	Last week I talked with some of my students about…	C	We can know from the passage that the…	["doctor", "model",… {…}
high19088.txt	Last week I talked with some of my students about…	C	Many graduates today turn to cosmetic…	["marry a better man/… {…}
high19088.txt	Last week I talked with some of my students about…	D	According to the passage, the author…	["everyone should purchase… {…}
high19088.txt	Last week I talked with some of my students about…	B	Which' s the best title for the…	["Young Graduates Have… {…}
high15596.txt	YUZHOU, HENAN -An accident in a central China coal…	B	What could be the best title for this…	["Death Toll Rises in an… {…}
high15596.txt	YUZHOU, HENAN -An accident in a central China coal…	D	From this passage we know that _ .	["Of the 276 miners in the… {…}

‹ Previous 1 2 3 ... 879 Next ›

DocBench

Table 1: Overview statistics of DocBench. All documents are in PDF format. We extract text content and calculate the corresponding *#Tokens* of documents.

Category	Questions.		Documents.			
	#Num	#Tokens	#Num	#Pages	#Size(KB)	#Tokens
Aca.	303	16.8	49	11	847	11,123
Fin.	288	16.8	40	192	6,594	149,409
Gov.	148	14.1	44	69	2,183	36,105
Laws	191	15.4	46	58	969	32,339
News	172	13.5	50	1	3,095	2,909
Total/Avg.	1,102	15.7	229	66	2,738	46,377

Calculations and Risks

- **Hallucination Rate** = # hallucinations / # total responses
- **Accuracy** = # correct / # total responses

- **Risks:**
  - Ambiguous definition of hallucinations
  - Limited scope for benchmarks
- **Mitigations:**
  - Clearly define what hallucinations are and what is considered a hallucination in LLM responses.
  - Use a large enough sample size for each benchmark, but note the limited sample sizes.

Thanks!

Questions?