Hallucinations

Tuesday, 29 August 2023 11:14 PM

https://medium.com/@hooman_66365/effective-methods-against-llm-hallucination-50249

Tools for constrained generation:
https://github.com/guidance-ai/guidance

LLM as labeler:
https://www.refuel.ai/blog-posts/llm-labeling-technical-report#:~:text=Takeaways%3A-,Sta
20LLMs%20can%20label%20text%20datasets%20at,of%20human%20annotators%20(86.2%

Techniques for Detecting LLM Hallucinations
https://www.rungalileo.io/blog/5-techniques-for-detecting-llm-hallucinations

https://medium.com/mlearning-ai/the-hallucination-problem-of-large-language-models-5d
~:text=SelfCheckGPT%20is%20based%20on%20the,diverge%20and%20contradict%20one%

Below are a few other interesting frameworks that offer constraine

1. Guidance
2. Guardrails
3. Outlines

Smart GPT:
GPT 4 is Smarter than You Think: Introducing SmartGPT
https://python.langchain.com/docs/use_cases/more/self_check/smart_llm
https://python.langchain.com/docs/use_cases/more/self_check/llm_checker
https://python.langchain.com/docs/use_cases/more/self_check/llm_summarization_checke

d53e08

te%20of%20the%20art%
25)

7ab1b0f37f#:
20another.

d generation:

er

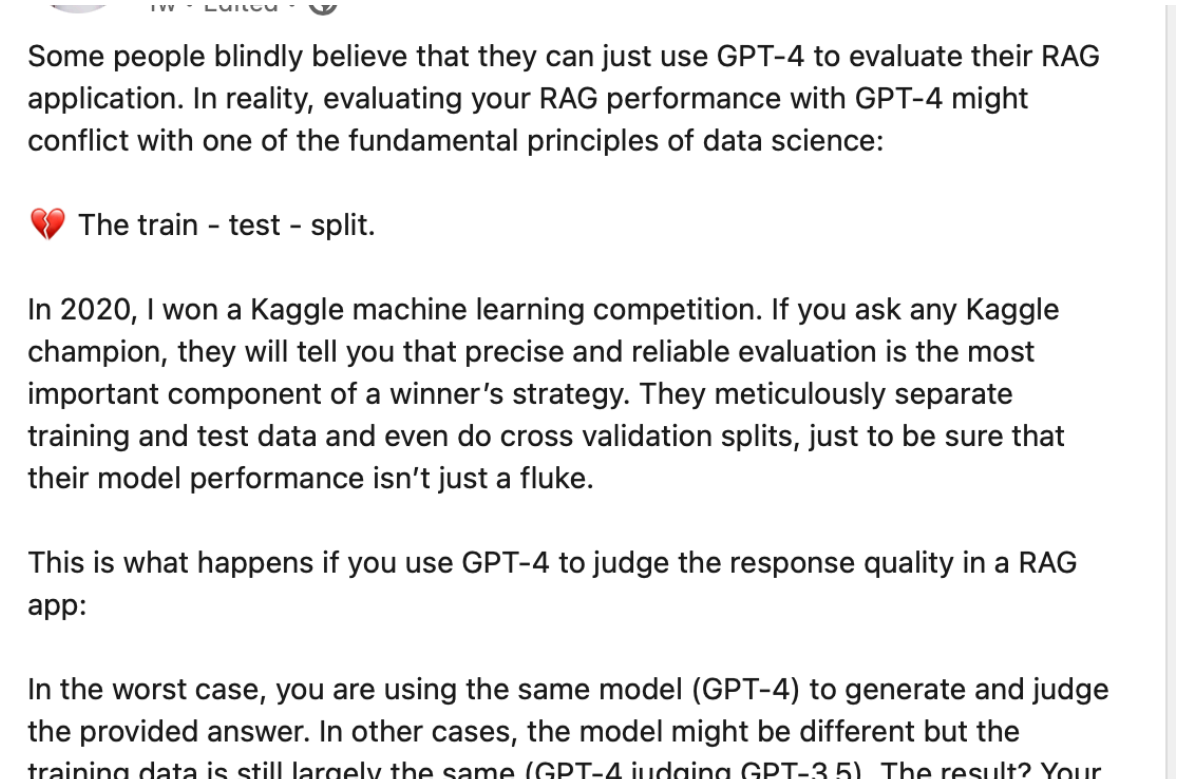
https://python.langchain.com/docs/use_cases/more/self_check/llm_summarization_checker

1. *Adjusting LLM generation parameters:* By controlling the output of LLMs through techniques like temperature adjustment, precision can be achieved in applications that require accuracy.
2. *Different decoding techniques:* Libraries like LMQL offer advanced constraints on the length and content of generated output, allowing for more controlled generation.
3. *Self-checking:* Techniques like SmartLLMChain and Self-checking involve multiple passes of the LLM output, critiquing and refining the generated ideas to improve response quality.
4. *Causal program-aided language (CPAL) chain:* CPAL

enhances PAL language chain by incorporating a causal graph representation, enabling the LLM to handle complex questions involving mathematical operations.

5. *Retrieval Augmented Generation (RAG):* RAG retrieves relevant documents to provide up-to-date knowledge, reducing hallucination. However, hallucination can still occur, particularly in mathematical reasoning. To minimize hallucination in RAG pipelines, prompt engineering plays a crucial role. Fine-tuning smaller LLMs for attribution evaluation can also help assess the reliability of LLM responses. Creative fine-tuning on specialized data and expert validation can further enhance attribution models.

2)



Some people blindly believe that they can just use GPT-4 to evaluate their RAG application. In reality, evaluating your RAG performance with GPT-4 might conflict with one of the fundamental principles of data science:

💔 The train - test - split.

In 2020, I won a Kaggle machine learning competition. If you ask any Kaggle champion, they will tell you that precise and reliable evaluation is the most important component of a winner's strategy. They meticulously separate training and test data and even do cross validation splits, just to be sure that their model performance isn't just a fluke.

This is what happens if you use GPT-4 to judge the response quality in a RAG app:

In the worst case, you are using the same model (GPT-4) to generate and judge the provided answer. In other cases, the model might be different but the training data is still largely the same (GPT-4 judging GPT-3.5). The result? Your

judgement comes with big fat error bars that you should be aware of!

In the attached graphic, you can see results from a paper where GPT-4 always preferred GPT-3.5 generated summaries over human written summaries: **https://lnkd.in/eVNBcj7R**

Hugging Face and **Scale AI** have partnered to investigate potential biases when using GPT-4 as an evaluator: **https://lnkd.in/e2QHqw-y**

They found that GPT-4:
- always prefers models that were trained on GPT-4 output (think instruction generation with GPT-4)
- has position bias (preference judgments are strongly influenced by the order of candidates)
- correlates badly with human judgements if the task requires factual correctness

Wang et al. find that even after applying extensive mitigation strategies, GPT-4 judgements correlate poorly with voted multiway human rated judgments (0.37 Cohen's Kappa):

\

TRICKS :

1)

Parallel response generation is also interesting in the context of hallucination. Nvidia recently presented their guardrails framework, the second part after the fact check is similar: if you feed a quote to multiple models and feed the different answers to another LLM instance and ask it to cluster and check for consistency, you can distinguish knowledge from hallucination. If the model knows something, the answer will be similar; if it is making things up, the hallucinations will usually go in different (inconsistent) directions. The consistent result can be fed back into a feedback loop for self-reflection. If no answer is consistent, the model doesn't know anything which could be given out as answer . Or the model could then actively search online or be tasked to identify a procedure to get the answer (an experiment or something), which would go in the direction of auto-GPT