Cambridge ALTA
Institute for Automated Language Teaching and Assessment

UNIVERSITY OF CAMBRIDGE

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

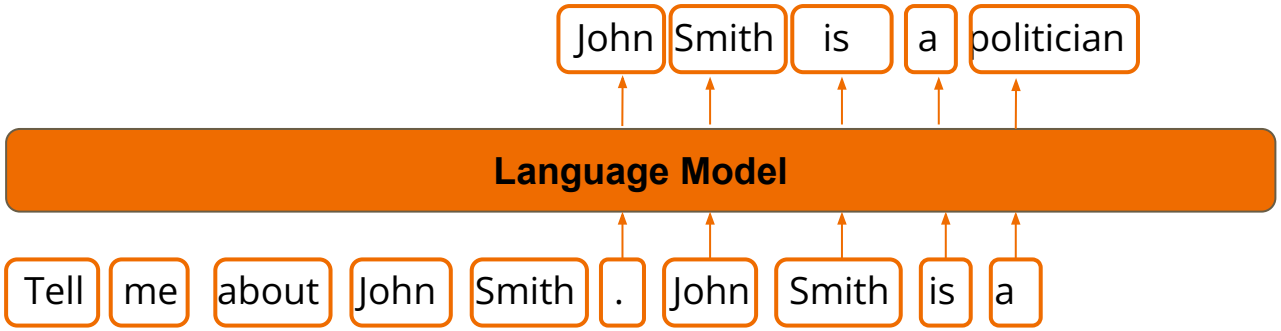
Hallucinations

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

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

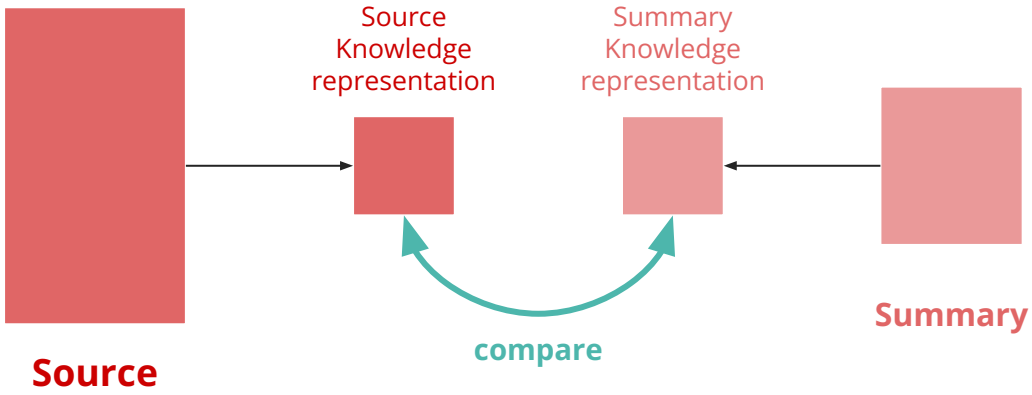
- Why LLMs Hallucinate?
  - Autoregressive language model

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

  - 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*

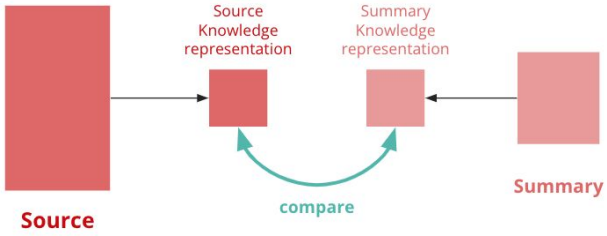
- Assessing Information Consistency in Summarization



- 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

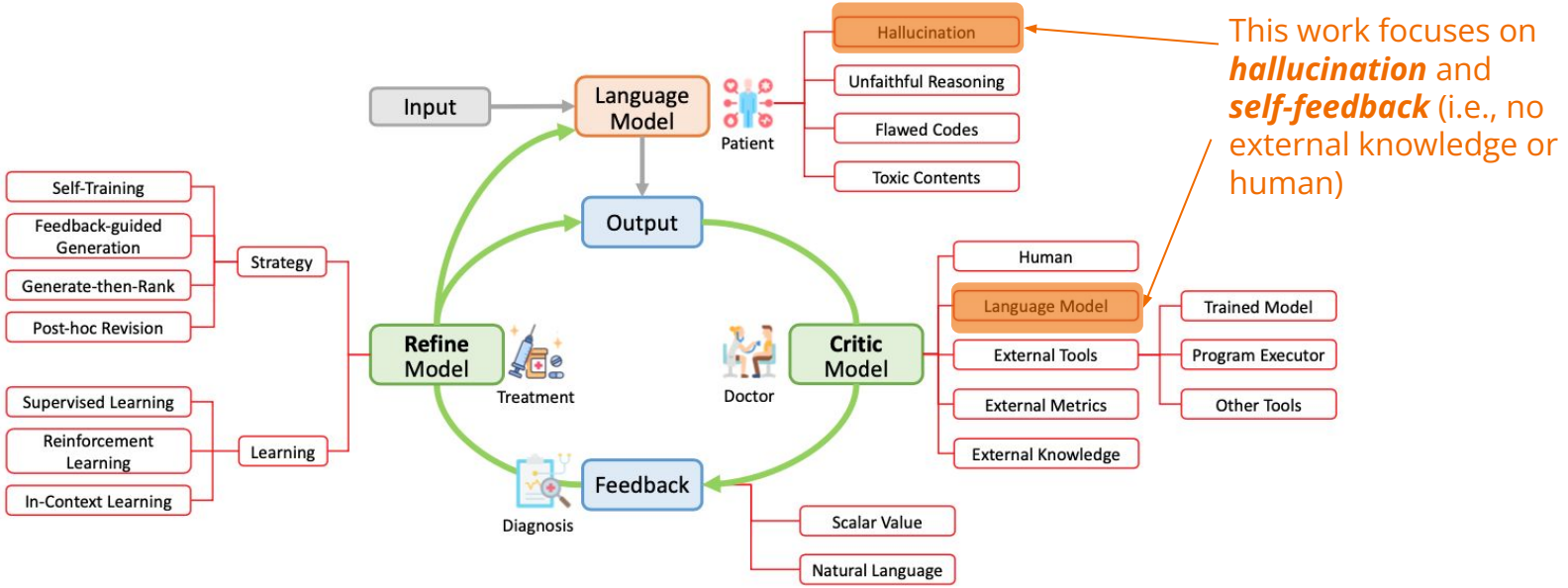


Detecting Hallucinations in *LLM generation*

- 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*



This work focuses on ***hallucination*** and ***self-feedback*** (i.e., no external knowledge or human)

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

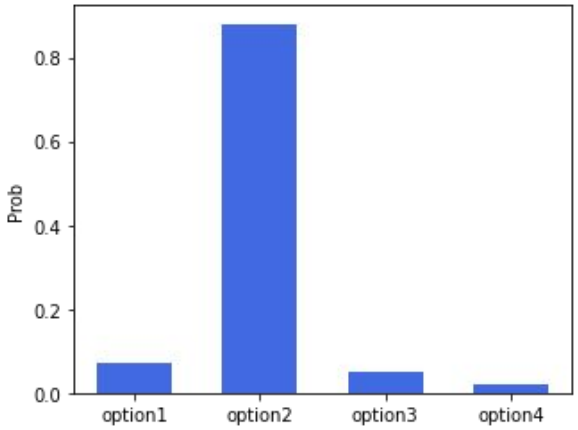
- White-box systems
  - Full access to the model's outputs and internal states
  - e.g., LLaMA, GPT-NeoX, various other open-source LLMs.
- Grey-box systems
  - Limited access to model outputs
  - e.g., GPT-3 API outputs texts and top-5 token probabilities (no internal states)
- Black-box systems
  -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

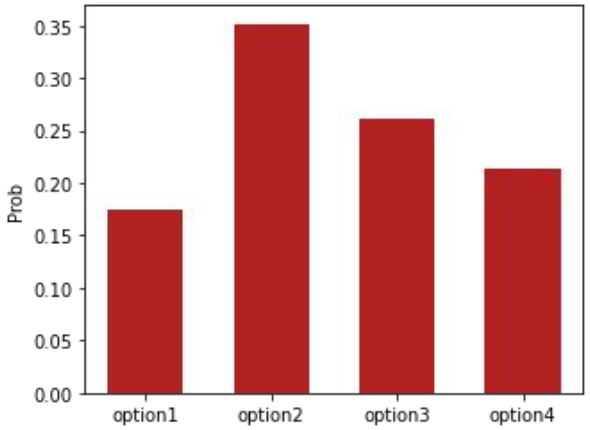
- **Zero-resource**
  - Does not require additional database / external knowledge
  - Allows one to assess any domain of LLM generation

Uncertainty Measures



Low Entropy (low uncertainty)

- The model is *more* certain about its prediction

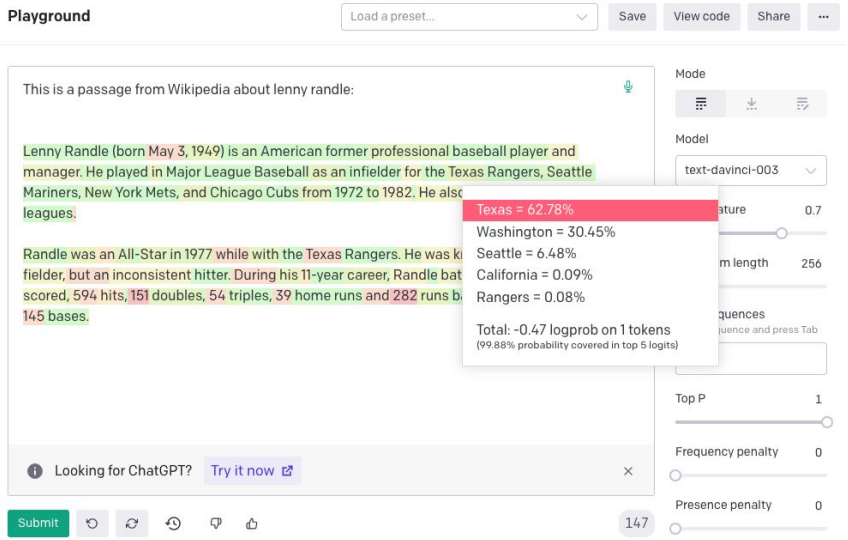


High Entropy (high uncertainty)

- The model is *less* certain about its prediction

Token-level Probability

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



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

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

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

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

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

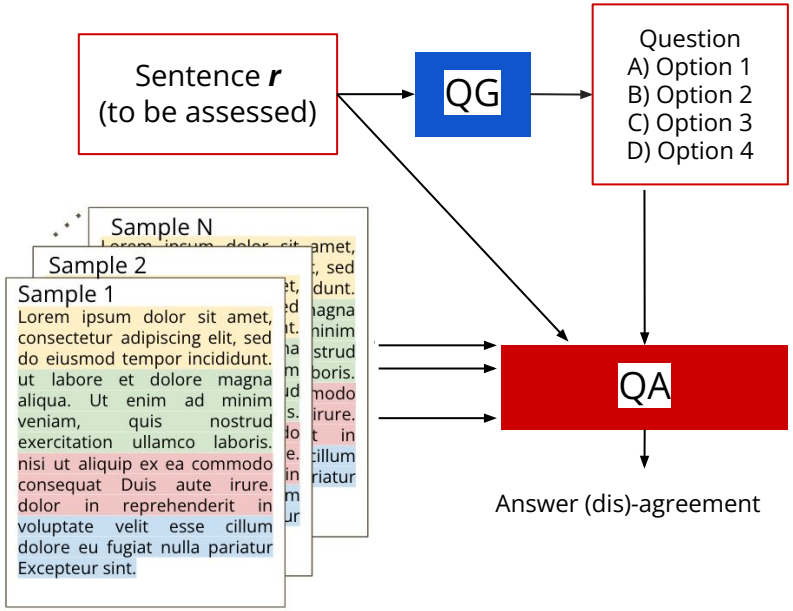
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt. ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris. nisi ut aliquip ex ea commodo consequat Duis aute irure. dolor in reprehenderit in voluptate velit esse cillum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



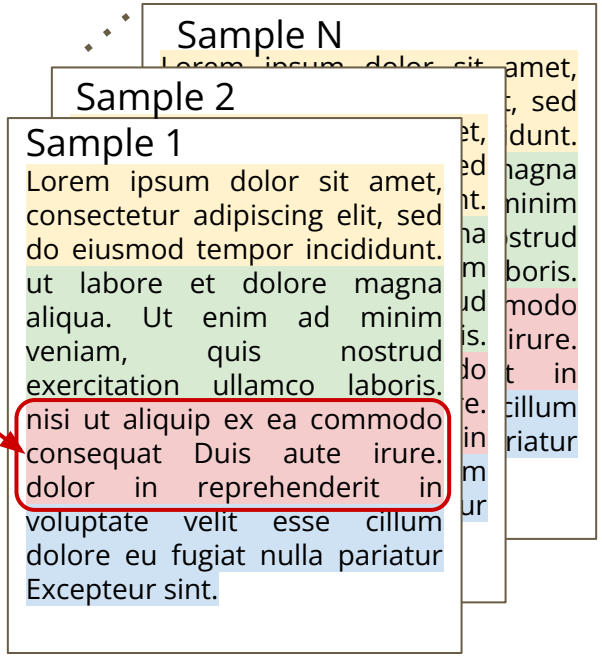
- For each sentence ***r***
  - Generate a set of {Question, Options}
- For each sample s^n and ***r***
  - Obtain the answer given s^n
  - Obtain the answer given ***r***
- Inconsistency Score

$$= \frac{\#\text{mismatches}}{(\#\text{mismatches} + \#\text{matches})}$$

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

SelfCheck with BERTScore

Sentence ***r***
(to be assessed)

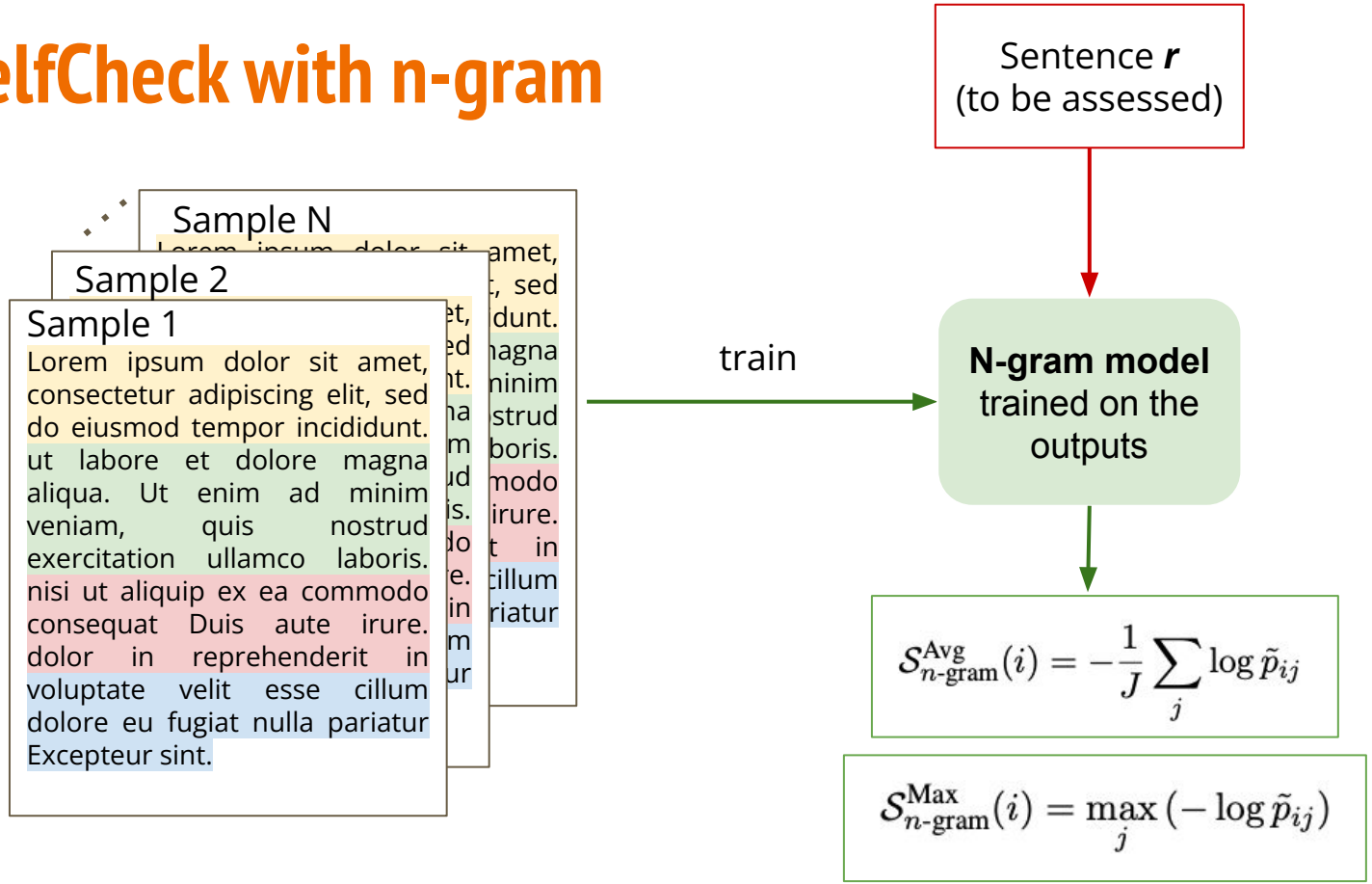


$$\mathcal{S}_{\text{BERT}}(i) = 1 - \frac{1}{N}\sum_{n=1}^{N} \max_{k} \left(\mathcal{B}(r_i, s_k^n)\right)$$

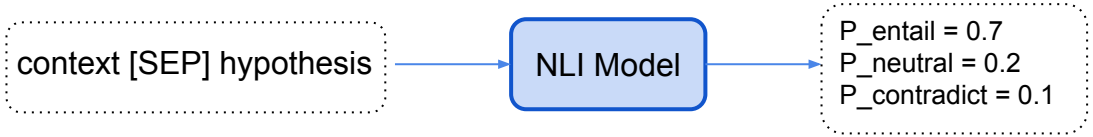
1. Obtain BERTScore of the sentence s_k that is the most similar to ***r***

2. Repeat step1 for all outputs 1…N

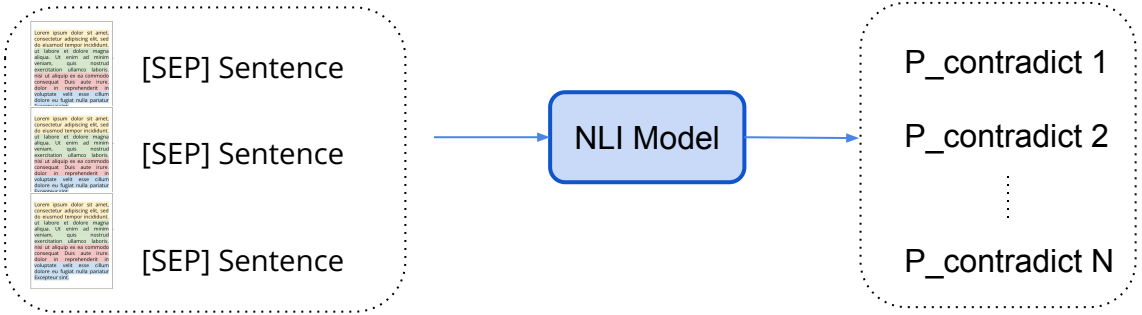
SelfCheck with n-gram



SelfCheck with Natural Language Inference (NLI)

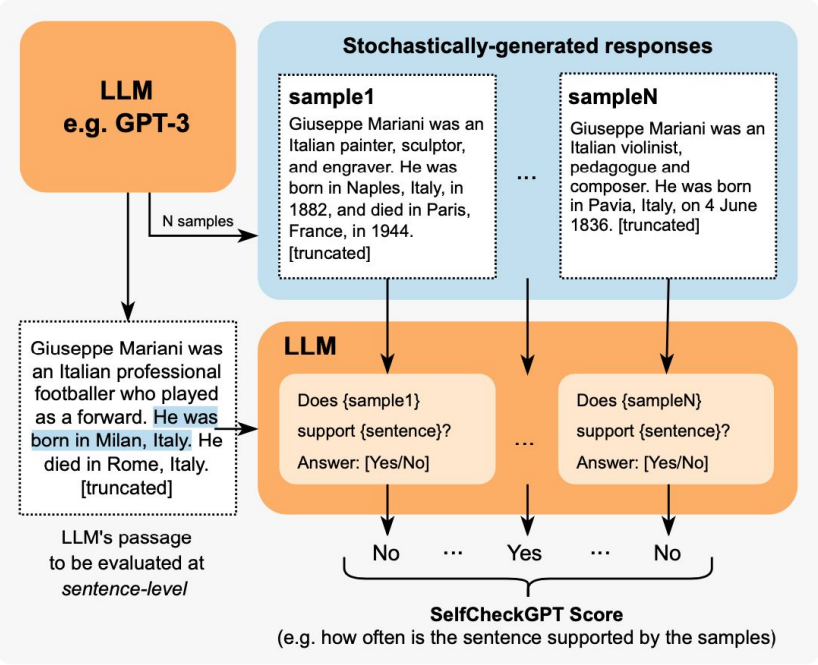


DeBERTa-v3 fine-tuned on Multi-NLI



Score = Avg(P_contradict)

SelfCheck with LLM prompting

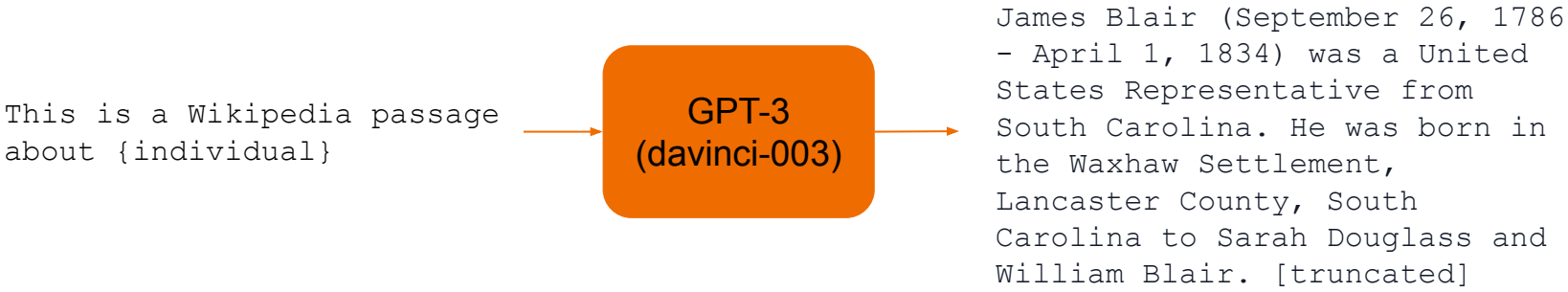



```
---------------------------------------------
Context: {}
Sentence: {}
Is the sentence supported by the context above?
Answer Yes or No:
---------------------------------------------
```

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

Experiments

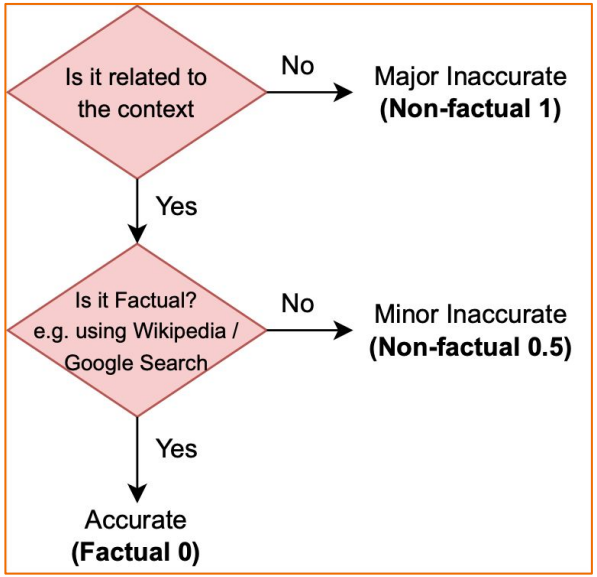
- Data
  - Previous work investigated perturbed/corrupted texts, e.g., HaDes*
  - More realistic LLM hallucination requires annotation for a specific LLM
    - We prompt GPT-3 (davinci-003) to generate passages about individuals in WikiBio



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

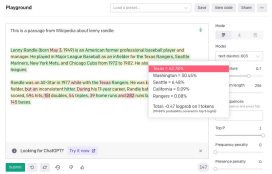
Experiments

- 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	**53.97**	53.93
GPT-3 Avg(H)	80.73	52.07	50.87
GPT-3 Max(-logP)	**87.51**	50.46	**55.69**
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	**53.97**	53.93
GPT-3 Avg(H)	80.73	52.07	50.87
GPT-3 Max(-logP)	**87.51**	50.46	**55.69**
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
SelfCheck-Prompt	**93.42**	**67.09**	**78.30**

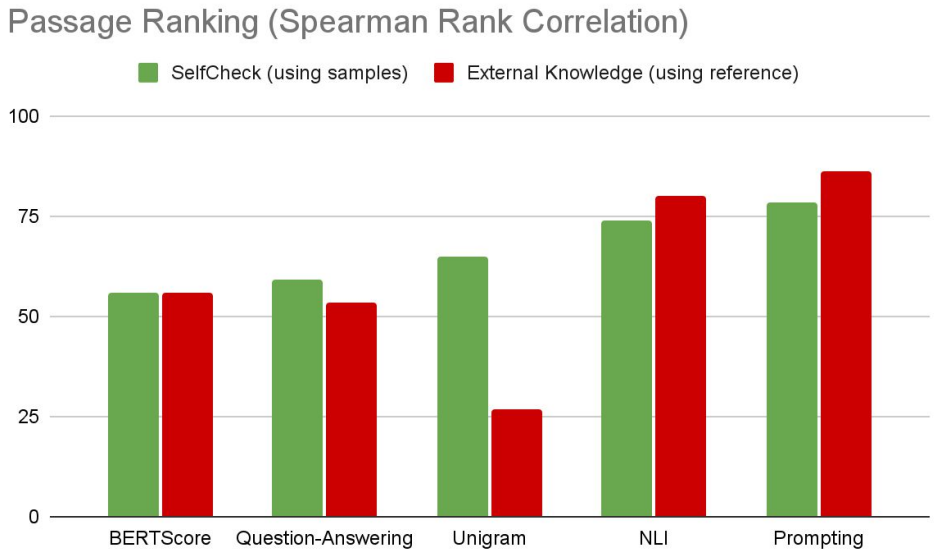
Grey-box approach

Proxy LLM probability

SelfCheck methods (black-box)

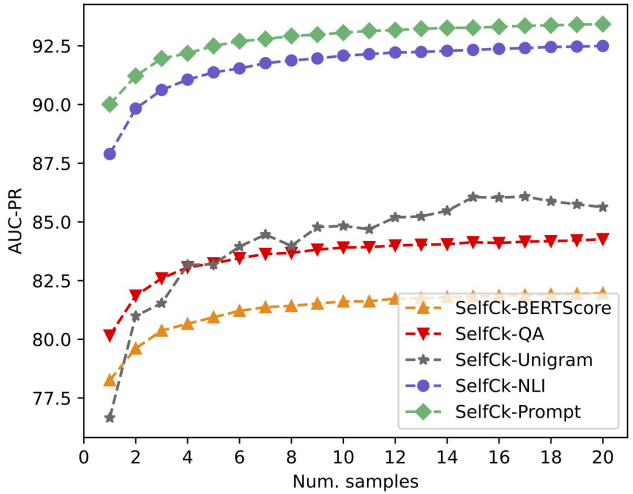
Experimental Results

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

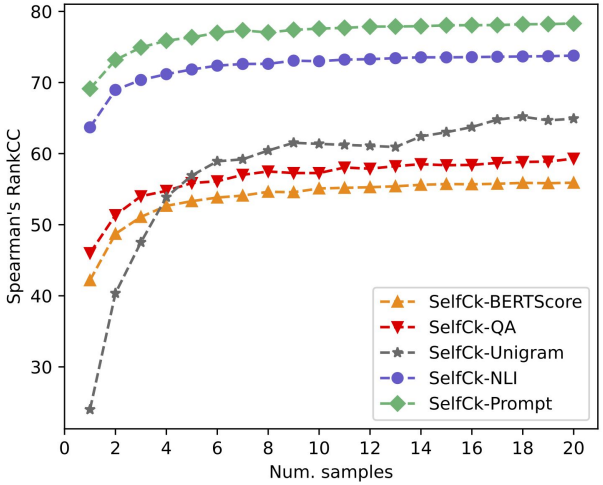


Experimental Results

- 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