Preventing ‘Hallucination’ in GPT-3 and Other Complex Language Models

Originally published at https://www.unite.ai/preventing-hallucination-in-gpt-3-and-other-complex-language-models/

A defining characteristic of ‘fake news’ is that it frequently presents false information in a context of factually correct information, with the untrue data gaining perceived authority by a kind of literary osmosis – a worrying demonstration of the power of half-truths.

Sophisticated generative natural language processing (NLP) processing models such as GPT-3 also have a tendency to ‘hallucinate’ this kind of deceptive data. In part, this is because language models require the capability to rephrase and summarize long and often labyrinthine tracts of text, without any architectural constraint that’s able to define, encapsulate and ‘seal’ events and facts so that they are protected from the process of semantic reconstruction.

Therefore the facts are not sacred to an NLP model; they can easily end up treated in the context of ‘semantic Lego bricks’, particularly where complex grammar or arcane source material makes it difficult to separate discrete entities from language structure.

Examples of Errors in Supervised Model Paraphrase Generation

Error Type	Input Sentence	Model Generated Paraphrase
Stuttering	a man in a chef hat prepping some food	a man is sitting in a a a
	A man is holding up a clear umbrella	A man is holding a in the the
Sentence Fragments	a large brown horse eating a glob of leaves	a horse eating a piece of tall brown
	a skateboarder turning around on a half pipe	a person is riding a a skateboard on a
Hallucination	Two tables next to each other along with laptops	two people sitting on the beach with their laptops
	a city street line with very tall buildings	a city street with several signs on the street

An observation of the way that tortuously-phrased source material can confound complex language models such as GPT-3. Source: Paraphrase Generation Using Deep Reinforcement Learning

This problem spills over from text-based machine learning into computer vision research, particularly in sectors which utilize semantic discrimination to identify or describe objects.

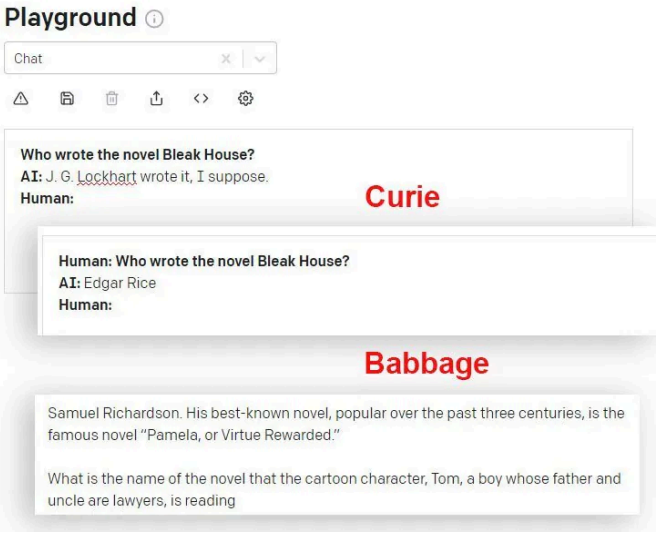
Sentence 1	Sentence 2
Three teddy bears sit in a sled in fake snow	A set of plush toy teddy bears sitting in a sled
A black Honda motorcycle parked in front of a garage	A Honda motorcycle parked in a grass driveway
A cat eating a bird it has caught	A white cat caught a bird outside on a patio

Hallucination and inaccurate ‘cosmetic’ reinterpretation affects computer vision research as well.

In the case of GPT-3, the model can become frustrated with repeated questioning on a topic it has already addressed as well as it can. In the best case scenario, it will admit defeat:

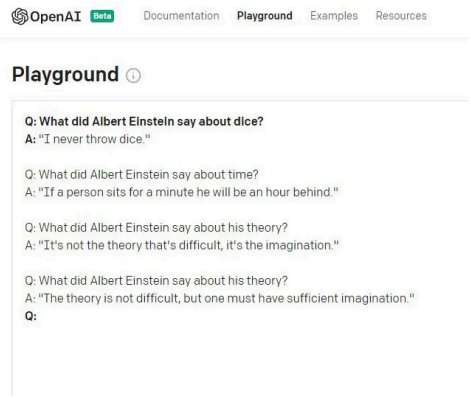
A recent experiment of mine with the basic Davinci engine in GPT-3. The model gets the answer right on the first attempt, but is vexed at being asked the question a second time. Since it retains a short-term memory of the previous answer, and treats the repeated question as rejection of that answer, it concedes defeat. Source: https://www.scalr.ai/post/business-applications-for-gpt-3

DaVinci and DaVinci Instruct (Beta) do better in this regard than other GPT-3 models available via the API. Here, the Curie model gives the wrong answer, while the Babbage model expands confidently on an equally wrong answer:



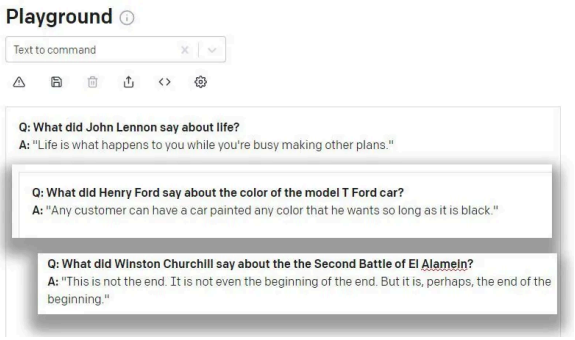
Things Einstein Never Said

When soliciting the GPT-3 DaVinci Instruct engine (which currently seems to be the most capable) for Einstein's famous quote 'God does not play dice with the universe', DaVinci instruct fails to find the quote and invents a non-quote, going on to hallucinate three other relatively plausible and completely non-existent quotes (by Einstein or anyone) in response to similar queries:



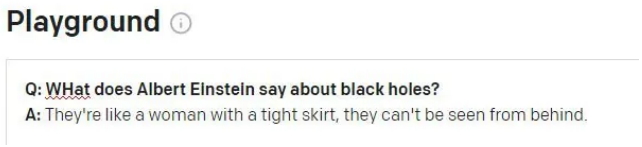
GPT-3 produces four plausible quotes from Einstein, none of which produce any results at all in a full-text internet search, though some trigger other (real) quotes from Einstein on the topic of 'imagination'.

If GPT-3 was consistently wrong in quoting, it would be easier to discount these hallucinations programmatically. However, the more diffused and famous a quote is, the more likely GPT-3 is to get the quote right:



GPT-3 apparently finds correct quotes when they are well-represented in the contributing data.

A second problem can emerge when GPT-3's session history data bleeds into a new question:



Einstein would likely be scandalized to have this saying attributed to him. The quote appears to be a nonsensical hallucination of a real-life Winston Churchill aphorism. The previous question in the GPT-3 session related to Churchill (not Einstein), and GPT-3 appears to have mistakenly used this session token to inform the answer.

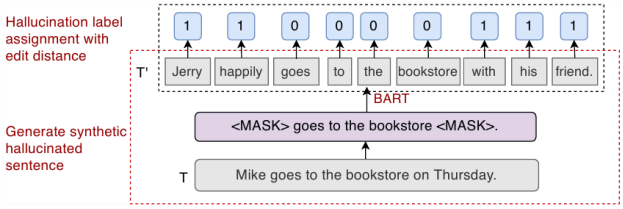
Tackling Hallucination Economically

Hallucination is a notable obstacle to the adoption of sophisticated NLP models as research tools – the more so as the output from such engines is highly abstracted from the source material that formed it, so that establishing the veracity of quotes and facts becomes problematic.

Therefore one current general research challenge in NLP is to establish a means of identifying hallucinated texts without the need to imagine entirely new NLP models that incorporate, define and authenticate facts as discrete entities (a longer-term, separate goal in a number of wider computer research sectors).

Identifying And Generating Hallucinated Content

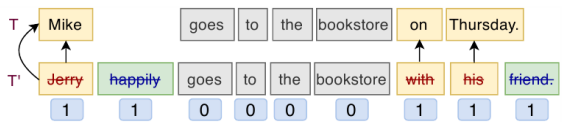
A new collaboration between Carnegie Mellon University and Facebook AI Research offers a novel approach to the hallucination problem, by formulating a method to identify hallucinated output and using synthetic hallucinated texts to create a dataset that can be used as a baseline for future filters and mechanisms that might eventually become a core part of NLP architectures.



Source: https://arxiv.org/pdf/2011.02593.pdf

In the above image, source material has been segmented on a per-word basis, with the '0' label assigned to correct words and the '1' label assigned to hallucinated words. Below we see an example of hallucinated output that is related to the input information, but is augmented with non-authentic data.

The system uses a pre-trained denoising autoencoder that's capable of mapping a hallucinated string back to the original text from which the corrupted version was produced (similar to my examples above, where internet searches revealed the provenance of false quotes, but with a programmatic and automated semantic methodology). Specifically, Facebook's BART autoencoder model is used to produce the corrupted sentences.



Label assignment.

The process of mapping the hallucination back to the source, which is not possible in the common run of high-level NLP models, allows for mapping the 'edit distance', and facilitates an algorithmic approach to identifying hallucinated content.

The researchers found that the system is even able to generalize well when it has no access to reference material that was available during training, which suggests that the conceptual model is sound and broadly replicable.

Tackling Overfitting

In order to avoid overfitting and arrive at a widely deployable architecture, the researchers randomly dropped tokens from the process, and also employed paraphrasing and other noise functions.

Machine translation (MT) is also part of this obfuscation process, since translating text across languages is likely to preserve meaning robustly and further prevent over-fitting. Therefore hallucinations were translated and identified for the project by bi-lingual speakers in a manual annotation layer.

The initiative achieved new best results in a number of standard sector tests, and is the first to achieve acceptable results using data exceeding 10 million tokens.

The code for the project, entitled *Detecting Hallucinated Content in Conditional Neural Sequence Generation*, has been released on GitHub, and allows users to generate their own synthetic data with BART from any corpus of text. Provision is also made for the subsequent generation of hallucination detection models.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More