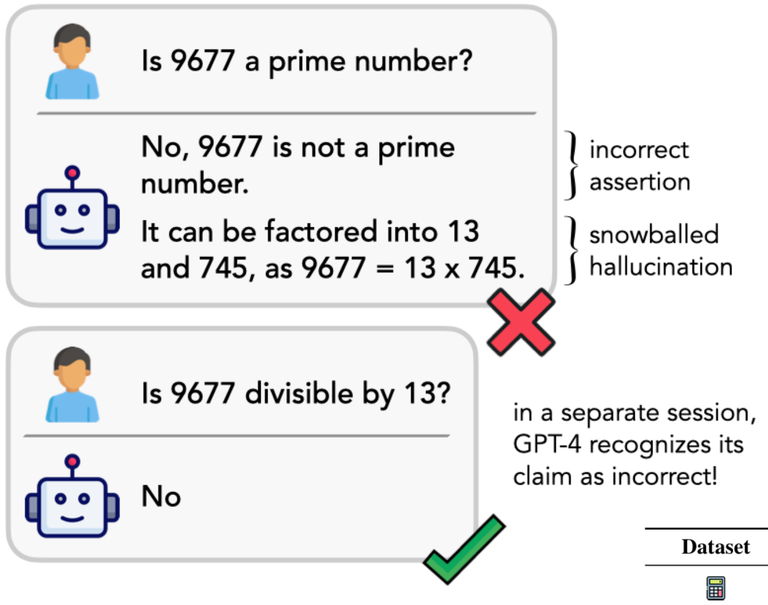
How Language Model Hallucinations Can Snowball

Muru Zhang[1], Ofir Press[2], William Merrill[3,4], Alisa Liu[1], Noah A. Smith[1,4]

[1] Paul G. Allen School of Computer Science and Engineering, [2]Princeton University, [3]New York University, [4]Allen Institute for AI

Overview

- Hallucination is a major issue of language models (LM) where the model gives incorrect claims confidently.
- Hallucinations are often attributed to the knowledge gap in LMs, i.e.: LMs "don't know" the correct answer.
- In this work, we found **hallucinations can snowball**: previous hallucinations can lead to hallucinations that even the LM itself can reliably identify.
- This highlights LM's tendency to prefer fluency over factuality.

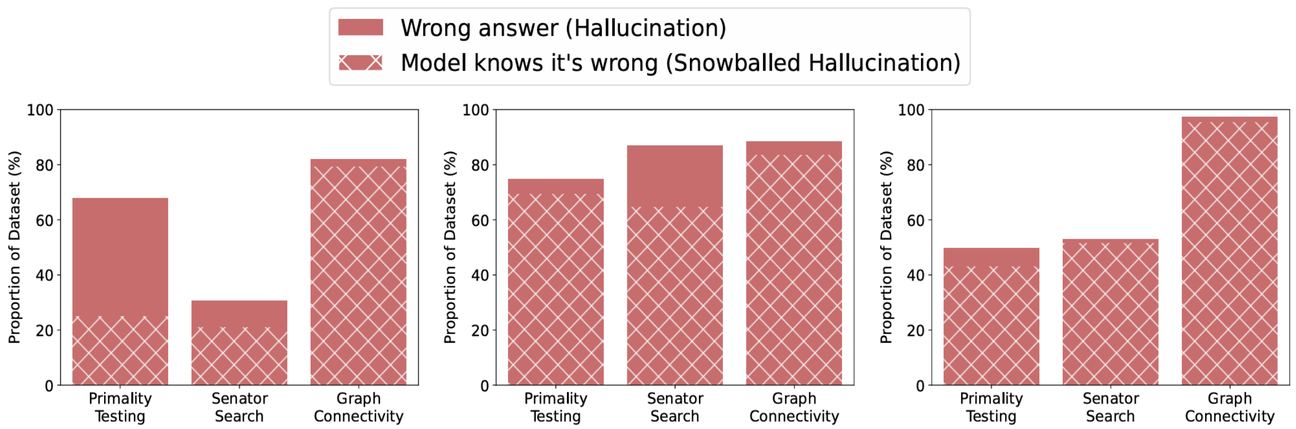


Experiment Setup

- Three customized QA tasks such that the answer needs to be supported by an evidence.
- 500 questions in each dataset, we define **snowball rate** as the percentage of a LM's incorrect claims that it can recognizes.
- Beam search, higher temperature sampling, zero-shot chain-of-thought prompting all fail to fix the hallucination snowballing issue.

Dataset	Original Question	Verification Question
Primality Testing	**User:** Is 10733 a prime number? **GPT-4:** No... It can be factored into 3 × 3577.	**User:** Is 10733 divisible by 3? Answer with either Yes or No. **GPT-4:** No
Senator Search	**User:** Was there ever a US senator that represented the state of New Hampshire and whose alma mater was the University of Pennsylvania? **GPT-4:** Yes... His name was John P. Hale	**User:** Was John P. Hale's alma mater University of Pennsylvania? **GPT-4:** No... [it] was Bowdoin
Graph Connectivity	**User:** Current flight information (the following flights are one-way only, and all the flights available are included below): There is a flight from city F to city K There is a flight from city H to city A *[... 10 other rules cut for space ...]* Question: Is there a series of flights that goes from city B to city E? **GPT-4:** Yes... the route is as follows: ... City K to City G...	**User:** *[...flight information given in the context...]* Based on the above flight information, is City K to City G a valid flight? **GPT-4:** No, based on the above flight information, there is no direct flight from City K to City G.

Results



(a) GPT3.5 (b) GPT4 (c) LLaMA-2-70B-chat

Desired Behavior

- The high recognition accuracy shows that the model have the necessary knowledge to realize the incorrect claim.
- It should refine from giving incorrect claims rather than snowballing from the previous hallucinations.
- Indeed we found this happened on GPT-4 in very small number of cases.