VerbaNexAI at SemEval-2025 Task 3: Fact Retrieval with Google Snippets for LLM Context Filtering to identify Hallucinations

Anderson Morillo and **Edwin Puertas** and **Juan Carlos Martinez-Santos**
Universidad Tecnologica de Bolivar, Cartagena Colombia
amorillo@utb.edu.co, epuertas@utb.edu.co, jcmartinezs@utb.edu.co

Abstract

This paper presents two complementary approaches for hallucination detection in large language model outputs, developed by the VerbaNexAI team for SemEval-2025 Task 3. The first approach leverages advanced LLMs, employing a chain-of-thought prompting strategy with one-shot learning and Google snippets for context retrieval, demonstrating superior performance. The second approach utilizes traditional NLP analysis techniques, including semantic ranking, token-level extraction, and rigorous data cleaning, to identify hallucinations. Evaluation of an English dataset comprising both labeled and unlabeled examples shows that the LLM-based system achieved competitive results, ranking 25th out of 41 in Intersection over Union and 28th in Spearman correlation. At the same time, the NLP approach provided valuable qualitative insights despite lower quantitative performance. These findings highlight the potential of our methods, along with challenges such as snippet availability and prompt optimization, paving the way for future improvements through enhanced snippet extraction and fine-tuning strategies.

1 Introduction

In the era of large language models (LLMs), the generation of fabricated or non-factual information often referred to as hallucinations (Ji et al., 2022; Tonmoy et al., 2024) poses a significant challenge to the reliability and trustworthiness of automated systems. SemEval-2025 Task 3 addresses this issue by identifying hallucination spans in generated texts, thereby promoting more accurate and contextually grounded responses (Vázquez et al., 2025). We centered our participation in this task on the English language, where we aim to mitigate hallucinations by incorporating external factual evidence retrieved via Google snippets (Strzelecki).

Our system builds on a multi-stage approach that integrates several key components. First, it retrieves relevant context through Google snippets, which are semantically ranked to select the most pertinent pieces of information (Strzelecki and Rutecka, 2020a). Next, a specially formatted prompt optimized through a one-shot chain-of-thought strategy guides the LLM in extracting hallucinations from the generated responses. Data cleaning and token extraction methods further refine this process, ensuring the system retains only meaningful hallucination candidates. It ultimately enhances factual verification and the robustness of LLM outputs (Tonmoy et al., 2024).

Preliminary results indicate that while our approach shows promise in identifying hallucinations, challenges remain regarding data availability and prompt efficiency. Notably, our system ranked 25th out of 41 teams in Intersection over Union and 28th in Spearman correlation. These findings underscore both the potential and limitations of our current method, motivating ongoing improvements. Our code is publicly available at https://github.com/VerbaNexAI.

2 Background

This section presents the current state of hallucination identification as proposed in Semeval Task 3 (Vázquez et al., 2025). This task aims to identify spans of hallucinations in text generated by instruction-tuned Large Language Models (LLMs) in a multilingual context. It represents the second iteration of the SHROOM task, which sought to determine whether a sentence generated by a generative model was a hallucination yes or no (Mickus et al., 2024). Authors defined hallucinations as "An unreal perception that feels real" (Ji et al., 2022).

The task involves evaluating language model outputs across 14 languages. Our participation was limited to English language datasets structured in JSON format.

Our proposal uses Google Snippets Boxes to

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1534–1541

retrieve relevant web-based information. Google developed this system, which provides quick answers to factual questions (Strzelecki and Rutecka, 2020b). Also, Google has better performance than Bing, another search engine that also offers this tool (Musa and Isa, 2021).

There are various approaches to mitigate hallucinations. (Tonmoy et al., 2024) summarizes these strategies, highlighting prompt engineering and model development as the primary methodologies.

2.1 Prompt Engineering

This technique focuses on experimenting with different instructions to obtain optimal results (Tonmoy et al., 2024). One prominent approach within prompt engineering is Retrieval Augmented Generation (RAG), which integrates relevant contextual information into the model's response generation. We can apply RAG at different stages: before (Peng et al.), during (Varshney et al.), and after (Rawte et al., 2573) generation. Additionally, end-to-end RAG solutions have been explored (Lewis et al.).

Another method is Self-Refinement through Feedback and Reasoning, where the model generates feedback on its responses to improve future iterations (Madaan et al.).

Finally, Prompt Tuning involves adjusting instructions using techniques such as fine-tuning to generate more effective responses tailored to specific tasks (Lester et al., 2021).

2.2 Developing Models

This approach focuses on improving language models through various architectural techniques to reduce the generation of incorrect information (Tonmoy et al., 2024).

One key strategy is introducing new decoding strategies, which optimize text generation to minimize errors. A method such as Context-Aware Decoding (CAD) (Shi et al., 2024).

Another approach is the utilization of knowledge graphs, where authors integrated structured information representations to improve response coherence. Examples include RHO (Ji et al., 2023) and FLEEK (Bayat et al., 2023).

Furthermore, introducing faithfulness-based loss functions helps train more reliable models by penalizing the generation of unverifiable content.

Finally, supervised fine-tuning reduces hallucinations like (Tian et al., 2023) that model learns how to respond by selecting the more factual between two responses.

For this model, we used Google snippets to analyze website content and extract the most relevant information for display in a featured snippet at the top of search results. These snippets summarize key web page details in response to user queries and can appear as lists, paragraphs, or tables. Their selection follows a structured approach to ensure accurate retrieval based on the query (Strzelecki; Strzelecki and Rutecka, 2020a). Snippets effectively provide factual context to user questions. Therefore, we propose using Google snippets, as most evaluation questions are content-based, as shown in the dataset section.

Most models discussed in (Tonmoy et al., 2024) perform binary classification to determine whether a response contains a hallucination ("yes" or "no"). However, only a few specifically address the identification of hallucination spans within the generated text. For instance, (Quevedo et al., 2024) employs two LLMs: one for generating responses and another for analyzing logs to estimate the probability of hallucination in the generated tokens. Additionally, (Liu et al.) detects hallucinations in free-form text using a token-level, reference-free approach.

3 System Overview

One of the main challenges in developing this system was designing a computationally efficient solution. To address this issue, we proposed two approaches: one based on feature extraction using linguistic analysis techniques **NLP base system** and another relying on large language models **LLM base system** the current state of the art.

3.1 Data Description

For model development, we utilized two datasets. The first was the English test dataset, comprising 50 labeled examples for initial evaluation. We used a separate test dataset with 150 examples to assess model performance. Additionally, a validation dataset containing 154 unlabeled examples was employed to analyze system behavior. These unlabeled examples could be evaluated using the platform proposed by (Vázquez et al., 2025), which applies Intersection over Union (IoU) and Spearman Correlation metrics for assessment.

The test dataset includes essential fields such as *id*, *lang*, *model_input*, *model_output_text*, and various annotations. The structure of *soft_labels* and

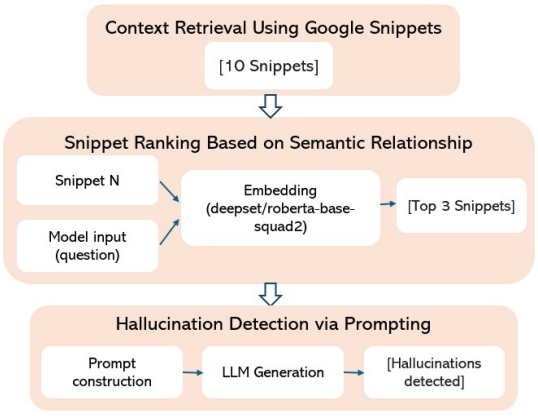

Figure 1: LLM base system

hard_labels, including keys such as start, end, and prob, provides a detailed representation of hallucination spans along with their associated probabilities, ranging from 0 to 1.

3.2 LLM base system

Figure 1 shows the model that consists of three parts. The Context Retrieval Using the Google Snippets component is essential for providing context to the system's responses and identifying hallucinations within the text. Next, the most relevant snippets are filtered using embeddings. Then, we generated a specially formatted prompt to help the language model better understand the task. Finally, we extracted the hallucinations using the LLMs and refined the results.

3.2.1 Context Retrieval Using Google Snippets

We used Google Snippets to provide context to the LLMs by retrieving search results based on the model's input. Specifically, we performed a Google search using the given query and extracted ten snippets. For example, when asked, "What are the colors of the United States flag?" Google retrieves relevant excerpts from web pages and presents them as snippets. The relevant snippets are then collected and organized into a list.

We required Proxy IP rotation to scrape the snippets, as described in (Patel, 2020), using the ScrapeOps service. Additionally, the system iterates up to two times to retrieve the data.

3.2.2 Snippet Ranking Based on Semantic Relationship

We used a semantic relationship method based on Morillo et al. (2024) to identify the most relevant snippets. This approach employs *deepset/roberta-base-squad2* embedding, which we derived from the RoBERTa model (Liu et al., 2019) and trained on Question Answering Dataset (SQuAD) from Rajpurkar et al. (2018). The method computes the semantic similarity between the snippets and the input query using cosine similarity, as proposed by Morillo et al. (2024) and Gomaa (2019). We retained snippets with a similarity score above 0.45% and selected the three most similar ones.

3.2.3 Hallucination Detection via Prompting

The next stage involves constructing the prompt shown in Figure 2, which illustrates its structure. The prompt consists of three main components: (1) an example demonstrating the expected extraction process by the model, (2) a dynamic section that adapts based on the snippets and model output, including its tokenized version that segments the text and indicates position and (3) a final instruction to organize the execution order, ensuring coherence, particularly for long-text evaluations. We tested the system using a chain-of-thought approach with one-shot and few-shot learning. The one-shot approach achieved the best performance, while the other methods were largely ineffective due to poor results.

Finally, we cleaned the data to ensure the extracted hallucinations were meaningful. We disregarded if an identified hallucination exceeded the actual tokens of the evaluated response, the probability exceeded one, or fell below 0.

3.3 NLP Techniques System

The system employs the same elements proposed in **LLM Base System** in section 3.2, **Context Retrieval Using Google Snippets**, and **Snippet Ranking Based on Semantic Relationship**, except for **Hallucination Detection via Prompting**. Instead, multiple techniques are implemented for data cleaning, followed by a token extraction process that identifies the position of each sentence within the text.

The key stage is lexical comparison. We evaluated the generated responses based on a *left outer join* operation between the six best snippets' word sets. This process helps identify potential hallucinations in the responses. However, a filtering step is applied using *Part of Speech* (PoS) analysis to avoid false positives. We considered only words belonging to the categories *NOUN, PROPN, VERB, NUM*, and *X*, using Spacy. The filtered words are

```
Here are 1 examples of hallucination detection:
Example 1:
Context: The Olympic Games are a major international sports competition.
Generated: Petra van Stoveren won a silver medal in the 2008 Summer Olympics in Beijing, China.
Generated with positions: <0>Petra<4> <6>van<8> <10>Stoveren<17> <19>won<21> <23>a<23> <25>silver<30> <32>medal<36>
<38>in<39> <41>the<43> <45>2008<48> <50>Summer<55> <57>Olympics<64> <66>in<67> <69>Beijing,<76> <78>China.<83>
Output: {"soft_labels": [{"start": 25, "end": 31, "prob": 0.9}, {"start": 45, "end": 49, "prob": 1.0}, {"start": 69,
"end": 83, "prob": 0.9}]}
Explanation: The specific medal, year and location details are not supported by context.

Now analyze this case:
Context: {['best Snippet 1', 'best Snippet 2', 'best Snippet 3']}
Generated Text:  {No, Albero Foulois was not in any of the FIFA World Cup finals.}
Generated with positions: {<1>No,<3> <5>Albero<10> <12>Foulois<18> <20>was<22> <24>not<26> <28>in<29> <31>any<33>
<35>of<36> <38>the<40> <42>FIFA<45> <47>World<51> <53>Cup<55> <57>finals.<63>}

Task:
Analyze the generated text and identify potential hallucinations by:
1. Compare the generated text with the given context
2. Mark spans that contain information not supported by the context
3. Assign probability scores (0.0-1.0) to potential hallucinations
4. Return the results in JSON format with start/end positions
```

Figure 2: One example of prompt construction for the LLM base system.

classified with a minimum probability of 90%, as shown in Figure 3.

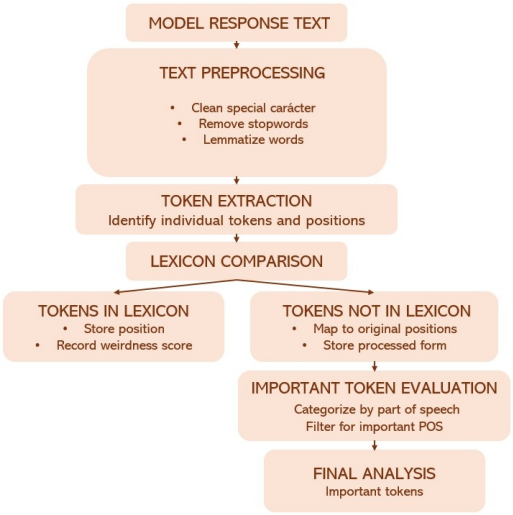

Figure 3: NLP base system proposal

4 Results

We obtained some of the results after the competition had ended due to internal issues with the code. However, tests with different models are shown in Table 4. The best model during the evaluation phase was the LLM system using *deepseek-r1-distill-llama-70b*, as proposed by (DeepSeek-AI et al., 2025). This model ranked 25th out of 41 for the IoU and 28th out of 41 in Spearman correlation on the English Dataset, as shown in Table 1.

4.1 Intersection over Union (IoU)

The IoU metric measures the overlap between predicted hallucination and reference spans. The following definitions apply:

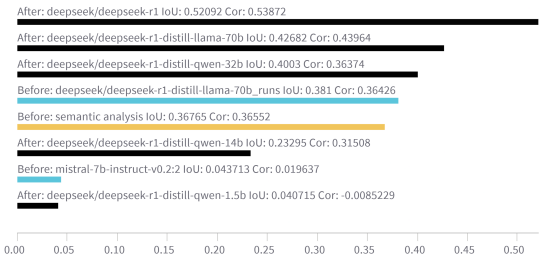

Figure 4: Performance comparison of different models before and after the evaluation phase for the English dataset in SemEval Task 3. The figure presents Intersection over Union (IoU) and correlation (Cor) scores for each model. Black bars represent post-evaluation results, while blue and yellow indicate evaluation results from different baseline approaches.

- S_R represents the set of indices corresponding to the reference hallucination spans R, which are the ground-truth hallucination positions in the text.
- S_P represents the set of indices for the predicted hallucination spans P, as identified by the model.
- IoU quantifies the similarity between S_R and S_P, ensuring that both precision and recall are considered.

$$S_R, S_P = \bigcup_{\text{span} \in R} \left\{ i \mid i \in \left[\text{span}_{\text{start}}, \text{span}_{\text{end}}\right) \right\}$$

$$\text{IoU}(S_R, S_P) = \begin{cases} 1, & \text{if } S_R = S_P = \emptyset, \\ \frac{|S_R \cap S_P|}{|S_R \cup S_P|}, & \text{otherwise.} \end{cases}$$

Table 1: Final Ranking for Intersection over Union.

Team	Position	Intersection over Union	Spearman Correlation
iai_MSU (Top performance)	1/41	0.650899	0.629443
Ours (LLM system)	25/41	0.380997	0.364264
Ours (NLP system)		0.3655	0.367

Spearman Correlation

Spearman correlation assesses how well the predicted hallucination probabilities align with the reference labels. Given soft labels with text length L, probability vectors $\mathbf{r}$ and $\mathbf{p}$ are constructed:

$$r_k, p_k = \begin{cases} r^i_{\text{prob}} & \text{if } k \in [r^i_{\text{start}}, r^i_{\text{end}}), \\ 0.0 & \text{otherwise,} \end{cases} \quad (1)$$

The Spearman correlation (ρ) is computed as:

$$\text{Cor} = \begin{cases} 1.0 & \text{if Var}(\mathbf{r}) = 0 \text{ and Var}(\mathbf{p}) = 0, \\ 0.0 & \text{if Var}(\mathbf{r}) = 0 \text{ or Var}(\mathbf{p}) = 0, \\ \rho(\mathbf{r}, \mathbf{p}) & \text{otherwise.} \end{cases} \quad (2)$$

The Spearman rank correlation coefficient ρ is defined as:

$$\rho = 1 - \frac{6 \sum_{k=1}^{L} d_k^2}{L(L^2 - 1)} \quad (3)$$

5 Ethical Considerations

The primary ethical consideration in this article is the potential bias in Google's snippet answers and the LLM responses, which can affect users' credibility judgments of the presented information (Bink et al., 2022). The system that generates featured snippets should ensure the accuracy of retrieved information. It is also essential to recognize that the filtering process created by the LLM may introduce biases due to the nature of its responses (Gallegos et al., 2024).

6 Conclusion

The system has the potential to identify hallucinations and resolve them based on context, as demonstrated in previous executions after the competition ends. Despite its low performance during the competition, We can improve the snippet extraction system to ensure data availability for each iteration. Additionally, we can optimize the prompt by testing different variations. Finally, we could apply fine-tuning techniques to train the models on the expected response format and the necessary processes to generate accurate answers.

The primary limitations encountered were related to scraping snippet boxes. Excessive requests could lead to an IP ban, requiring proxy rotation services to retrieve the information. Despite this, some snippets were still unavailable. During the testing phase, we utilized a dataset where snippets were absent for 31 of 154 data points.

Acknowledgments

We dedicate this work to the master's degree scholarship program in Engineering at the Universidad Tecnológica de Bolívar (UTB) in Cartagena, Colombia.

We want to express our gratitude to the team at the VerbaNex AI Lab [1] for their dedication, collaboration, and ongoing support of our research endeavors.

References

Farima Fatahi Bayat, Kun Qian, Benjamin Han, Yisi Sang, Anton Belyi, Samira Khorshidi, Fei Wu, Ihab F. Ilyas, and Yunyao Li. 2023. Fleek: Factual error detection and correction with evidence retrieved from external knowledge. *Preprint*, arXiv:2310.17119.

Markus Bink, Steven Zimmerman, and David Elsweiler. 2022. Featured snippets and their influence on users' credibility judgements. In *CHIIR 2022 - Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, pages 113–122. Association for Computing Machinery, Inc.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li,

[1] https://github.com/VerbaNexAI

Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Adobe Research, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey md mehrab tanjim under a creative commons attribution-noncommercial-noderivatives 4.0 international (cc by-nc-nd 4.0) license. *Computational Linguistics*, 50.

Wael Gomaa. 2019. A multi-layer system for semantic relatedness evaluation. *Journal of Theoretical and Applied Information Technology*, 97:3536.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation.

Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. 2023. RHO: Reducing hallucination in open-domain dialogues with knowledge grounding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4504–4522, Toronto, Canada. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-Tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. Technical report.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. A token-level reference-free hallucination detection benchmark for free-form text generation. Technical report.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. Technical report.

Timothee Mickus, Elaine Zosa, Raul Vazquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. SemEval-2024 task 6: SHROOM, a shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1979–1993, Mexico City, Mexico. Association for Computational Linguistics.

Anderson Morillo, Daniel Peña, Juan Carlos Martinez Santos, and Edwin Puertas. 2024. VerbaNexAI lab at SemEval-2024 task 1: A multilayer artificial intelligence model for semantic relationship detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1344–1350, Mexico City, Mexico. Association for Computational Linguistics.

Farouk Musa and Yusuf Isa. 2021. An investigation of the accuracy of knowledge graph-base search engines: Google knowledge graph, bing satori and wolfram alpha. *International Journal of Scientific and Engineering Research*, 12.

Jay Patel. 2020. *Getting Structured Data from the Internet: Running Web Crawlers/Scrapers on a Big Data Production Scale*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. Check your facts and try again: Improving large language models with external knowledge and automated feedback *. Technical report.

Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomas Cerny. 2024. Detecting hallucinations in large language model generation: A token probability approach.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *Preprint*, arXiv:1806.03822.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S M Towhidul, Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2573. The troubling emergence of hallucination in large language models-an extensive definition, quantification, and prescriptive remediations. Technical report.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Artur Strzelecki. Featured snippets results in google web search: An exploratory study. Technical report.

Artur Strzelecki and Paulina Rutecka. 2020a. Direct answers in google search results. *IEEE Access*, 8:103642–103654.

Artur Strzelecki and Paulina Rutecka. 2020b. Direct answers in google search results. *IEEE Access*, 8:103642–103654.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. *Preprint*, arXiv:2311.08401.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. Technical report.

Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sánchez-Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona de Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 Task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes.

A Appendix Google snippet example

Google snippet example shown in the web browser figure 5

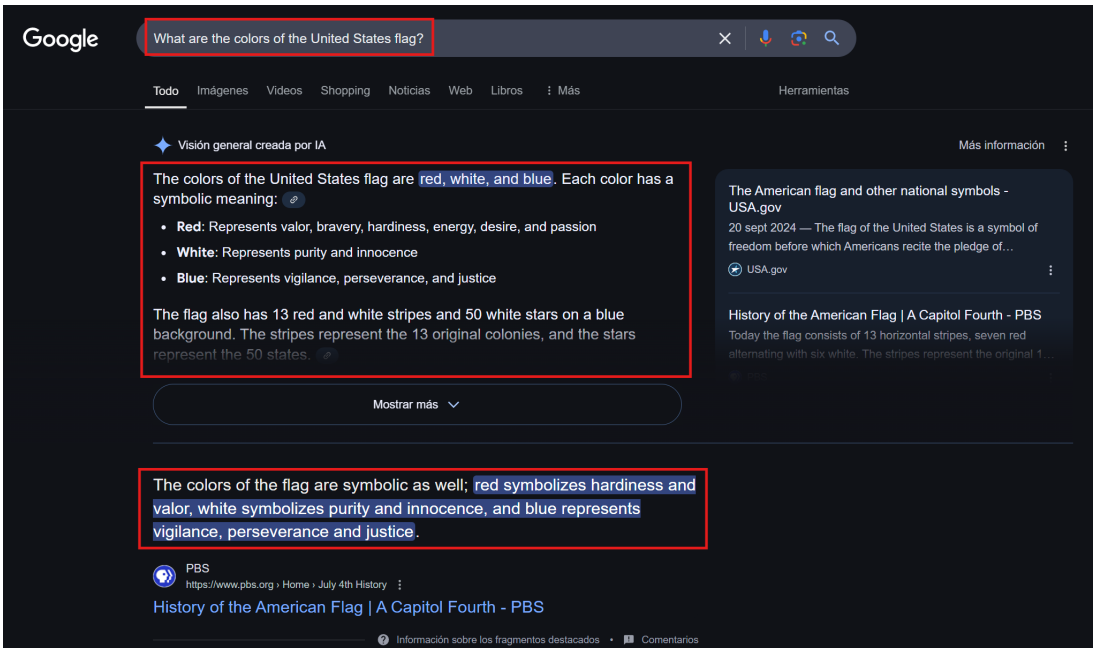

Figure 5: Google Snippets Example for the question "What are the colors of the United States flag?".