# REFIND at SemEval-2025 Task 3: Retrieval-Augmented Factuality Hallucination Detection in Large Language Models

**DongGeon Lee, Hwanjo Yu**[*]
Pohang University of Science and Technology (POSTECH)
Pohang, Republic of Korea
{donggeonlee, hwanjoyu}@postech.ac.kr

## Abstract

Hallucinations in large language model (LLM) outputs severely limit their reliability in knowledge-intensive tasks such as question answering. To address this challenge, we introduce REFIND (Retrieval-augmented Factuality hallucINation Detection), a novel framework that detects hallucinated spans within LLM outputs by directly leveraging retrieved documents. As part of the REFIND, we propose the *Context Sensitivity Ratio (CSR)*, a novel metric that quantifies the sensitivity of LLM outputs to retrieved evidence. This innovative approach enables REFIND to efficiently and accurately detect hallucinations, setting it apart from existing methods. In the evaluation, REFIND demonstrated robustness across nine languages, including low-resource settings, and significantly outperformed baseline models, achieving superior IoU scores in identifying hallucinated spans. This work highlights the effectiveness of quantifying context sensitivity for hallucination detection, thereby paving the way for more reliable and trustworthy LLM applications across diverse languages. Our code is available at https://github.com/oneonlee/REFIND.

## 1 Introduction

Detecting hallucinated information in responses generated by large language models (LLMs) has emerged as a critical challenge in the field of natural language generation (Ji et al., 2023; Zhang et al., 2023). Hallucination, in this context, refers to the generation of content that is factually incorrect or lacks grounding in verifiable sources (Li et al., 2024). This issue is particularly pronounced in knowledge-intensive tasks that demand high factual accuracy, such as question answering (Lee et al., 2022; Sun et al., 2024). The consequences of unmitigated hallucination are significant, ranging from the propagation of misinformation to a decline in trust in AI systems, underscoring the

---
[*]Corresponding author



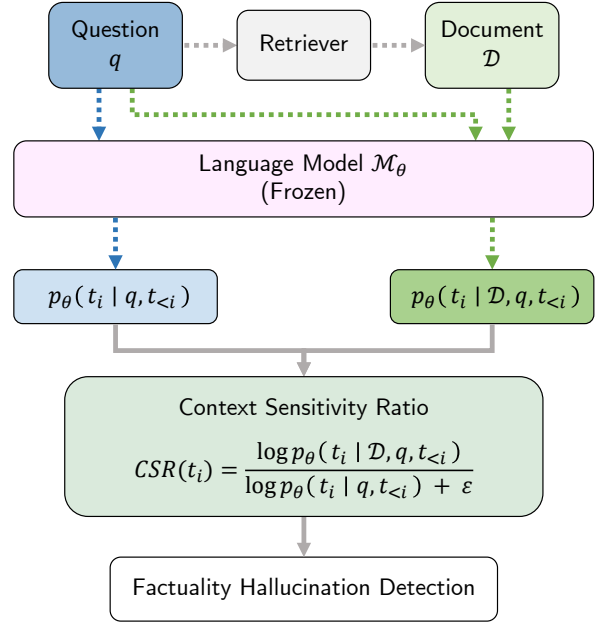Figure 1: An overview of the proposed **REFIND** method. (1) Given a question $q$, a set of relevant documents $\mathcal{D}$ is retrieved using a retriever $\mathcal{R}$. (2) A frozen language model $\mathcal{M}_\theta$ computes token probabilities $p_\theta(t_i \mid \cdot)$ for each token $t_i$, with and without the retrieved context $\mathcal{D}$. (3) The Context Sensitivity Ratio (CSR) is calculated for each token $t_i$. Tokens with the CSR exceeding a predefined threshold $\delta$ are classified as hallucinations.

need for effective hallucination detection for the development of safe and trustworthy AI.

Prior research has explored various approaches for hallucination detection. Token-level classifiers, for example, leveraging pre-trained language models like RoBERTa (Liu et al., 2019), have been employed for binary classification, labeling individual tokens as either factual or hallucinated (Liu et al., 2022). However, these models often exhibit limitations when applied to low-resource languages and tend to rely heavily on internal knowledge without effectively utilizing external evidence, which can hinder their performance. Extrinsic methods,

such as retrieval-augmented models, aim to mitigate hallucinations by integrating external knowledge. Nevertheless, existing retrieval-augmented approaches, such as FAVA (Mishra et al., 2024), can potentially lead to inaccuracies in aligning the modified responses with the original LLM output, due to their multi-step processes involving retrieval, comparison, and editing.

To address these limitations, we introduce **REFIND** (**RE**trieval-augmented **F**actuality halluc**IN**ation **D**etection), a novel framework specifically designed to identify hallucinated spans within LLM-generated text. REFIND achieves this by quantifying the context sensitivity of each token at the token level. By leveraging retrieved documents, REFIND calculates a Context Sensitivity Ratio (CSR) for each token in the LLM's response, measuring the token's dependence on external contextual information. Tokens exhibiting high CSR values are identified as likely hallucinations, offering a more direct and efficient approach to factuality verification.

Our contributions can be summarized as follows:

- We present REFIND, a novel framework for detecting hallucinated spans in LLM responses by leveraging an external retriever and calculating the CSR at the token level.

- We conduct a comprehensive evaluation of REFIND using the SemEval 2025 Task 3: Mu-SHROOM dataset (Vázquez et al., 2025), a multilingual benchmark for detecting hallucinated spans. REFIND is rigorously tested across nine diverse languages – Arabic, Czech, German, Spanish, Basque, Finnish, French, Italian, and English – demonstrating its robustness in both high- and low-resource settings.

- Experimental results demonstrate that REFIND significantly outperforms baseline models such as token-level classifiers and FAVA, achieving superior Intersection-over-Union (IoU) scores. This highlights the efficacy of the CSR in accurately identifying hallucinated content.

## 2 Related Work

**Detection of Hallucinated Responses** Several studies have proposed methods to detect whether a response contains hallucinated information. Farquhar et al. (2024); Han et al. (2024); Arteaga et al. (2025) leveraged semantic entropy (Kuhn et al., 2023) to estimate uncertainty and identify hallucinations. These approaches utilize entropy-based

metrics to assess the reliability of generated responses. SelfCheckGPT (Manakul et al., 2023) introduces a method that employs the language model itself to sample multiple responses and detect inconsistencies among them, thus identifying hallucinated outputs. However, this method relies solely on the internal knowledge of the language model, making it less effective when the model's knowledge is limited or incomplete.

**Detection of Hallucinated Spans** Beyond identifying whether a response is hallucinated, other works aim to detect specific spans of hallucinated content within a response of LLMs. Token-level classification approaches (Liu et al., 2022) utilized pre-trained language models to classify individual tokens as factual or hallucinated. These methods focus on analyzing attention patterns, demonstrating that query input tokens (defined as constraint tokens) exhibit strong correlations with factual answer tokens (Yuksekgonul et al., 2024).

FAVA (Mishra et al., 2024) proposes a retrieval-augmented pipeline that integrates retrieval, comparison, and editing steps to identify and correct hallucinated spans. While effective, the multi-step process introduces complexity and alignment challenges, particularly in ensuring that the corrected responses remain consistent with the semantics of the original output.

## 3 Method

### 3.1 Task Description

The SemEval 2025 Task 3: Mu-SHROOM (Vázquez et al., 2025) focuses on detecting hallucinated spans in responses generated by LLMs. Given an input question $q$ and its corresponding LLM-generated response (along with the model's identifier), the goal is to identify spans in the response that are hallucinated. Details of the Mu-SHROOM dataset are provided in Section 4.1.

### 3.2 Retrieval-Augmented Factuality Hallucination Detection

To address the challenge of factual hallucination detection in LLM outputs, we introduce **REFIND** (**RE**trieval-augmented **F**actuality halluc**IN**ation **D**etection). The overall workflow of the REFIND method is illustrated in Figure 1. REFIND leverages external knowledge retrieved from a relevant document set to assess the context sensitivity of each generated token.

The core principle behind REFIND is to quantify the influence of external context on the token generation process. We do this by measuring the change in the conditional probability of generating a token as information from retrieved documents is incorporated. This change is captured by the Context Sensitivity Ratio (CSR). It quantifies the degree to which the conditional probability of generating a token is altered by the inclusion of external contextual information from retrieved documents.

Let $\mathcal{M}_\theta$ denote an LLM parameterized by $\theta$, $q$ represent the input question, and $t_i$ denote the $i$-th token in the LLM's response to $q$. We use $p_\theta(t_i \mid \cdot)$ to represent the probability of generating token $t_i$ given the input. Furthermore, let $\mathcal{R}$ be a retriever that provides relevant documents based on $q$, and let $\mathcal{D} = \mathcal{R}(q)$ be the set of retrieved documents. The CSR for each token $t_i$ is defined as:

$$CSR(t_i) = \frac{\log p_\theta(t_i \mid \mathcal{D}, q, t_{<i})}{\log p_\theta(t_i \mid q, t_{<i}) + \varepsilon} \qquad (1)$$

where $t_{<i}$ represents the sequence of preceding tokens. The numerator computes the log-probability of generating $t_i$ conditioned on the question $q$, the preceding tokens $t_{<i}$, and the retrieved document set $\mathcal{D}$. The denominator computes the log-probability of generating $t_i$ conditioned solely on the question $q$ and preceding tokens $t_{<i}$, excluding the retrieved documents.[1]

By comparing these two probabilities, the CSR effectively quantifies the sensitivity of $t_i$ to the external context provided by the $\mathcal{D}$. A higher CSR indicates a stronger influence of the retrieved context on the generation of the token.

Finally, to determine whether a token is a hallucination, we compare its CSR value to a predefined threshold, denoted as $\delta$. If the CSR value for the given token $t_i$ is greater than or equal to the threshold $\delta$, we classify that the token as a hallucination. Conversely, if the CSR value is less than $\delta$, the token is not considered a hallucination. This threshold $\delta$ serves as a hyperparameter that can be tuned to optimize the balance between precision and recall in hallucination detection.

## 4 Experimental Setup

### 4.1 Dataset

We conduct our experiments on the Mu-SHROOM dataset (Vázquez et al., 2025), which consists of

outputs generated by various LLMs in response to specific input questions. Each output is annotated by human annotators to identify spans that correspond to hallucinations.

The dataset includes multiple languages, and for our study, we focus on the following nine languages: Arabic (AR), Czech (CS), German (DE), English (EN), Spanish (ES), Basque (EU), Finnish (FI), French (FR), and Italian (IT). This multilingual diversity enables a comprehensive evaluation of our method across diverse linguistic contexts.

Each data point in the dataset contains the language identifier, the input question posed to the LLM, the model name, the generated output text, and its token-level probabilities. Additionally, binary annotations specify the start and end indices of hallucinated spans, marking each such span as a hallucination.

### 4.2 Evaluation Metric

To evaluate the performance of our hallucination detection method, we adopt the IoU metric, a standard measure for span-based evaluation.

Given the set of character indices predicted as hallucinations, $\mathcal{H}_{pred}$, and the set of character indices labeled as hallucinations in the gold reference, $\mathcal{H}_{gold}$, the IoU is calculated as:

$$\text{IoU} = \frac{|\mathcal{H}_{pred} \cap \mathcal{H}_{gold}|}{|\mathcal{H}_{pred} \cup \mathcal{H}_{gold}|} \qquad (2)$$

This metric quantifies the overlap between the predicted and ground truth hallucinated spans. To handle cases where both $\mathcal{H}_{pred}$ and $\mathcal{H}_{gold}$ are empty (i.e., no hallucinations are present in either prediction or reference), we define $\text{IoU} = 1.0$ to signify perfect agreement.

### 4.3 Baseline Models

**Token-level Hallucination Classifier (XLM-R)**
We employ a token-level hallucination classifier (Liu et al., 2022) based on XLM-RoBERTa (XLM-R) (Conneau et al., 2020), a multilingual transformer model. The model is fine-tuned to perform binary classification at the token level, where each token is labeled as either hallucinated or non-hallucinated.

**FAVA** We also include FAVA (Mishra et al., 2024) as a baseline model. FAVA is a retrieval-augmented language model designed to detect and correct hallucinations in outputs generated by LLMs. The model is built upon Llama2-Chat 7B (Touvron

---

[1]To prevent division by zero, we use a small constant $\varepsilon$, which is set to $10^{-8}$.

| Method | AR | CS | DE | EN | ES | EU | FI | FR | IT | Average |
|--------|------|------|------|------|------|------|------|------|------|---------|
| XLM-R | 0.0418 | 0.0957 | 0.0318 | 0.0310 | 0.0724 | 0.0208 | 0.0042 | 0.0022 | 0.0104 | 0.0345 |
| FAVA | 0.2168 | 0.2353 | **0.3862** | 0.2812 | **0.2348** | 0.3869 | 0.2300 | 0.2120 | **0.3255** | 0.2787 |
| REFIND | **0.3743** | **0.2761** | 0.3518 | **0.3525** | 0.2152 | **0.4074** | **0.5061** | **0.4734** | 0.3127 | **0.3633** |

Table 1: Evaluation results on the Mu-SHROOM dataset (Vázquez et al., 2025) using the IoU metric across eight languages: Arabic (AR), Czech (CS), German (DE), English (EN), Spanish (ES), Basque (EU), Finnish (FI), French (FR), and Italian (IT). The proposed method, REFIND, achieves the highest average IoU score, outperforming the baselines XLM-R and FAVA in most languages, demonstrating its effectiveness for multilingual hallucination detection.

et al., 2023) and employs a two-step process: retrieval and editing. To detect hallucinations in text, we compare the edited text produced by FAVA with the original text and get the span of $\mathcal{H}_{pred}$.

### 4.4 Implementation Details

The retriever $\mathcal{R}$ used to retrieve context for RE-FIND and FAVA employs a hybrid approach, combining sparse and dense retrieval methods. Initially, a Wikipedia corpus is preprocessed for each language, including chunking, to serve as the retrieval corpus. The retriever first retrieves the top 10 relevant documents using BM25 (Robertson and Zaragoza, 2009). Subsequently, a document reranking step is performed using a pre-trained language model to select the final 5 documents to $\mathcal{D}$. To maintain consistency across the multilingual setting, we utilize `multilingual-e5-large`[2] (Wang et al., 2024) for the reranking process.

When calculating $p_\theta(t_i \mid q, t_{<i})$ in REFIND, we utilize the token probabilities of the LLM's output response provided in the Mu-SHROOM dataset. The computation of $p_\theta(t_i \mid \mathcal{D}, q, t_{<i})$ is performed using PyTorch 2 (Ansel et al., 2024). The specific prompt template employed for REFIND is illustrated in Figure 4 (Appendix A.1). More details for baselines will be discussed in Appendix A.

## 5 Result and Analysis

### 5.1 Performance Comparison

Table 1 presents the evaluation results of our proposed method, alongside the baseline models, XLM-R and FAVA, on the Mu-SHROOM dataset. The results are reported across nine languages (AR, CS, DE, EN, ES, EU, FI, FR, IT) and averaged to provide an overall assessment of performance.

REFIND outperforms the baseline models in terms of average IoU scores. The improvements are

particularly notable in low-resource languages such as Arabic, Finnish, and French, where REFIND achieves IoU scores of 0.3743, 0.5061, and 0.4734, respectively, compared to significantly lower scores from the baselines. This indicates that REFIND effectively leverages retrieval-augmented information to enhance hallucination detection in diverse linguistic settings.

### 5.2 Baseline Comparison

The XLM-R-based token classifier performs poorly on average, with an IoU of 0.0345. Its reliance solely on intrinsic model knowledge without leveraging external context limits its ability to identify hallucinated spans accurately, particularly in low-resource languages.

FAVA exhibits better performance than XLM-R, with an average IoU of 0.2787. This improvement can be attributed to its use of retrieval-augmented information for detecting and editing hallucinated text. However, FAVA's two-step process introduces complexity and potential inaccuracies in aligning the edited text with the original output.

REFIND outperforms both baselines with an average IoU of 0.3633, highlighting its superior ability to integrate retrieved context directly into the token generation process for hallucination detection. This streamlined approach ensures accurate and efficient identification of hallucinated spans.

### 5.3 Analysis of Multilingual Performance

REFIND demonstrates robust performance across both high-resource and low-resource languages. This indicates the generalizability of its retrieval-augmented approach to varying linguistic contexts. Notably, performance varies considerably across languages for all methods; for instance, XLM-R and FAVA struggle significantly with low-resource languages like Arabic, Finnish, and French. In contrast, REFIND's integration of external retrieval

---

[2] https://huggingface.co/intfloat/multilingual-e5-large

Figure 2: Example result of REFIND's hallucination detection. The gold reference $\mathcal{H}_{gold}$ highlights the correct hallucinated span, while REFIND successfully identifies the hallucinated span in the output, demonstrating its alignment with the gold annotations. The complete text of the retrieved documents is available in Appendix B.
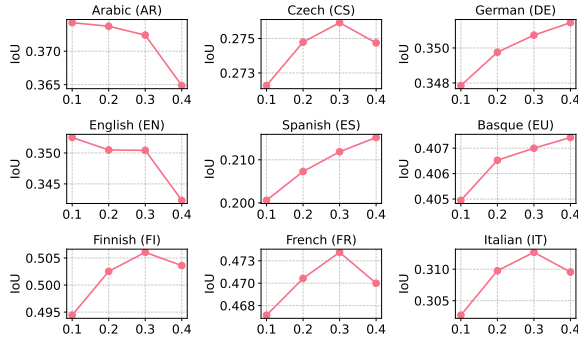


Figure 3: Analysis of IoU scores across different threshold values ($\delta \in 0.1, 0.2, 0.3, 0.4$). Each subplot represents a different language, showing the relationship between threshold values and IoU scores.

with the LLM's internal knowledge helps mitigate performance drops in these settings.

### 5.4 Analysis of Threshold Sensitivity

Figure 3 illustrates the performance of REFIND across varying threshold values (0.1-0.4) for nine languages. Most languages exhibit consistent IoU scores, indicating robustness to threshold changes. High-resource languages like English and German maintain stable scores around 0.35, while low-resource languages such as Arabic and Finnish show slightly larger variations, especially at lower thresholds. This suggests that the choice of threshold may have a more significant impact on low-resource languages, potentially due to their inherent linguistic challenges and data scarcity. Overall, these findings emphasize REFIND's ability to maintain reliable performance across a range of threshold values while highlighting potential areas for optimization in low-resource scenarios.

### 5.5 Case Study

Figure 2 illustrates REFIND's ability to detect hallucinations by utilizing retrieved evidence. The question asks about Chance the Rapper's debut year. The LLM's output contains a hallucinated span ("<mark>2011</mark>"), which is inconsistent with the retrieved documents. By comparing the generated output with external knowledge, REFIND effectively identifies spans that deviate from factual information.

## 6 Conclusion

In this study, we introduced REFIND, a novel framework for detecting hallucinated spans in LLM-generated outputs by leveraging retrieved documents to compute the Context Sensitivity Ratio (CSR) at the token level. REFIND was rigorously evaluated on the multilingual SemEval 2025 Task 3: Mu-SHROOM dataset, demonstrating superior performance across nine languages, including low-resource settings, compared to baseline approaches. By directly integrating retrieved context into the token probability calculation, REFIND effectively identifies hallucinated spans with greater precision and efficiency.

Our experimental results highlight the robustness and scalability of REFIND in multilingual environments, offering a promising solution for enhancing the factuality of LLM outputs. Moreover, the streamlined detection process avoids the complexities associated with multi-step frameworks, enabling practical deployment in real-world applications.

For future work, we aim to extend REFIND by exploring adaptive thresholding mechanisms to further optimize the balance between precision and recall in hallucination detection.

## Limitations

While REFIND achieves notable improvements in hallucination detection, there are limitations to consider. First, the reliance on retrieved documents means that the quality of the retriever directly impacts performance. Errors in retrieval or limited availability of relevant documents may lead to suboptimal CSR calculations and misclassification of hallucinated spans. Second, the approach involves computational overhead associated with calculating token probabilities with and without retrieved context, which could pose challenges in low-latency applications. Lastly, REFIND focuses on detecting factual hallucinations, and its performance in non-factoid question answering (Bolotova et al., 2022; Lee et al., 2025) remains unexplored. Further studies are needed to assess its ability to detect hallucinations in non-factoid QA tasks.

## References

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '24, page 929–947, New York, NY, USA. Association for Computing Machinery.

Gabriel Y. Arteaga, Thomas B. Schön, and Nicolas Pielawski. 2025. Hallucination detection in LLMs: Fast and memory-efficient finetuned models. In *Northern Lights Deep Learning Conference 2025*.

Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W. Bruce Croft, and Mark Sanderson. 2022. A nonfactoid question-answering taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1196–1207, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Jiatong Han, Jannik Kossen, Muhammed Razzak, Lisa Schut, Shreshth A Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. In *ICML 2024 Workshop on Foundation Models in the Wild*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

DongGeon Lee, Ahjeong Park, Hyeri Lee, Hyeonseo Nam, and Yunho Maeng. 2025. TypedRAG: Type-aware multi-aspect decomposition for non-factoid question answering. *arXiv preprint arXiv:2503.15879*.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 34586–34599. Curran Associates, Inc.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection

benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *The First Conference on Language Modeling*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Hao Sun, Hengyi Cai, Bo Wang, Yingyan Hou, Xiaochi Wei, Shuaiqiang Wang, Yan Zhang, and Dawei Yin. 2024. Towards verifiable text generation with evolving memory and self-reflection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8211–8227, Miami, Florida, USA. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. 2024. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. In *The Twelfth International Conference on Learning Representations*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

# A  Implementation Details

All experiments are conducted using NVIDIA A100 80GB GPUs.

For training the XLM-R-based (Conneau et al., 2020) system, we leverage the Trainer from the Hugging Face Transformers library (Wolf et al., 2020). We train the model using token-aligned hallucination annotations from our dataset, with the model parameters optimized using cross-entropy loss and AdamW optimizer with a learning rate of 2e-5 for 5 epochs.

Inference for FAVA (Mishra et al., 2024) is conducted using vLLM (Kwon et al., 2023), adhering to the original settings with *temperature*=0, *top_p*=1.0, and *max_tokens*=1024. The prompt template used for FAVA inference is detailed in Figure 5 (Appendix A.1).

## A.1  Prompt Details

---

**Prompt template for REFIND**

You are an assistant for answering questions.
Refer to the references below and answer the following question.

### References
{reference_passages}

### Question
{question}

### Answer

---

Figure 4: Prompt template of REFIND used to compute per-token probabilities under the conditions provided in the input context.

---

**Prompt template for FAVA**

Read the following references:
{reference_passages}
Please identify all the errors in the following text using the information in the references provided and suggest edits if necessary:
[Text] {output}
[Edited]

---

Figure 5: Prompt template for using FAVA (Mishra et al., 2024).

## B   Full Text of Retrieved Documents $\mathcal{D}$ for Case Study (§5.5)

**Document 1.** Chance the Rapper discography he discography of American rapper Chance the Rapper consists of one studio album, five mixtapes and 27 singles (including 14 singles as a featured artist). Chance the Rapper released his debut mixtape, "10 Day" on April 3, 2012. The mixtape was followed up with the release of "Acid Rap" the following year, which saw universal acclaim from music critics. Chance the Rapper then released his third mixtape, "Coloring Book" on May 13, 2016. The mixtape peaked at number eight on the "Billboard" 200 chart to continued acclaim and was supported by the singles "Angels"

**Document 2.** Juice (Chance the Rapper song) "Juice" is a song by American rapper Chance the Rapper, released on January 31, 2013 as the lead single from his second mixtape "Acid Rap" (2013). It was written by Chance and Nate Fox, who also produced the song. "Juice" is a midtempo song, built around a loop of Donny Hathaway's live performance of "Jealous Guy" by John Lennon. Chance the Rapper sings and raps in a comedic manner; his verses in the song have been described as having a "freewheeling, bluesy sway" that "gives way to raucous call-and-response choruses". He references the 1992 film "Juice" (of

**Document 3.** signs of advertisements and department stores appear in the background, some of which provide imagery and visual references of the lyrics. For example, when Chance lyrically alludes to the film "Juice", a portrait of rapper Tupac Shakur (who starred in the film) flashes across a billboard. When "Acid Rap" was first re-released on streaming services on June 28, 2019, "Juice" was replaced with a 30-second spoken message, in which Chance the Rapper explains the song is excluded from the mixtape because of an uncleared sample. Chance then adds that all streaming proceeds for the alternate

**Document 4.** Cocoa Butter Kisses "Cocoa Butter Kisses" is a song by American rapper Chance the Rapper from his second mixtape "Acid Rap" (2013). The song features American rappers Vic Mensa and Twista, and was produced by Cam O'bi and Peter Cottontale. It is one of Chance the Rapper's most popular songs to date. At the time when the song was written, Vic Mensa was staying at an apartment in Humboldt Park, Chicago with his manager Cody Kazarian. Chance the Rapper visited one day and showed Mensa a verse and hook he had written earlier. Soon, Mensa began composing his part for the song. In an interview

**Document 5.** (eight) in several of those categories. One of the most closely watched races will be Best New Hip-Hop Artist, whose nominees including Anderson .Paak, Bryson Tiller (who won that award and Best Male R&B/Pop Artist at June's BET Awards), Chance the Rapper, Desiigner and Tory Lanez. Drake – "Hotline Bling" Fat Joe & Remy Ma featuring French Montana & Infared – "All the Way Up" Kendrick Lamar Kendrick Lamar Director X DJ Khaled Metro Boomin DJ Khaled "All the Way Up" – Produced by Cool & Dre and Edsclusive Drake – "Views" Chance the Rapper DJ Khaled Kanye West Chance the Rapper

Figure 6: Complete text of documents retrieved for the input question "*When did Chance the Rapper debut?*" as referenced in the case study in Section 5.5.