



---

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



Potsawee Manakul



Adian Liusie



Mark Gales

Department of Engineering, University of Cambridge

---

# Outline

- What are Hallucinations?
- Existing Hallucination Detection approaches in NLG
- SelfCheckGPT
- Experiments
- Conclusion

# Hallucinations

- What are Hallucinations?
  - Generated text is not *faithful* to the input context
  - Generated text is not *factually correct* with respect to world knowledge



**Bill Murphy Jr.**  
@BillMurphyJr

I asked ChatGPT to write a short bio of [@billmurphyjr](#) (aka, me)

Apparently I:

- \* went to [@BrownUniversity](#)\* & [@MITSloan](#)\*
- \* worked in investment banking\* & management consulting\*
- \* wrote a book about medical breakthroughs\*

(\*not true)

When challenged it insisted it was correct.

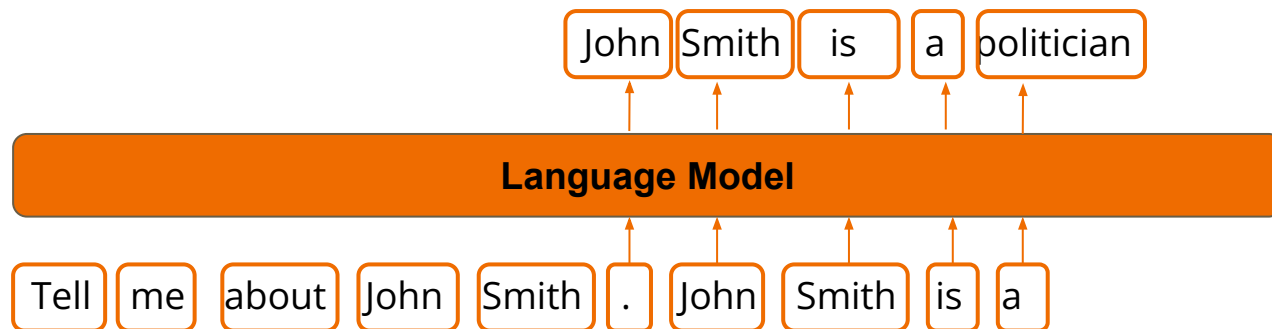
write a short biography of bill murphy jr.

Bill Murphy Jr. is a journalist, author, and entrepreneur based in Washington, D.C. He is the author of several books, including "The Intelligent Entrepreneur" and "Breakthrough: How the 10 Greatest Discoveries in Medicine Saved Millions and Changed Our View of the World." He is also the founder of the online news and commentary site, The Daily Update. Prior to his writing career, Murphy worked as an investment banker and management consultant. He received his undergraduate degree from Brown University and his MBA from the MIT Sloan School of Management.

Source: @BillMurphyJr on Twitter on 15 Feb 2023

# Hallucinations

- Why language models hallucinate?
  - Pre-training Data: Contradicting information, Inaccurate Information
  - Decoding Strategy:
    - Autoregressive nature of language modelling
    - Inability to verify facts



# Hallucinations

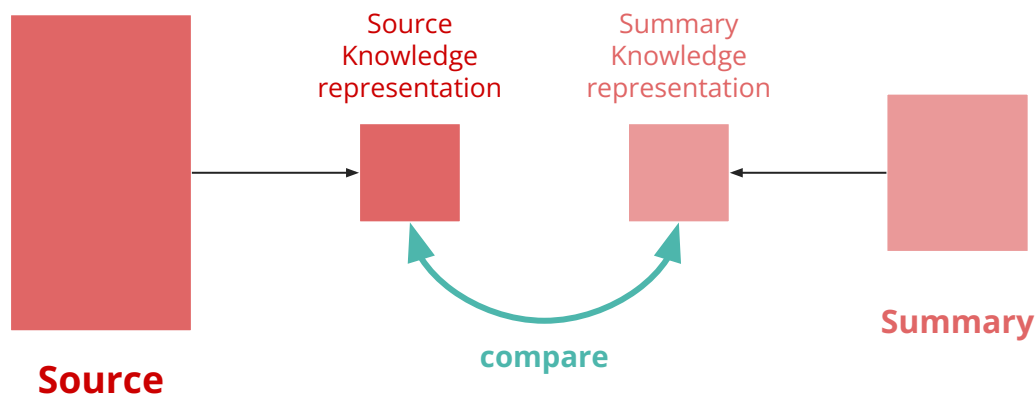
- Why LLMs Hallucinate?
  - Autoregressive language model

$$\hat{x}_i \sim P(x|\hat{x}_{1:i-1}, \text{prompt})$$

- The model is forced to generate the *next* token
  - *Deterministic v.s. Stochastic*
    - creativity can be controlled by temperature
- Teacher-forced training → Exposure Bias  $\hat{x}_{1:i-1}$

# Detecting Hallucinations in *NLG*

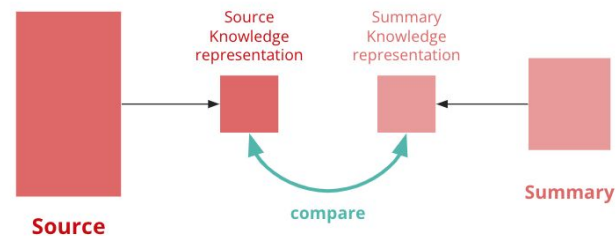
- Assessing Information Consistency in Summarization



- Recent survey by Ji et al. 2023, "*Survey of Hallucination in Natural Language Generation*"

# Detecting Hallucinations in *NLG*

- Existing methods for Measuring Faithfulness
  - Textual Overlap
    - N-gram – ROUGE/BLEU, Embedding – BERTScore
  - Information Extraction Based
    - Comparing Triple(*Source*) against Triple(*Sum*)
  - Question Answering Based
    - QAGS, QuestEval, MQAG
  - Language Model Score
    - BARTScore
  - Natural Language Inference (NLI)
    - Entailment Model

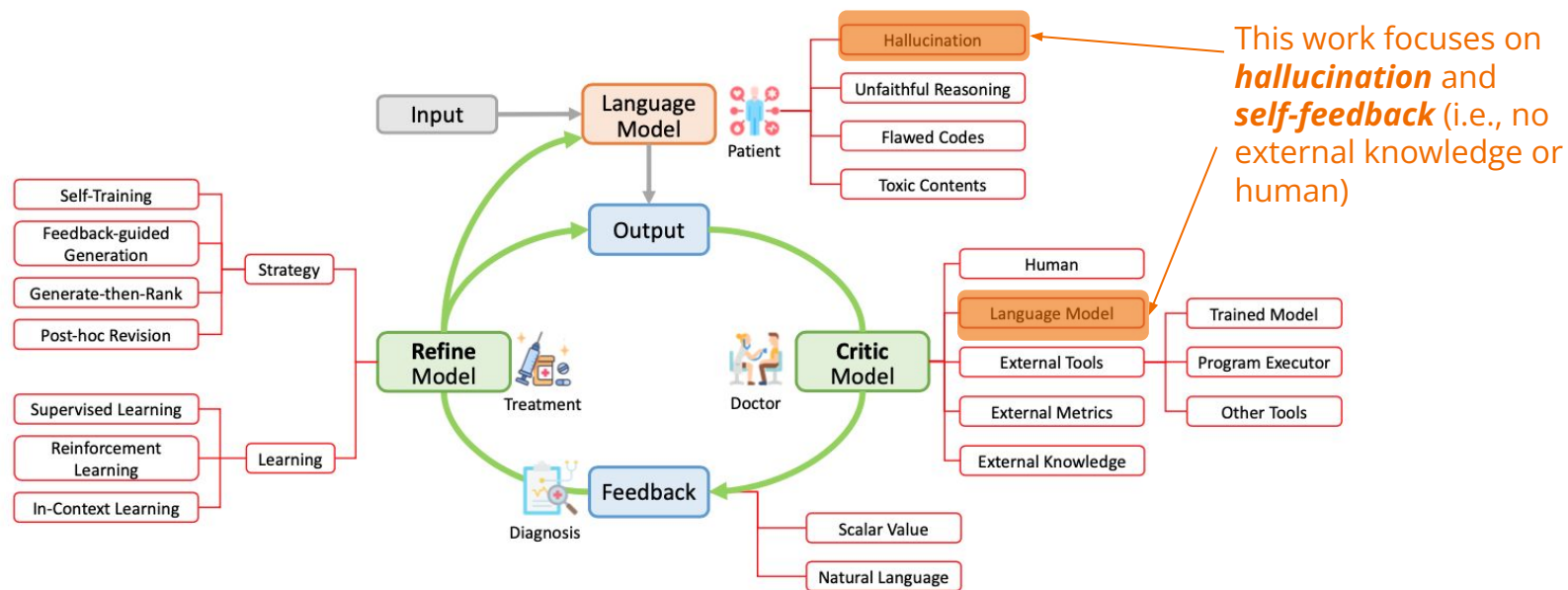


# Detecting Hallucinations in *LLM generation*

- Hallucination detection methods evaluate **Generated Text** against **Source** (e.g., document in summarization or previous context in dialogue generation) – **Faithfulness**
- In LLM generation, there is *little/no* source, especially in long-form generation – **Factuality**



# Detecting Hallucinations in *LLM* generation



Source: Pan et al. 2023, Automatically Correcting Large Language Models: *Surveying the landscape of diverse self-correction strategies*

# LLMs: White-box, Grey-box, Black-box

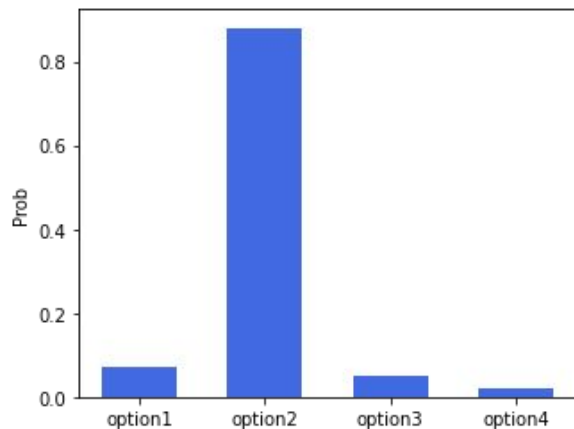
- White-box systems
  - Full access to the model's outputs and internal states
  - e.g., LLaMA, GPT-NeoX, various other open-source LLMs.
- Grey-box systems
  - Limited access to model outputs
  - e.g., GPT-3 API outputs texts and top-5 token probabilities (no internal states)
- Black-box systems
  - Only access to text-based outputs
  - e.g., ChatGPT (API, web interface)



# Scope of this work

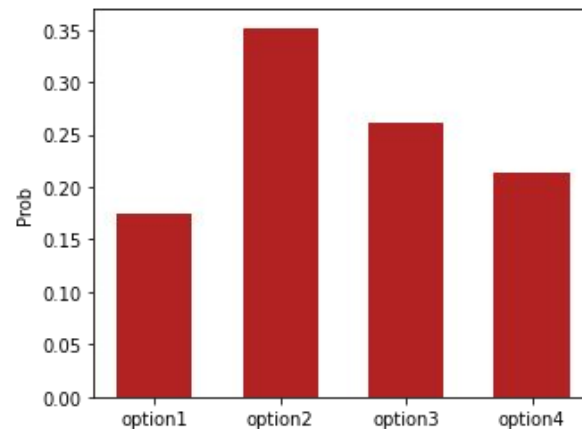
- **Black-box**
  - As some LLMs only outputs texts to the users
- **Zero-resource**
  - Does not require additional database / external knowledge
  - Allows one to assess any domain of LLM generation

# Uncertainty Measures



## Low Entropy (low uncertainty)

- The model is *more* certain about its prediction



## High Entropy (high uncertainty)

- The model is *less* certain about its prediction

# Token-level Probability

$$\hat{x}_i \sim P(x|\hat{x}_{1:i-1}, \text{prompt})$$

The screenshot shows the OpenAI Playground interface. The prompt is "This is a passage from Wikipedia about lenny randle:". The model is "text-davinci-003". The output text is highlighted in green, and a tooltip shows the token-level probabilities for the word "Texas".

Looking for ChatGPT? [Try it now](#)

Submit [refresh] [undo] [redo] [share] [like] 147

State	Probability
Texas	62.78%
Washington	30.45%
Seattle	6.48%
California	0.09%
Rangers	0.08%

Total: -0.47 logprob on 1 tokens  
(99.88% probability covered in top 5 logits)

$$\text{Avg}(-\log p) = -\frac{1}{J} \sum_j \log p_{ij}$$

$$\text{Max}(-\log p) = \max_j (-\log p_{ij})$$

Note that token-level output probabilities are required.  
(Grey-box approach)

Source: <https://platform.openai.com/playground>

# Consistency between Generated Responses

**Question:** Is faithfulness (consistency) between generated samples related to factuality?

## Sample 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt. ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris. nisi ut aliquip ex ea commodo consequat Duis aute irure. dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur Excepteur sint.

## Sample 2

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt. ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris. nisi ut aliquip ex ea commodo consequat Duis aute irure. dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur Excepteur sint.

.....

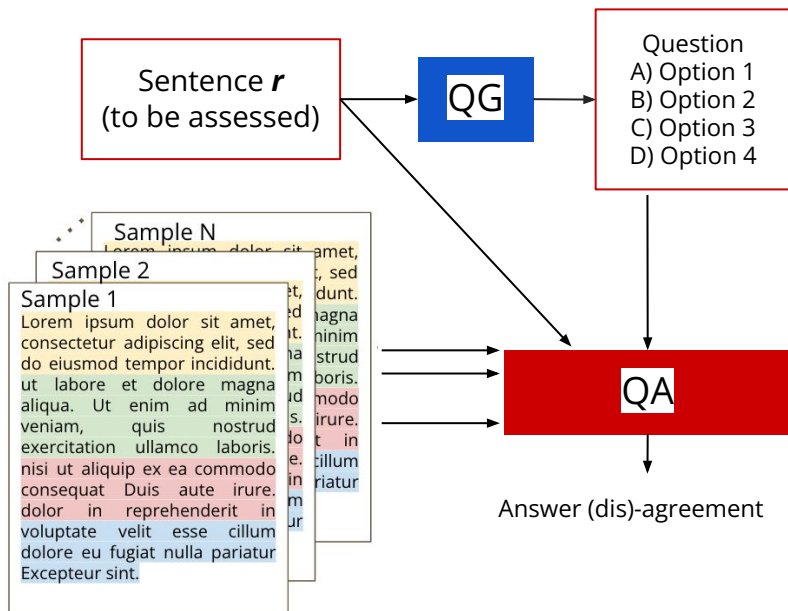
## Sample N

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt. ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris. nisi ut aliquip ex ea commodo consequat Duis aute irure. dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur Excepteur sint.

# SelfCheckGPT

- The measures are motivated by summary assessment methods
- Measure self-consistency between generated *samples*
  - Question-Answering
  - BERTScore
  - N-gram language model
  - Natural Language Inference (NLI)
  - LLM prompting

# SelfCheck with Question Answering

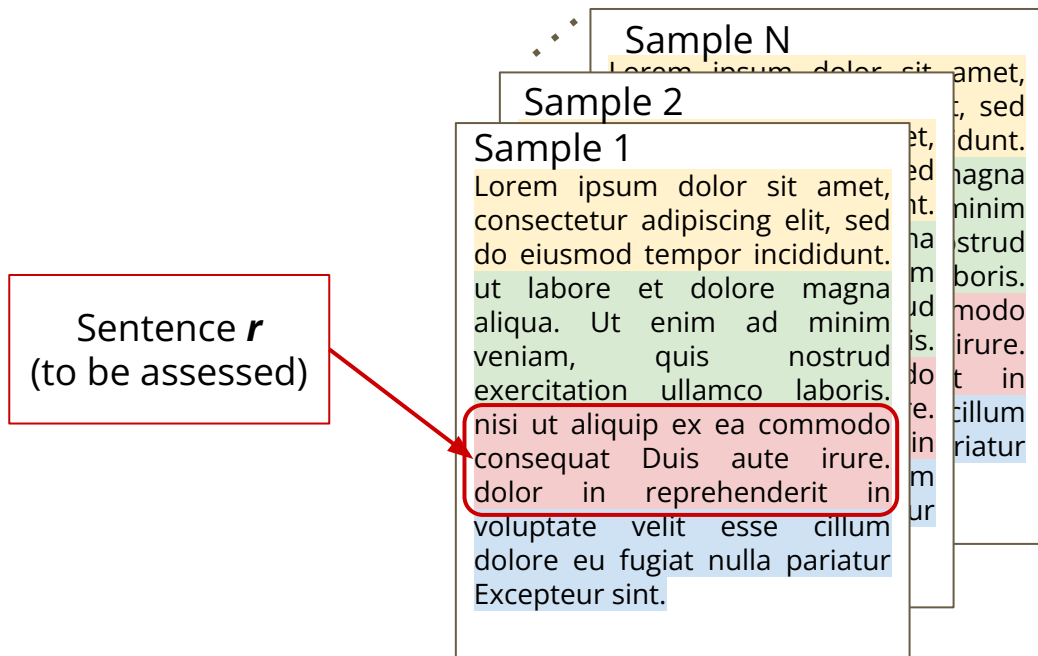


- For each sentence  $r$ 
  - Generate a set of {Question, Options}
- For each sample  $s^n$  and  $r$ 
  - Obtain the answer given  $s^n$
  - Obtain the answer given  $r$
- Inconsistency Score  
= 
$$\frac{\#mismatches}{(\#mismatches + \#matches)}$$

\*MQAG: Multiple-choice Question Answering and Generation for Assessing Information Consistency in Summarization, Manakul et al, 2023.



# SelfCheck with BERTScore

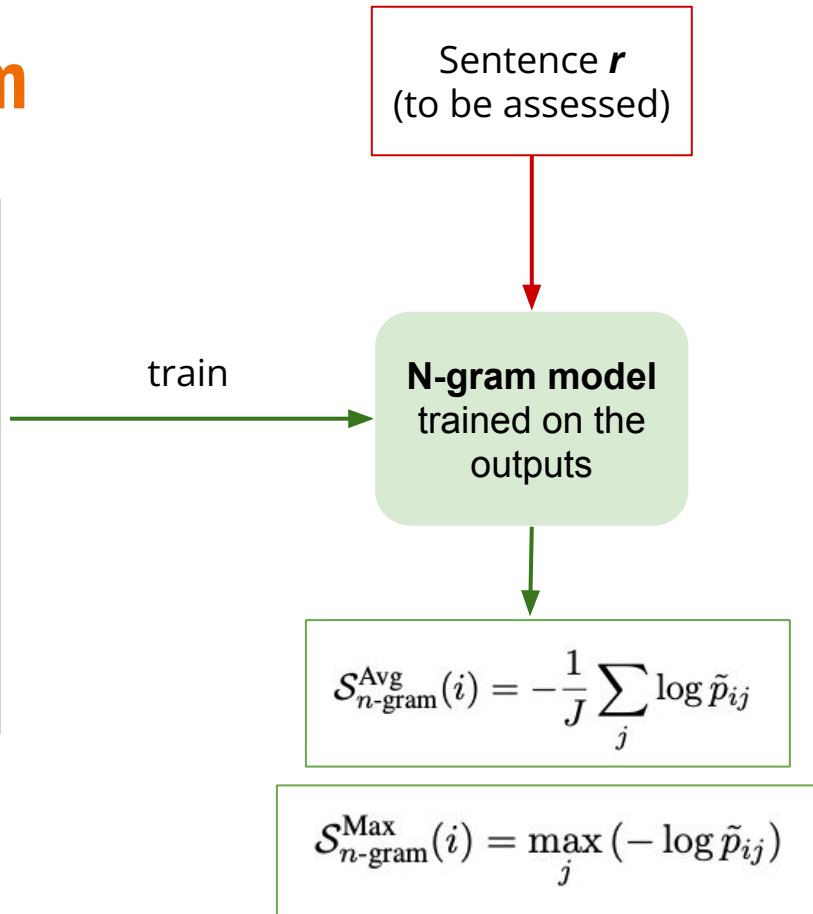
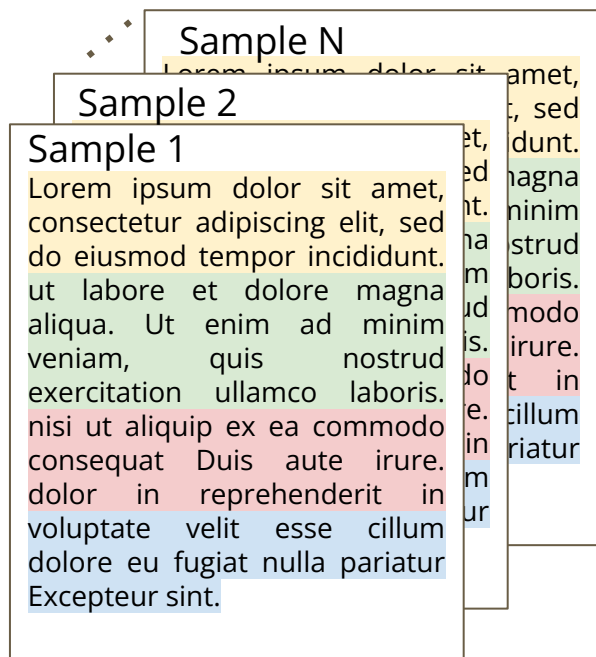


$$\mathcal{S}_{\text{BERT}}(i) = 1 - \frac{1}{N} \sum_{n=1}^N \max_k (\mathcal{B}(r_i, s_k^n))$$

1. Obtain BERTScore of the sentence  $s_k$  that is the most similar to  $r$

2. Repeat step1 for all outputs 1...N

# SelfCheck with n-gram



# SelfCheck with Natural Language Inference (NLI)

context [SEP] hypothesis

NLI Model

P\_entail = 0.7  
P\_neutral = 0.2  
P\_contradict = 0.1

DeBERTa-v3  
fine-tuned on  
Multi-NLI

[SEP] Sentence

[SEP] Sentence

[SEP] Sentence

NLI Model

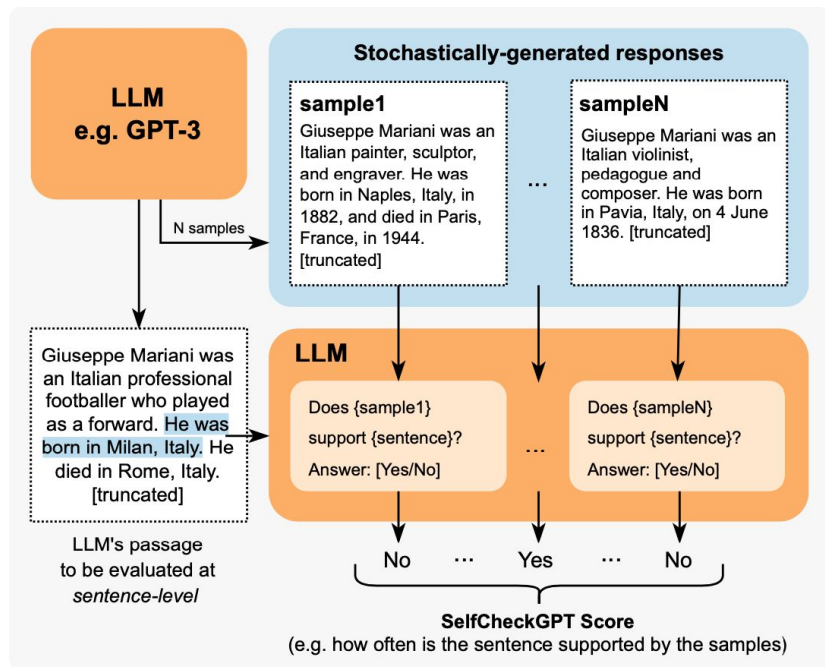
P\_contradict 1

P\_contradict 2

P\_contradict N

Score  
=  
Avg(P\_contradict)

# SelfCheck with LLM prompting



-----  
Context: {}  
Sentence: {}  
Is the sentence supported by the context above?  
Answer Yes or No:  
-----

$$S_{\text{Prompt}}(i) = \frac{1}{N} \sum_{n=1}^N x_i^n$$

# Experiments

- Data
  - Previous work investigated perturbed/corrupted texts, e.g., HaDes\*
  - More realistic LLM hallucination requires annotation for a specific LLM
    - We prompt GPT-3 (davinci-003) to generate passages about individuals in WikiBio

This is a Wikipedia passage  
about {individual}

GPT-3  
(davinci-003)

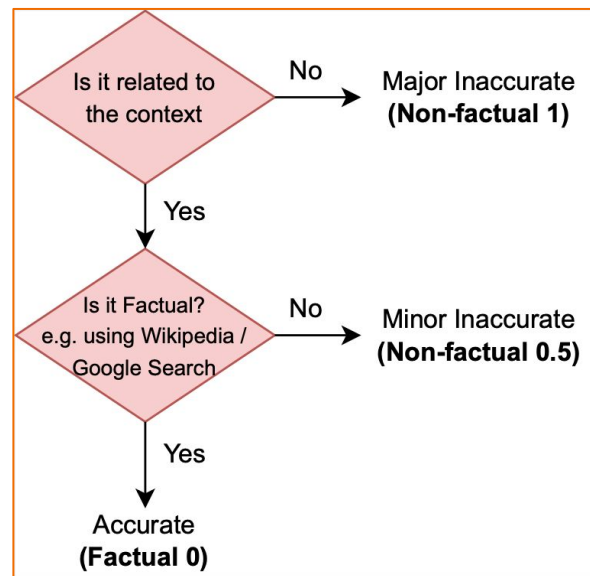
James Blair (September 26, 1786  
- April 1, 1834) was a United  
States Representative from  
South Carolina. He was born in  
the Waxhaw Settlement,  
Lancaster County, South  
Carolina to Sarah Douglass and  
William Blair. [truncated]

\*Liu et al, 2022. A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation

# Experiments

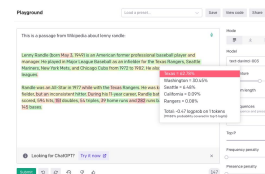
- Annotation

- Perform annotation at the sentence level
  - Manual annotation on 238 passages (1902 sentences)
- Passage-level score is obtained by averaging sentence-level scores
- Evaluation using AUC-PR (Sentence-level) or Correlation (Passage-level)



# Experimental Results

Method	Non-Factual (sent-lvl) AUC-PR	Factual (sent-lvl) AUC-PR	Ranking (passage-lvl) Spearman
Random Guessing	72.96	27.04	-
GPT-3 Avg(-logP)	83.21	<b>53.97</b>	53.93
GPT-3 Avg(H)	80.73	52.07	50.87
GPT-3 Max(-logP)	<b>87.51</b>	50.46	<b>55.69</b>
GPT-3 Max(H)	85.75	50.27	49.55



Grey-box approach

# Experimental Results

Method	Non-Factual (sent-lvl) AUC-PR	Factual (sent-lvl) AUC-PR	Ranking (passage-lvl) Spearman
Random Guessing	72.96	27.04	-
GPT-3 Avg(-logP)	83.21	<b>53.97</b>	53.93
GPT-3 Avg(H)	80.73	52.07	50.87
GPT-3 Max(-logP)	<b>87.51</b>	50.46	<b>55.69</b>
GPT-3 Max(H)	85.75	50.27	49.55
LLaMA-30B Avg(-logP)	75.43	41.29	20.20
LLaMA-30B Avg(H)	80.80	42.97	39.49
LLaMA-30B Max(-logP)	74.01	31.08	22.71
LLaMA-30B Max(H)	80.92	37.90	38.94

→ Grey-box approach  
with LLM probability

→ Proxy LLM probability



# Experimental Results

Method	Non-Factual (sent-lvl) AUC-PR	Factual (sent-lvl) AUC-PR	Ranking (passage-lvl) Spearman
Random Guessing	72.96	27.04	-
GPT-3 Max(-logP)	87.51	50.46	55.69
LLaMA-30B Avg(H)	80.80	42.97	39.49
SelfCheck-BERTS	81.96	44.23	55.90
SelfCheck-QA	84.26	48.14	59.29
SelfCheck-Unigram	85.63	58.47	64.91
SelfCheck-NLI	92.50	66.08	73.78
<b>SelfCheck-Prompt</b>	<b>93.42</b>	<b>67.09</b>	<b>78.30</b>

→ Grey-box approach

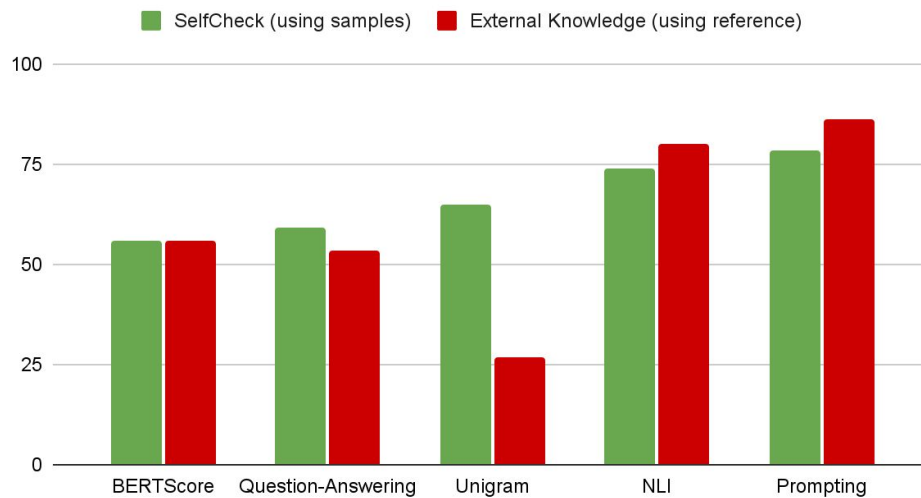
→ Proxy LLM probability

→ SelfCheck  
methods  
(black-box)

# Experimental Results

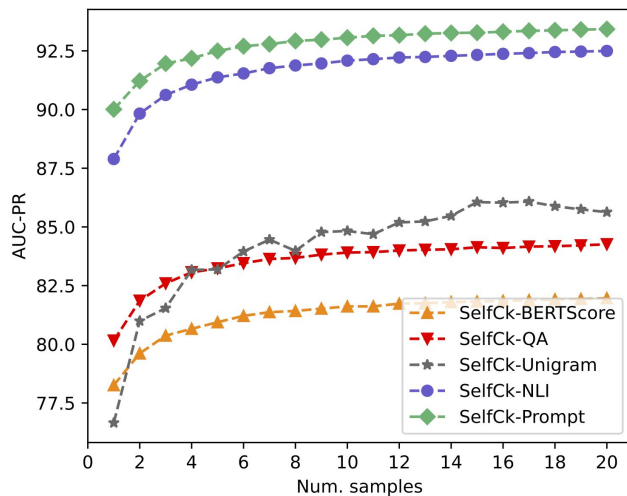
- *External Knowledge v.s. SelfCheck Samples*
  - We make use of the first paragraph in WikiBio as the reference

Passage Ranking (Spearman Rank Correlation)

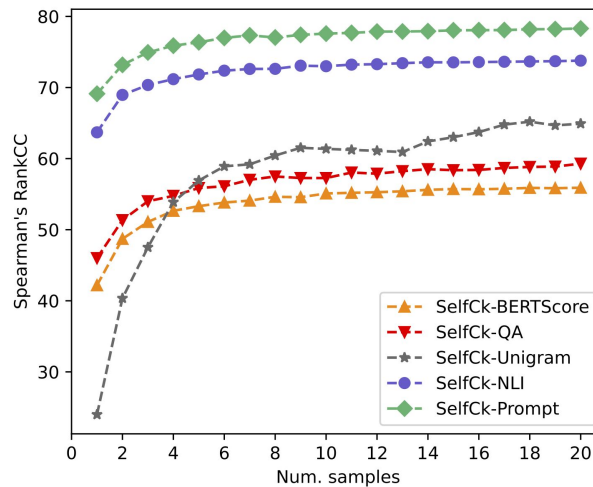


# Experimental Results

- Impact of the number of samples



Detecting Non-factual  
Sentence-level (AUC-PR)



Passage ranking  
Passage-level (Spearman)

# Conclusion

- We propose a sampling-based method that assesses factuality of LLM generation through self-consistency of generated samples
- The annotated dataset and the python package to run selfcheckgpt are available on our project page: <https://github.com/potsawee/selfcheckgpt>