# IN-CONTEXT LEARNING FOR SCALABLE AND ONLINE HALLUCINATION DETECTION IN RAGS

Nicolò Cosimo Albanese

Amazon Web Services (AWS), Milan, Italy

## ABSTRACT

*Ensuring fidelity to source documents is crucial for the responsible use of Large Language Models (LLMs) in Retrieval Augmented Generation (RAG) systems. We propose a lightweight method for real-time hallucination detection, with potential to be deployed as a model-agnostic microservice to bolster reliability. Using in-context learning, our approach evaluates response factuality at the sentence level without annotated data, promoting transparency and user trust. Compared to other prompt-based and semantic similarity baselines from recent literature, our method improves hallucination detection F1 scores by at least 11%, with consistent performance across different models. This research offers a practical solution for real-time validation of response accuracy in RAG systems, fostering responsible adoption, especially in critical domains where document fidelity is paramount.*

## KEYWORDS

*Large Language Models, Hallucinations, Prompt Engineering, Generative AI, Responsible AI.*

## 1. INTRODUCTION

Recent advancements in the field of Natural Language Processing (NLP) have witnessed the diffusion of Large Language Models (LLMs), such as BERT [1] and subsequent larger models, which have showcased remarkable capabilities in generating human-like text.

However, despite their prowess, LLMs encounter inherent limitations, particularly in handling events occurred after their training and retrieving rare or uncommon information [2, 3], as well as showing “hallucinations”, i.e., the creation of content that lacks either coherence or fidelity to the input source [4], hindering performances and posing obstacles for the adoption of LLMs in real-world applications.

To address these shortcomings, Retrieval Augmented Generation (RAG) techniques have emerged [5, 6, 7]. These methods enhance the capabilities of generative AI models by incorporating factual information obtained from external sources. In particular, RAG combines the text generation task with a retrieval operation. Given an input query, relevant facts are fetched from external sources. These facts are then used to enrich the prompt and manufacture the request to the LLM performing the text generation task. This augmentation aids the model in providing more reliable and accurate responses, as they are grounded in pertinent contextual data [8, 9].

David C. Wyld et al. (Eds): SEC, NLPML, CYBI, SIGL, CRIS, AIAPP, CoSIT- 2024
pp. 31-42, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.140703

While contextual information retrieval mitigates the generation of factually unreliable responses, the persistence of hallucinations represents an ongoing challenge. The quantification of hallucinations in responses generated by RAG remains a significant concern [4, 10].

This is a significant obstacle for the widespread adoption of Generative AI, particularly in sectors reliant on trustworthy information dissemination. For example, a medical application where a LLM generates hallucinated summaries of clinical records may pose risks for patients. In general, a customer-facing assistant in any industry may cause issues if it provides wrong information to clients.

In this paper, we propose a methodology for the assessment of factuality in RAG applications. Our approach consists in asking the model to self-evaluate the faithfulness of the generates response with respect to the context. The self-evaluation is achieved through in-context learning, i.e., by providing examples of similar evaluations enriched with a justification to elicit model reasoning.

The influence of prompt engineering on stimulating reasoning in generative AI has been thoroughly examined in prior studies [11, 12]. Through in-context learning, LLMs have demonstrated the capability to achieve performances comparable to humans across various tasks [13]. Previous research has also utilized prompting techniques to enable LLMs to autonomously evaluate their generated text across various dimensions, particularly to address concerns related to fairness and bias [14, 15].

Our unique contribution consists of a hallucination detection procedure that is simple and accurate, and can be used for online, real-time applications, improving trustworthiness and transparency of RAG solutions, and fostering their safe adoption in business scenarios. Our methodology does not require a labelled ground truth, nor training and fine-tuning procedures. We test this approach with different LLMs to prove its scalability and generalization, and benchmark its performance against existing techniques to assess its viability for industrial deployment. In conclusion, we propose an approach that can be easily deployed with a RAG solution as a microservice.

This paper is organized as follows. In section 2, we review the literature regarding hallucination detection techniques for LLMs and RAG solutions. We present our strategy for online evaluation of factuality in section 3. In section 4, we discuss the generation of a dataset for our experimental settings. In section 5, we describe the conducted experiments and achieved results, and in section 6 we draw our conclusions.

## 2. RELATED WORK

Several approaches were proposed in literature to detect and quantify hallucinations in text generation tasks. In general, we can distinguish the proposed methods as based on prompt engineering, training or fine-tuning a model, or relying on semantic similarity measures.

Several authors have suggested the idea of predicting factuality using a few-shot prompting strategy. Zhang et al., 2023 [16] proposed a few-shot prompting strategy to achieve a unified grounding entailment method for both fact and fairness checking in generated text. James et al. (2023) introduced the Retrieval Augmented Generation Assessment (RAGAS) framework, which assesses RAG applications on various criteria, including faithfulness. Faithfulness is evaluated using prompt engineering, wherein a large language model (LLM) is tasked with extracting additional single sentences from generated text and determining if they can be logically inferred from a provided context. The level of faithfulness is quantified by a numeric score derived from the ratio of inferred sentences to the total number of extracted sentences. In general, in-context learning methods provide simplicity and readiness, although they are sensitive to the prompt formulation and wording [17]. Manakul et al., 2023 [18] proposed the SelfCheckGPT framework, in which the same prompt is repeated multiple times to sample different answers. The rationale behind this approach is that a model will tend to provide the same answer to known questions, while hallucinated answers to unknown questions may diverge and be contradictory one another. The consistency of multiple collected responses can serve as a proxy for factuality. Despite the documented high performances in hallucination detection, it may be difficult to deploy this approach at scale in an online evaluation process, as the sampling strategy may be time consuming and costly due to the elevated number of repeated calls. Besides, the exact number of calls to perform may require tuning and vary based on the model, task and domain. Falcon et al. 2023 [19] proposed ARES (Automated RAG Evaluation System). By leveraging a human-annotated set, it first generates synthetic questions and answers pairs from in-domain passages, then fine-tunes lightweight LLM as judges to evaluate the RAG system. Although this approach reported high performances, companies adopting a RAG solution for the first time or experimentation may lack the resources to generate a high-quality annotated dataset for fine-tuning, and may want a faster and automated approach to model validation. Asai et al. 2023 [20] introduced Self-RAG, a framework that trains a language model to retrieve relevant text passages, generate responses, and critique its own text generation. It does this by having the model predict tokens from its original vocabulary as well as new reflection tokens. The reflection tokens allow the model to critique and reflect on retrieved passages and its own generated text. Khrisna et al. 2024 [21] proposed GenAudit, a tool trained and evaluated on the fact-checking tasks to assist the detection of hallucinated content in LLM responses for document-grounded tasks.

A different approach to hallucination detection insists on comparing the model generated answer with a human validated reference in terms of semantic similarity. The idea behind this approach is that the higher is the hallucination degree, the lower will be the semantic similarity. BERTScore [22] generates contextual embeddings for both generated and reference text using a pre-trained BERT model. Then, it computes the similarity of two sentences as a sum of cosine similarities between their tokens' embeddings. Although BERTScore correlates better with human judgments than existing metrics, it requires a human validated reference to verify against the generated text. Such prerequisite prevents from adopting BERTScore in an online validation strategy. Similarly, BARTScore [23] leverages BART's pre-trained contextual embeddings to return a score that can measure faithfulness of a model generated text against a human reference. It estimates the log token probability of the generated text given the reference, then weighting the results and returning a score.

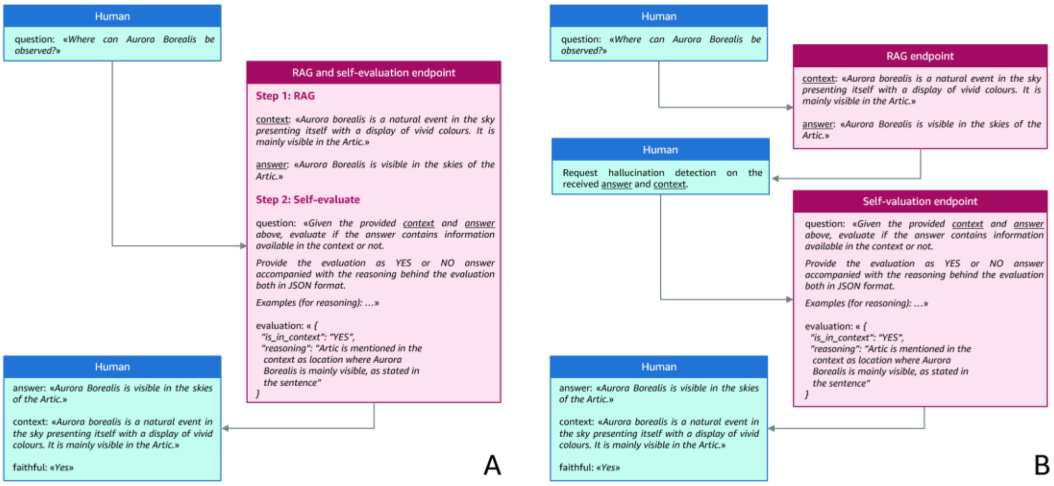
## 3. OUR METHOD FOR ONLINE HALLUCINATION DETECTION

Drawing from prior research, the proposed methodology leverages prompt engineering to stimulate reasoning in generative AI systems. Our approach centers on employing self-evaluation techniques facilitated by in-context learning, wherein the model evaluates the faithfulness of its generated responses in reference to the provided context aided by significant examples.

We directly judge the factuality of the generated answer sentence by sentence. For online applications, the input text can be split in sentences before parallelized calls to the model. The final evaluation would produce a sentence-level hallucination assessment.

With this strategy, we aim at capitalizing on the vast linguistic knowledge and representation capabilities of foundation models. In particular, we avoid fine-tuning procedures to minimize resource requirements and expediting deployment. Moreover, we do not rely on human generated ground truths, whose manufacturing is a time-consuming effort that demands the engagement of specialized subject matter experts.

This methodology presents significant advantages. Firstly, it is simple to implement. Moreover, it is scalable and can be applied with different generator models. For these reasons, this methodology can easily integrate in an existing conversational or search application, becoming a self-consistent microservice for hallucination detection to promote widespread adoption of trustworthy RAG systems. Figure 1 displays the proposed workflow for integrating this approach in a conversational service.



question: «Where can Aurora Borealis be observed?»

context: «Aurora borealis is a natural event in the sky presenting itself with a display of vivid colours. It is mainly visible in the Artic.»

answer: «Aurora Borealis is visible in the skies of the Artic.»

question: «Given the provided context and answer above, evaluate if the answer contains information available in the context or not.

Provide the evaluation as YES or NO answer accompanied with the reasoning behind the evaluation both in JSON format.

Examples (for reasoning): …»

```
evaluation: « {
  "is_in_context": "YES",
  "reasoning": "Artic is mentioned in the
   context as location where Aurora
   Borealis is mainly visible, as stated in
   the sentence"
}
```

answer: «Aurora Borealis is visible in the skies of the Artic.»

context: «Aurora borealis is a natural event in the sky presenting itself with a display of vivid colours. It is mainly visible in the Artic.»

faithful: «Yes»

question: «Where can Aurora Borealis be observed?»

context: «Aurora borealis is a natural event in the sky presenting itself with a display of vivid colours. It is mainly visible in the Artic.»

answer: «Aurora Borealis is visible in the skies of the Artic.»

Request hallucination detection on the received answer and context.

question: «Given the provided context and answer above, evaluate if the answer contains information available in the context or not.

Provide the evaluation as YES or NO answer accompanied with the reasoning behind the evaluation both in JSON format.

Examples (for reasoning): …»

```
evaluation: « {
  "is_in_context": "YES",
  "reasoning": "Artic is mentioned in the
   context as location where Aurora
   Borealis is mainly visible, as stated in
   the sentence"
}
```

answer: «Aurora Borealis is visible in the skies of the Artic.»

context: «Aurora borealis is a natural event in the sky presenting itself with a display of vivid colours. It is mainly visible in the Artic.»

faithful: «Yes»

Figure 1. Hallucination detection workflow.

Starting from a context and answer received from a previous step, in-context learning enables an online evaluation of the factuality through eliciting reasoning. The two steps of RAG and self-evaluation can be synchronous (A), i.e., returned together with the response during a conversation at expense of higher latency, or triggered by different application endpoints (B), as the self-evaluation step can be a self-consistent REST API with context and answer as input, and it can be functionally de-coupled from the conversation.

## 4. Data Preparation

For the purpose of this study, we generated a synthetic dataset made of:

- "*context*": a snippet of a document source.
- "*question*": a query related to the available context.
- "*answer*": a response to the question generated using the input context.
- "*hallucination*": binary variable indicating whether the answer is factual or not.

The data source for this dataset consisted of four e-books on different topics freely available for download from Project Gutenberg [24], and detailed in Table 1.

Table 1. Sources associated to their respective identifier on Project Gutenberg [24].

| Source ID | Subject | Title | Author and Translator |
|---|---|---|---|
| 6400 | History | The Lives of the Twelve Caesars, Complete | Suetonius (69 -122)<br>Alexander Thomson M.D. (1767-1801) |
| 14988 | Theology, Political Science | Cicero's Tusculan Disputations | Marcus Tullius Cicero (107 BC -44 BC)<br>Charles Duke Yonge (1812-1891) |
| 21076 | Science, Mathematics | The First Six Books of the Elements of Euclid | Euclid (300 BC)<br>John Casey (1820-1891) |
| 1232 | Political Science, Ethics | The Prince | Niccolò Machiavelli (1469-1527)<br>William Kenaz Marriott (1847 – 1927) |

The books were processed by splitting the textual sources in sequential chunks of around 2,000 characters each. For each chunk, we asked a LLM to generate a plausible question related to the available text, together with the factual response extrapolated from the source and a hallucinated one. In the case of the hallucinated response, the content was changed in a way to alter the meaning of the input text, while preserving the original topic. All generated answers were one sentence long. We instructed the LLM through in-context learning, i.e., manufacturing a prompt that contained examples of the data generation process given an input context. All results were reviewed by a human annotator. After human review, we obtained 4,498 answers, of which 50.07% (2,252 / 4,498) factually correct and 49.93% (2,246 / 4,498) hallucinated. The model used for this task was *Anthropic Claude v2* [25], and it was queried with low temperature (0.1) to reduce randomness of responses and increase reproducibility. Examples of contextual questions, together with factual and hallucinated answers obtained from the data generation process, are available in Appendix A.

## 5. EXPERIMENTS

### 5.1. Experimental Set-Up

The objective of this study is to assess the applicability of an in-context learning approach in detecting hallucinatory content for online applications. Our primary focus is on establishing a binary evaluation mechanism that operates at the granularity of individual sentences. In pursuit of this goal, we formulate a generic prompt intended to elicit model reasoning and obtain a self-evaluation on a generated answer with respect to its original document source. The prompt is enriched with explained examples of binary evaluations of factuality from document snippets and extrapolated sentences. The prompt template for this task is shared in Appendix B. It is important to mention that, while manufacturing the prompt, attention was put in avoiding using examples extracted from the sources in Table 1 and available as context. Domain specific examples were also deliberately excluded to ensure broader applicability and transferability across diverse domains.

As prompt-based approaches are sensitive to the wording and model being used, to verify the scalability of the approach we tested it on different LLMs: *Anthropic Claude v2* [25], *Anthropic Claude Instant* [25], *Amazon Titan Text Express* [26], *Cohere Command* [27] and *AI21 Jurassic-2 Ultra* [28]. All models are accessed through Amazon Bedrock [29] on Amazon Web Services (AWS) Cloud.

Importantly, we compare our findings with RAGAS, a framework for evaluation which is also prompt-based, and BERTScore, a numeric score based on semantic similarity.

As for RAGAS, we compare our method with the faithfulness evaluation dimension of the framework. While concerning BERTScore, although it was designed to compare the generated text with a reference, since we want to work in conditions of absence of a ground truth, we calculate the score comparing the generated answer with the original documentary source. Our assumption is that hallucinated answers may have more pronounced semantic dissimilarities with the context reference compared to factually correct text samples.

During the validation phase we observed that a minor portion of cases in all tested LLMs resulted in responses not adhering to the requested output format, despite the prompt explicitly requesting for JSON-formatted answers (Appendix B). Nevertheless, these instances were still considered in the presented analysis, as the generated texts were easily extractable through basic string post-processing techniques.

## 5.2. Results

To compare the performances of both RAGAS faithfulness and BERTScore against our strategy, we transformed their numeric output in a binary hallucination classification by finding the threshold that maximizes the Youden's J statistic, defined as (sensitivity + specificity - 1) for each possible threshold. The Receiver Operating Characteristic (ROC) curves for each metric, their Area Under the Curve (AUC) and cut-offs are displayed in Figure 2.

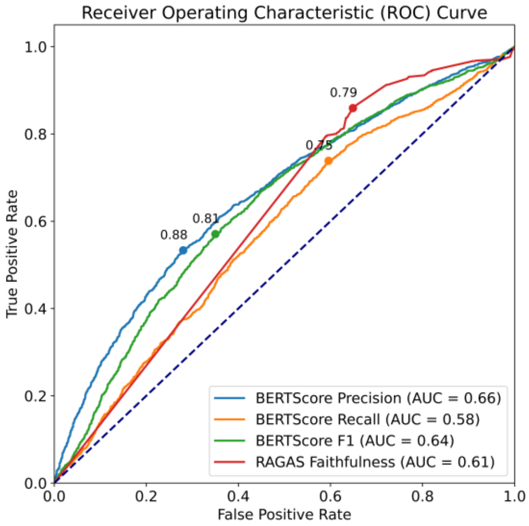


Figure 2. Receiver Operating Characteristic (ROC) curves in the hallucination detection task.

For each metric, the Area Under the Curve (AUC) is provided. The curves display the cut-off that maximizes the Youden's J statistic, defined as (sensitivity + specificity - 1), showing its numerical value. From observing the chart, we can conclude that the analyzed metrics are underfitting on the given dataset.

The performances assessed for each metric in the hallucination detection task are reported in Table 2.

Table 2. Results achieved on the hallucination detection task.

| Metrics | | Outcomes | | | |
|---|---|---|---|---|---|
| | | Accuracy | Precision | Recall | F1 Score |
| BERTScore Precision < 0.88 | | 0.62 | 0.67 | 0.48 | 0.56 |
| BERTScore Recall < 0.75 | | 0.57 | 0.55 | 0.72 | 0.63 |
| BERTScore F1 < 0.81 | | 0.61 | 0.61 | 0.61 | 0.61 |
| RAGAS faithfulness < 0.79 | | 0.61 | 0.57 | 0.86 | 0.69 |
| Our approach (in-context learning) | Anthropic Claude v2 | 0.92 | 0.88 | 0.97 | 0.92 |
| | Anthropic Claude Instant | 0.91 | 0.87 | 0.96 | 0.91 |
| | Amazon Titan Text Express | 0.82 | 0.90 | 0.73 | 0.80 |
| | Cohere Command | 0.82 | 0.81 | 0.83 | 0.82 |
| | AI21 Labs Jurassic-2 Ultra | 0.84 | 0.78 | 0.93 | 0.85 |

Our in-context learning approach showed strong gains on the hallucination detection task, increasing F1 Score by at least +11% compared to other methods, as shown in Figure 3. We also verified consistent high performance across different models without any prompt variation, suggesting our strategy could enable a model-agnostic microservice for hallucination detection.

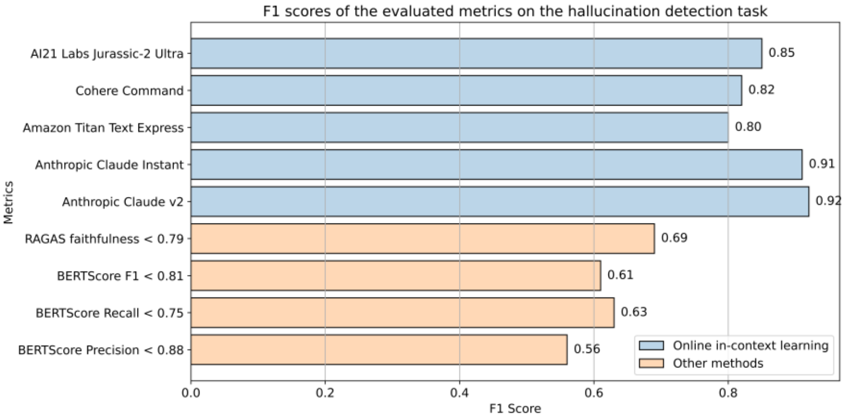


Figure 3. The F1 score of different metrics obtained on the hallucination detection task.

The online in-context learning approach provides a significant performance outlift across the tested models.

While RAGAS shares similarities with our methods, key differences drove our superior performance. Unlike RAGAS, we avoid sampling statements through an additional generative step we hypothesize introduces bias. This assumption was supported by our results. RAGAS demonstrated high recall (86%), meaning it can detect hallucinations. However, its low precision (57%) indicates many factual sentences were incorrectly flagged as hallucinated. While it may be assumed that minimizing type II errors (false negatives) over type I errors (false positives) would increase trustworthiness in RAG-based conversational systems, it is essential to recognize that both types of errors would erode user trust and hinder the adoption of the application.

The similarity-based BERTScore metrics underfitted in the hallucination detection task, with a maximum F1 score of 61%. We intentionally designed the hallucinated answers to have high surface form similarity to factual ones, as shown in Appendix A. This type of hallucination scenario, wherein the content is factually inaccurate yet contains numerous correct contextual references, is characteristic of RAG solutions, where the documents used as context in the prompt are an integral part of the response even if incorrect. While this accounts for the poorer performance of semantic similarity solutions, it also suggests that relying on these methodologies to identify hallucinations in RAGs may lead to significant inaccuracies.

It is important to mention that BERTScore was designed to compare two generated textual samples rather than evaluating a generated response against the original source, as we do in this research to simulate an online application, where a human validated response is unavailable. To provide a comprehensive analysis, we assessed BERTScore using the available ground truth, comparing hallucinated responses with factual ones. The precision, recall, and F1 scores obtained were 0.94 ± 0.03, 0.92 ± 0.04, and 0.93 ± 0.03, respectively. Therefore, despite the availability of ground truth data, BERTScore encountered challenges in distinguishing between hallucinated and factual responses within our dataset, highlighting the inherent limitations of similarity-based methodologies when facing challenging hallucination mechanisms.

The distributions of the RAGAS faithfulness and BERTScore metrics by factuality are shown in Figure 4 and 5.

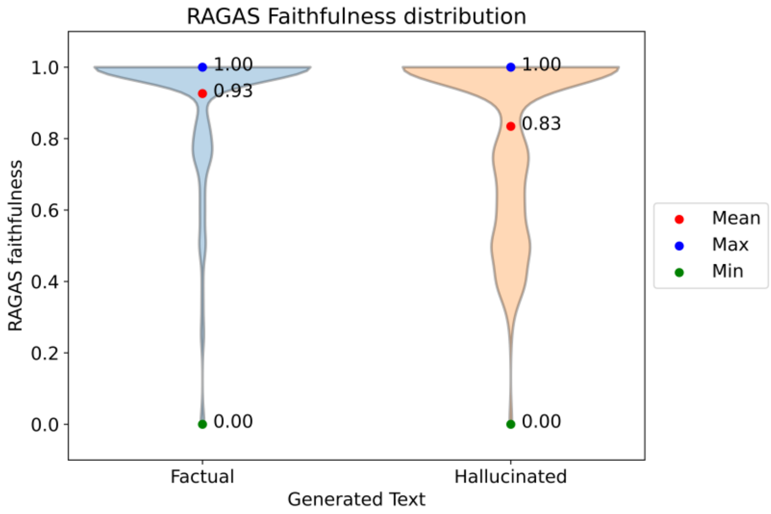


Figure 4. RAGAS faithfulness distribution by answers factuality.

Although hallucinated answers report a more prominent shift towards low scores, the distribution is still skewed towards high perceived factuality.

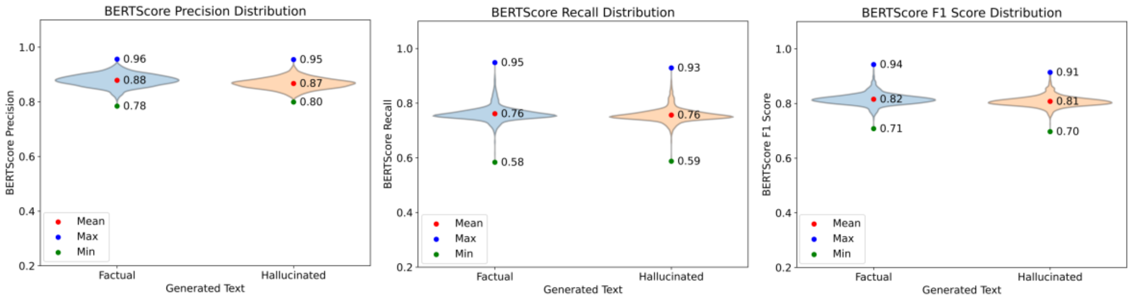


Figure 5. BERTScore precision, recall and F1 score distribution by answers factuality.

The score is assessed comparing the generated responses with the original document source.

## 6. CONCLUSIONS

In this study, we presented a promising new approach for detecting hallucinations in responses generated in RAG applications, where adherence to the knowledge base is critical.

Our in-context learning strategy yielded significant improvements compared to current techniques, boosting F1 scores by a minimum of +11% in detecting hallucinations. This performance remained consistent across various models, suggesting that our methodology could facilitate the development of a model-agnostic microservice tailored for assessing faithfulness in real-world RAG systems.

The key contribution of this work lies in the proposal for a lightweight yet accurate hallucination detection for RAGs: unlike prior work [10, 18], our approach avoids sampling bias by directly soliciting responses from the model rather than generating statements, nor it requires additional model tuning or a ground truth [19, 20, 21]. Therefore, the resulting methodology is scalable for online usage. Its simple construction, high accuracy and ease of implementation favour integration in existing systems and workflows.

While preliminary, these results highlight the potential of prompt engineering and in-context learning for improving trust and fidelity in RAG systems. Further research should explore additional datasets to avoid selection bias, even including more specialised and domain specific sources. Moreover, a wider range of LLMs should be tested. Optimization and ablation studies on the prompt template would provide useful insights into robustness to word and sentence variation, which is a known limitation of prompt-based techniques. To this aid, recent frameworks such as DSPy [30] could streamline a systematic and abstract approach to prompt optimization through a declarative module supported by the definition symbolic text transformation graph.

## REFERENCES

[1] Jacob Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding", in NAACL-HLT 2019, Minneapolis, USA, Jun 2019.

[2] Nikhil Kandpal et al., "Large language models struggle to learn long-tail knowledge", ICML 2023, Honolulu, Hawaii, USA, Jul 2023.

[3] Alex Mallen et al., "When not to trust language models: Investigating effectiveness of parametric and non-parametric memories", in ACL 2023, Toronto, Canada, Jul 2023.

[4] Ziwei Ji et al., "Survey of Hallucination in Natural Language Generation", ACM Comput. Surv., Vol. 1, No. 1, 2022.

[5] Kenton Lee et al., “Latent retrieval for weakly supervised open domain question answering”, in ACL 2019, Florence, Italy, Jul 2019.

[6] Patrick S. H. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks", in NeurIPS 2020, virtual, 2020.

[7] Kelvin Guu et al., "Retrieval augmented language model pre-training", in ICML 2020, virtual, Jul 2020.

[8] Fabio Petroni et al., "KILT: a benchmark for knowledge intensive language tasks", in NAACL-HLT 2021, virtual, Jun 2021.

[9] Alex Wang et al., "Superglue: A stickier benchmark for general-purpose language understanding systems", in NeurIPS 2019, Vancouver, Canada, 2019.

[10] Jithin James et al., "Ragas: Evaluation Framework for your Retrieval Augmented Generation (RAG) Pipelines", arXiv preprint, arXiv:2309.15217, 2023.

[11] Jason Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models", NeurIPS 2022, New Orleans, USA, Nov 2022.

[12] Karl Cobbe et al., "Training Verifiers to Solve Math Word Problems", arXiv preprint, arXiv:2110.14168, 2021.

[13] Denny Zhou et al., "Least-To-Most Prompting Enables Complex Reasoning in Large Language Models", ICLR 2023, Kigali, Ruanda, May 2023.

[14] William Saunders et al., "Self-critiquing models for assisting human evaluators", arXiv preprint arXiv:2206.05802, 2022.

[15] Rui Wang et al., "Self-Critique Prompting with Large Language Models for Inductive Instructions", arXiv preprint, arXiv:2305.13733, 2023.

[16] Tianhua Zhang et al., "Interpretable unified language checking", arXiv preprint arXiv:2304.03728, 2023.

[17] Junyi Li et al., "Halueval: A largescale hallucination evaluation benchmark for large language models", EMNLP 2023, Singapore, Dec 2023.

[18] Potsawee Manakul et al., "Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models", EMNLP 2023, Singapore, Dec 2023.

[19] Jon Saad-Falcon et al., "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems", arXiv preprint arXiv:2311.09476, 2023.

[20] Akari Asai et al., "Self-RAG: Learning to Retrieve, Generate, and Critique Through Self-Reflection", in ICLR 2024, Vienna, Austria, May 2024.

[21] Kundan Krishna et al., "GENAUDIT: Fixing Factual Errors in Language Model Outputs with Evidence", arXiv preprint arXiv:2402.12566, 2024.

[22] Tianyi Zhang et al., “Bertscore: Evaluating text generation with BERT”, in ICLR 2020, Addis Ababa, Ethiopia, Apr 2020.

[23] Weizhe Yuan et al., "Bartscore: Evaluating generated text as text generation", in NeurIPS 2021, virtual, 2021.

[24] Project Gutenberg [Online]. Available: https://www.gutenberg.org/.

[25] Anthropic, Model Card and Evaluations for Claude Models [Online]. Available: https://cdn.sanity.io/files/4zrzovbb/website/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226.pdf.

[26] Amazon Titan models [Online]. Available: https://aws.amazon.com/bedrock/titan.

[27] Cohere models documentation [Online]. Available: https://docs.cohere.com/docs/foundation-models#command.

[28] AI21 Jurassic-2 Ultra [Online]. Available: https://docs.ai21.com/docs/jurassic-2-models.

[29] Amazon Bedrock [Online]. Available: https://aws.amazon.com/bedrock.

[30] Omar Khattab et al., "DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines", NeurIps 2023, New Orleans, USA, Dec 2023.

# APPENDIX

## Appendix A

Examples of observations manufactured through the data generation procedure are reported here:

```
```

Example n. 1:

- Context (excerpt): Having been elected military tribune, the first honour he [Caesar] received from the suffrages of the people after his return to Rome, he zealously assisted those who took measures for restoring the tribunitian authority, which had been greatly diminished during the usurpation of Sylla.
- Question: What did Caesar do when he was elected military tribune in Rome?
- Factual answer: When elected military tribune, Caesar assisted those working to restore the tribunitian authority that had been diminished during Sulla's rule.
- Hallucinated answer: When elected military tribune, Caesar opposed efforts to restore the tribunitian authority that had been diminished during Sulla's rule.

Example n. 2:

- Context (excerpt): On the other hand, Castruccio reached Montecarlo with his army; and having heard where the Florentines one's lay, he decided not to encounter it in the plains of Pistoia, nor to await it in the plains of Pescia, but, as far as he possibly could, to attack it boldly in the Pass of Serravalle.
- Question: What was Castruccio's plan for attacking the Florentines?
- Factual answer: Castruccio planned to attack the Florentines boldly in the Pass of Serravalle rather than encounter them in the plains of Pistoia or Pescia.
- Hallucinated answer: Castruccio planned to await the Florentines in the plains of Pistoia rather than attack them boldly in the Pass of Serravalle.

```

## Appendix B

The prompt template used for the hallucination detection step follows:

```
Given the provided context and sentence, evaluate if the sentence contains information available in the context or not.

Provide the evaluation as YES or NO answer accompanied with the reasoning behind the evaluation both in JSON format.

Examples :

Context: "The philosophical tradition of Stoicism emerged and gained popularity in ancient Greece and Rome. The Stoic thinkers held the view that living virtuously is sufficient for attaining eudaimonia, which is a life lived excellently."
Sentence: "According to Stoicism, the secret to eudaimonia is living virtuously."
Answer: {{"is_in_context": "YES",
"explanation": "The context explicitly states that living virtuously is the path to eudaimonia as reported in the sentence."}}

Context: "The philosophical tradition of Stoicism emerged and gained popularity in ancient Greece and Rome. The Stoic thinkers held the view that living virtuously is sufficient for attaining eudaimonia, which is a life lived excellently."
Sentence: "According to Stoicism, the secret to eudamonia is living lasciviously."
Answer: {{"is_in_context": "NO",
"explanation": "The context explicitly states that living virtuously is the path to eudaimonia, as opposite to living lasciviously as stated in the sentence."}}

Context: "France is a country located primarily in Western Europe. France is a unitary semi-presidential republic with its capital in Paris."
Sentence: "Paris is the capital of France."
Answer: {{"is_in_context": "NO",
"explanation": "The context explicitly states Paris is the capital of France, as stated in the sentence."}}

Context: "France is a country located primarily in Western Europe. France is a unitary semi-presidential republic with its capital in Paris."
Sentence: "France is a parliamentary republic with its capital in Paris."

```
Answer: {{"is_in_context": "NO",
"explanation": "The context explicitly states France is a semi-presidential republic, and not a parliamentary republic as stated in the sentence."}}

Context: "France is a country located primarily in Western Europe. France is a unitary semi-presidential republic with its capital in Paris."
Sentence: "Paris has been one of the world's major centres of finance, diplomacy, commerce, culture, fashion, and gastronomy."
Answer: {{"is_in_context": "NO",
"explanation": "The sentence provides information that is not explicitly stated or deductible from the context."}}

Context: "Aurora borealis is a natural event in the sky presenting itself with a display of vivid colors. It is mainly visible in the Artic."
Sentence: "Aurora borealis is the result of disturbances in the magnetosphere caused by the solar wind."
Answer: {{"is_in_context": "NO",
"explanation": "The sentence provides information that is not explicitly stated or deductible from the context."}}

Context: "Aurora borealis is a natural event in the sky presenting itself with a display of vivid colors. It is mainly visible in the Artic."
Sentence: "Aurora borealis can be seen in the Artic but also in other countries such as Spain, Italy, Germany"
Answer: {{"is_in_context": "NO",
"explanation": "The sentence provides a list of countries that are not explicitly mentioned in the context."}}

Context: "{context}"
Sentence: "{sentence}"
Answer:
```

## AUTHOR

**Nicolò Cosimo Albanese** is a Data Scientist and Machine Learning Engineer for Amazon Web Services (AWS) Professional Services.