



Hallucinations in LLM's

Ashton Michelstein and Fariha Khan

Recap:

Research Question: 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.

Models

ChatGPT-4

- This is an older chat gpt model, publicly released on March 14, 2023
- Capable of reading from documents
- Faster text generation
- Had to be an old model still capable of reading documents

ChatGPT-4o

- This is a newer model, publicly released on May 13, 2024
- Capable of reading from documents
- Longer reasoning

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.

Method

- 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-4o and ChatGPT-4 through API calls using the code from the benchmark repos.
- Each response will be labeled as correct, incorrect, or hallucinatory based on the benchmark's hallucination criteria.
- Scores will be calculated as the percentage of hallucinated responses per benchmark and compared across both models.

Current Progress

Finished testings:

- DocBench
 - ChatGPT-4o
 - ChatGPT-4
- RACE
 - ChatGPT-4o
 - ChatGPT-4

In progress:

- TruthfulQA
 - ChatGPT-4o
 - ChatGPT-4

Our Findings

	DocBench	RACE	TruthfulQA
ChatGPT-4o	24/30 Correct 6/30 Incorrect <ul style="list-style-type: none">• 4/30 Hallucinatory	28/30 Correct 2/30 Incorrect <ul style="list-style-type: none">• 1/30 Hallucinatory	In Progress
ChatGPT-4	24/30 Correct 6/30 Incorrect <ul style="list-style-type: none">• 3/30 Hallucinatory	27/30 Correct <ul style="list-style-type: none">• 3/30 Hallucinatory 3/30 Incorrect	In Progress

DocBench

```
[
  {
    "question": "What method is used by the model to aggregate individual article credibility scores for a single claim?",
    "answer": "The model averages the per-article credibility scores from M reporting articles to generate the overall credibility score for the claim.",
    "type": "text-only",
    "evidence": "\"Therefore, once we have the per-article credibility scores from our model, we take an average of these scores to generate the overall c",
  },
  {
    "question": "How many total claims are reported by Snopes according to Table 1?",
    "answer": "4341 total claims are reported by Snopes.",
    "type": "multimodal-t",
    "evidence": "The number of total claims reported by Snopes is directly stated in the first row under the SN (Snopes) column."
  },
  {
    "question": "What is the percentage of unverified claims out of the total claims for the SE dataset?",
    "answer": "The percentage of unverified claims out of the total claims for the SE dataset is approximately 34.93%.",
    "type": "multimodal-t",
    "evidence": "This requires calculating the percentage by using the number of unverified claims (95) and the total number of claims (272) in the SE dat",
  },
  {
```


DocBench Outputs

```
{ } gpt-4o_eval_output.jsonl
```

```
1  {"question": "What specific measures does Prosper Africa intend to implement to enhance trade relationships between the U.S. and Africa?",
2   "answer": "Prosper Africa intends to increase domestic business outreach activities, provide platforms for showcasing investment opportuni
3   "type": "text-only",
4   "evidence": "Through Prosper Africa, we will increase domestic... more conducive to investment.",
5   "sys_ans": "Prosper Africa intends to enhance trade relationships between the U.S. and Africa by increasing domestic business outreach act
6   "file": "122",
7   "eval": "1"}
8  {"question": "When was the document approved?",
9   "answer": "On February 14, 2022.",
10  "type": "meta-data",
11  "evidence": "",
12  "sys_ans": "The document was approved on February 14, 2022.",
13  "file": "122",
14  "eval": "1"}
15 {"question": "What is the second footnote in the document?",
16  "answer": "The document does not have a second footnote.",
17  "type": "unanswerable",
18  "evidence": "",
19  "sys_ans": "The second footnote in the document is: \"Programs and activities implemented by State and USAID in support of the JRS for Afr
20  "file": "122",
21  "eval": "0"}
```

RACE

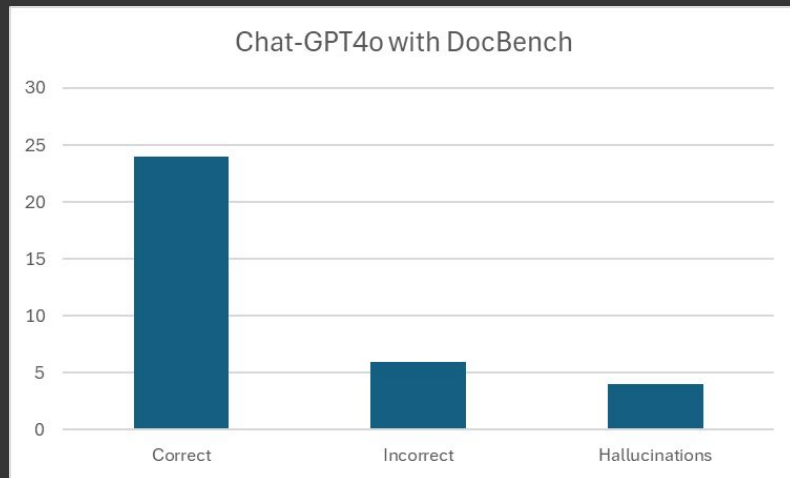
```
{
  "answers": ["A", "D", "B", "B", "C"],
  "options": [
    [
      "On a beach at night",
      "In a house near the dunes.",
      "On a front porch in the evening.",
      "In the shallow water of the ocean."
    ],
    [
      "Proud.",
      "Sad.",
      "Confused.",
      "Surprised."
    ],
    [
      "She read it in a book.",
      "Her grandmother told her.",
      "She learned about the turtles at school.",
      "Her parents taught her about the turtles."
    ],
    [
      "To get a flashlight",
      "To turn off the porch light.",
      "To communicate with her parents.",
      "To ask her grandmother about the lights."
    ],
    [
      "enjoying the ocean",
      "turning lights off at night",
      "saving the baby turtles",
      "learning from her grandparents"
    ]
  ],
  "questions": [
    "Where did the story happen?",
    "How did Mira feel when she saw the baby turtles appear?",
    "How did Mira know the lights confused the turtles?",
    "Why did Mira run to her house?",
    "This story is mainly about a girl _ ."
  ],
  "article": "One warm September night, Mira had come down to the beach to enjoy the night air. She closed her eyes and listened to the",
  "id": "high144.txt"
}
```

RACE Outputs

file	index	question	gold	pred	correct	hallucinated	justification	explanation
1279.txt	0	It can be inferred that ...	C	C	1	FALSE	0 The explanation does not contain hallucinations. It accurately references information from the passage and logically deduces the correct answer based on the details provided.	The passage provides information about a scene. Based on this information, we can infer that: A. Jackie Gleason is the director of the film Not B. Jackie Gleason is very angry when he is fired C. Max, a character in a film, is in very low spirit D. Jackie Gleason is the writer of the film Nothin' Therefore, the correct inference is option C: Max
23123.txt	1	Which of the following is not true according to the passage?	B	B	1	FALSE	0 The explanation does not contain hallucinations. It accurately evaluates each option based on the information provided in the passage, without introducing any new facts or external knowledge.	Based on the passage, let's evaluate each option: A. Everyone has an inborn number sense. - The passage states that humans are born with a number sense. B. Those who have a strong number sense are born with it. - The passage mentions that there is a link between a strong number sense and being born with it. C. Some of the math skills are inborn. - The passage explicitly states that not all math skills are inborn. D. Doing a lot of math exercises may help to learn math. - The passage concludes by stating that doing math exercises can help. Based on the passage, the statement that is not true is option B.
17784.txt	2	The passage is probably going to talk about ... if it continues.	A	A	1	FALSE	0 The explanation does not contain hallucinations. It accurately summarizes the passage and logically deduces that the passage might continue discussing other damage caused by the polar vortex, as suggested by the phrase "other damage".	The passage provides a detailed account of the damage caused by the polar vortex. Option A is the best choice because the passage mentions "other damage" which implies further discussion of the effects of the polar vortex.
15336.txt	3	What can we learn about Jonathan Saunders from the passage?	B	B	1	FALSE	0 The explanation accurately reflects the information in the passage, specifically that Jonathan Saunders' clothing collection is affordable, with prices ranging from \$14.99 to \$29.99. It correctly identifies that there is no information about his favorite color or his favorite food.	Based on the passage, we can learn that Jonathan Saunders' clothing collection is affordable. The clothes he designed are affordable.
3540.txt	4	We can know from Alice's words that ...	A	A	1	FALSE	0 The explanation accurately reflects the passage, which states that Tom had informed Alice about the signs and their meanings. There are no hallucinations present in the explanation.	Based on the passage, the correct option is A. Explanation: Alice explains to the gentleman that the signs on the wall mean that the weather is going to be bad.
24274.txt	5	After reading the passage, we know that Karen ...	A	A	1	FALSE	0 The model's explanation accurately reflects the information provided in the passage and the answer options without introducing any new facts or details.	After reading the passage, we can determine that: A. The passage states that Karen Rodger, 41, gave birth to the twin girls in her forties. B. The passage mentions that Karen has given birth to twins. C. The passage indicates that Karen was the one who gave birth to the twins. D. There is no indication in the passage that Karen gave birth to twins. Based on the passage, the correct answer is: A. gave birth to the twin girls in her forties
20515.txt	6	The two scientists found the color1 of the universe by ...	D	A	0	FALSE	0 The explanation accurately reflects the information in the passage without introducing any new facts. It correctly identifies the method used by the scientists and addresses why the other options are incorrect based on the passage.	The two scientists, Karl Glazebrook and Ivan B. Tsvetanov, discovered that the universe is a combination of light from about 200,000 star systems.
9.txt	7	Which of the following can we do in the zoo? [Zoo&Cont]	D	D	1	FALSE	0 The explanation accurately reflects the passage and the options, without introducing any new information or hallucinations.	Based on the passage, the zoo has specific rules for visitors: A. to give some food to the fish - This action is not allowed. B. to touch the monkey on the head - This action is not allowed. C. to throw things everywhere - This action is not allowed. D. to keep the zoo clean - This action is encouraged. Therefore, the only action that is allowed and encouraged is option D.
4757.txt	8	Why do some people enjoy the game?	A	A	1	FALSE	0 The explanation accurately reflects the passage and the answer options without introducing any new information.	Based on the passage, the reason some people enjoy the game is that it is a simple and fun game that can be played anywhere.
23637.txt	9	According to the officials, the name change of BeaverCollege ...	A	A	1	FALSE	0 The explanation does not contain hallucinations. It accurately references the passage's information about the reasons for the name change and the positive outcomes, such as increased applications and test scores, which supports the correct answer.	According to the passage, the name change of Beaver College was successful. Therefore, the most appropriate option is: A. turned out very successful

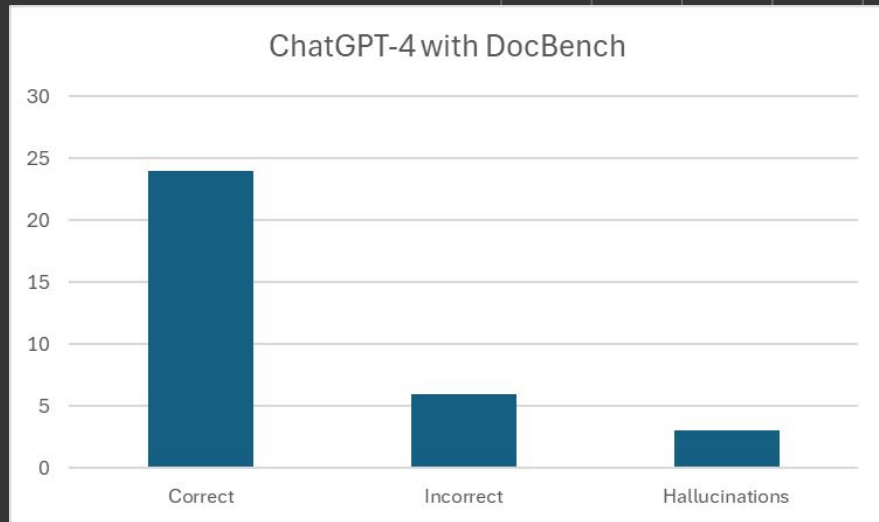
ChatGPT-4o with DocBench

- Answered correctly and accurately when asked about news articles, table referencing, and high level summarization
- 24/30 correct, 6 incorrect, 4 hallucinatory
- Model hallucinated most when asked about metadata, visual figures, and when given unanswerable questions
 - When asked how many figures there were in a document, ChatGPT responded with 6 when there were actually 2
 - Incorrectly answered the amount of words in the document
 - Generated a sentence from page 12 when there was no 12th page



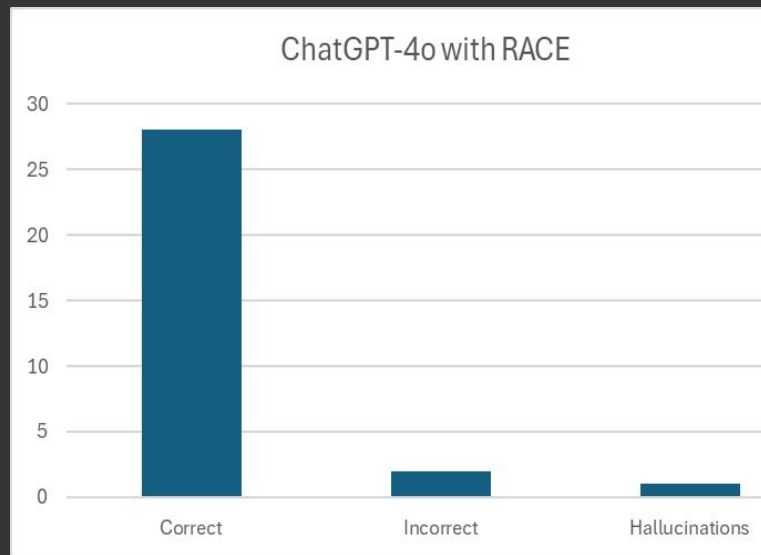
ChatGPT-4 with DocBench

- Got 24/30 answers, giving 80% accuracy
- The hallucination rate was 20%
- Out of the 6 incorrect answers, 2 were from the meta-data type, 2 were unanswerable, 1 was a question using figures, and 1 was a question using a table
- Hallucination example:
 - Question: "What is the first sentence on page 12?"
 - Correct Answer: "The paper does not have page 12."
 - Model's Answer: "The first sentence on page 12 is: \"For Snopes and PolitiFact, we need to extract relevant snippets from the reporting articles for a claim\"."



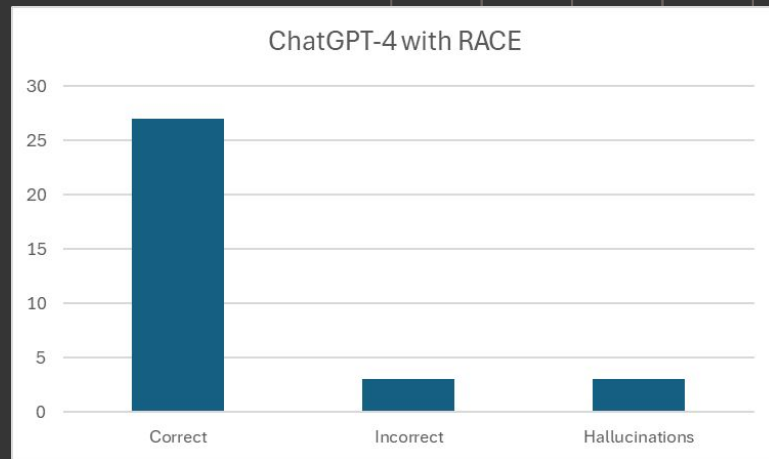
ChatGPT-4o with RACE

- Scored a 93%, or 28/30 questions correct
- Only one hallucination occurred
- The other incorrect answer had correct reasoning and did not fabricate any details, but still got the answer incorrect
- The hallucination occurred because the model invented information when reasoning.
 - 4o invented information about a change happening to the New York Times in 2006.
 - The passage and question never gave this year, yet 4o used 2006 for reasoning in its answer.



ChatGPT-4 with RACE

- Answered 27/30 questions correctly
- Used the same questions we asked ChatGPT-4o
- Exceeded in detail retrieval, vocabulary in context questions, and cause-and-effect questions
- Hallucinations occurred because the model merged information together and was over-inferring given some facts
- Specific Example:
 - “After seven weeks the woman he chose decided she didn’t want to marry him.”
 - “After leaving the show, later, months after, he proposed to Ulrika Jonsson.”
 - Model hallucinated that he proposed after seven weeks



Future

- We plan to finish our testing with the two GPT models on the TruthfulQA Benchmark.
- Increase the number of questions from each benchmark to improve data accuracy if time permits
- Highlight shortcomings of these benchmarks that could affect their effectiveness

Thanks!

