Unsupervised hallucination detection

GitHub.com/kliph/cursor-hallucination-detection-talk-2025

What's hallucination?

An anthropomorphized term for an observed tendency of Large Language Models (LLMs) to output content that is factually incorrect, nonsensical, or not entailed based on human interpretation of the prompt provided to the LLM

Hallucination Examples

What is the capital of Pennsylvania?

- Factually incorrect
  - "The capital Pennsylvania is Philadelphia"
- Nonsensical
  - "The capital abc13jskiay ; /a"
- Not entailed
  - "The capital of France is Paris"

How can it be prevented?

one way:

(other ways exist and are subject to active research)

- https://arxiv.org/pdf/2303.08896
- Performance similar to human raters

SELFCHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models

Potsawee Manakul, Adian Liusie, Mark J. F. Gales
ALTA Institute, Department of Engineering, University of Cambridge
pm574@cam.ac.uk, al826@cam.ac.uk, mjfg@eng.cam.ac.uk

Abstract

Generative Large Language Models (LLMs) such as GPT-3 are capable of generating highly fluent responses to a wide variety of user prompts. However, LLMs are known to hallucinate facts and make non-factual statements which can undermine trust in their output. Existing fact-checking approaches either require access to the output probability distribution (which may not be available for systems such as ChatGPT) or external databases that are interfaced via separate, often complex, modules. In this work, we propose "SelfCheckGPT", a simple sampling-based approach that can be used to fact-check the responses of black-box models in a zero-resource fashion, i.e. without an external database. SelfCheckGPT leverages the simple idea that if an LLM has knowledge of a given concept, sampled responses are likely to be similar and contain consistent facts. However, for hallucinated facts, stochastically sampled responses are likely to diverge and contradict one another. We investigate this approach by using GPT-3 to generate passages about individuals from the WikiBio dataset, and manually annotate the factuality of the generated passages. We demonstrate that SelfCheckGPT can: i) detect non-factual and factual sentences; and ii) rank passages in terms of factuality. We compare our approach to several baselines and show that our approach has considerably higher AUC-PR scores in sentence-level hallucination detection and higher correlation scores in passage-level factuality assessment compared to grey-box methods.[1]

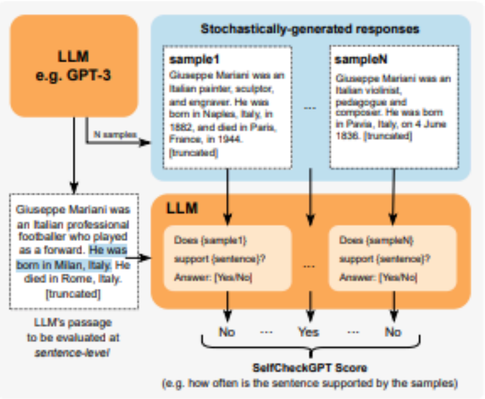

Figure 1: SelfCheckGPT with Prompt. Each LLM-generated sentence is compared against stochastically generated responses with no external database. A comparison method can be, for example, through LLM prompting as shown above.

tools to draft reports, virtual assistants and summarization systems. Despite the convincing and realistic nature of LLM-generated texts, a growing concern with LLMs is their tendency to hallucinate facts. It has been widely observed that models can confidently generate fictitious information, and worryingly there are few, if any, existing approaches to suitably identify LLM hallucinations.

A possible approach of hallucination detection is to leverage existing intrinsic uncertainty metrics to determine the parts of the output sequence that the system is least certain of (Yuan et al., 2021; Fu et al., 2023). However, uncertainty metrics such

How does it work?

- Hallucinations are operationalized as **Major** and **Minor Inaccuracies**
- **Major Inaccuracies**
  - defined by answering "no" to the question "Is this sentence related to the context?"
- **Minor Inaccuracies**
  - "Is this sentence factual (in relation to some context/ documents)?" which are assessed in series after determining that the sentence does not contain major inaccuracies

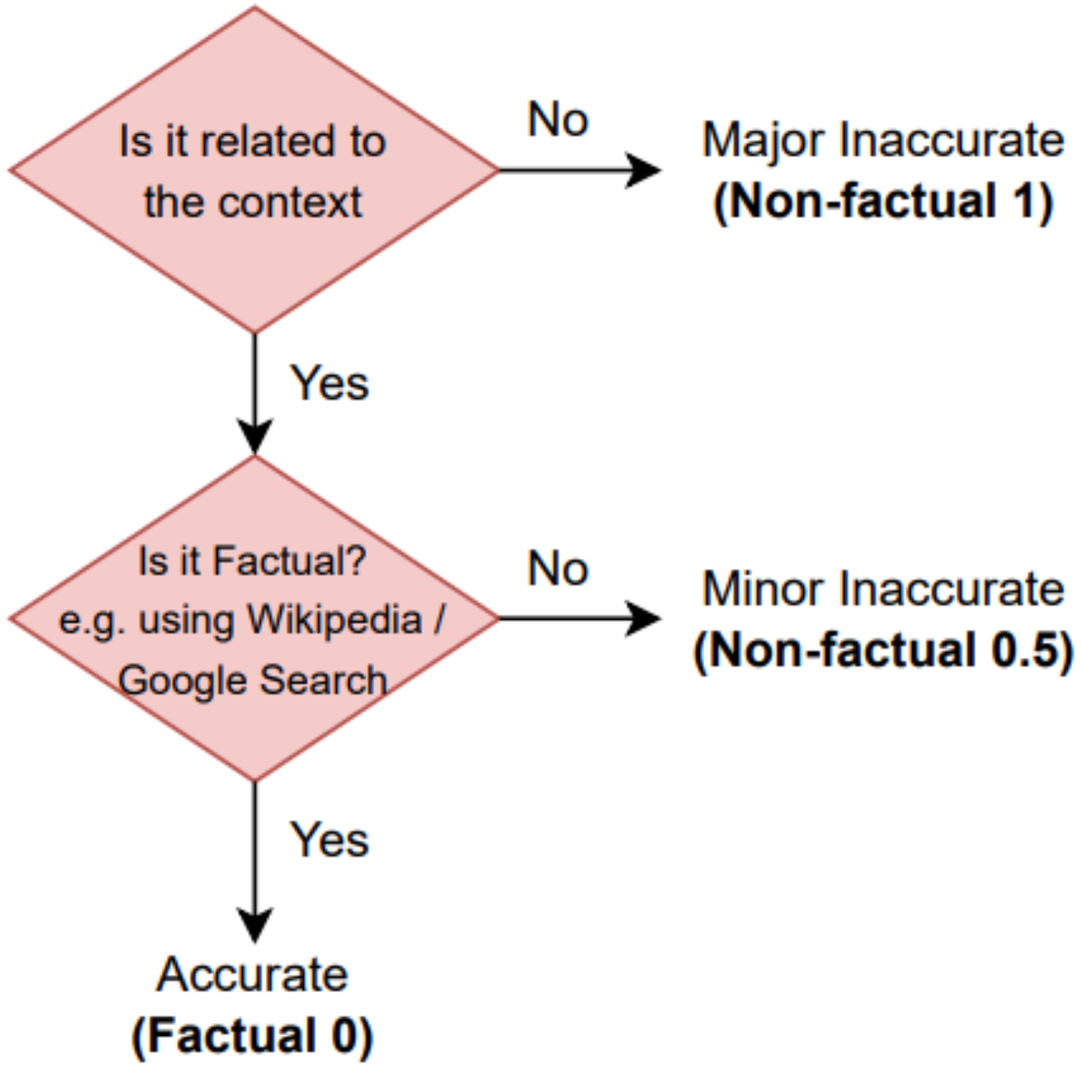

Figure 3: Flowchart of our annotation process

Yes

in cases where real time accuracy is important and humans are not in the loop

Is it useful?