Deloitte (Drocks) at SemEval-2025 Task 3: Fine-Grained Multi-lingual Hallucination Detection Using Internal LLM Weights

Alex Chandler[1], Harika Abburi[2], Sanmitra Bhattacharya[1], Edward Bowen[1], Nirmala Pudota[2]

[1]Deloitte & Touche LLP, USA
[2]Deloitte & Touche Assurance & Enterprise Risk Services India Private Limited India

{achandler, abharika, sanmbhattacharya, edbowen, npudota}@deloitte.com

Abstract

While Large Language Models (LLMs) have driven significant progress in Natural Language Generation (NLG), their propensity to hallucinate—generating factually incorrect content—remains a barrier to wider adoption. Most existing hallucination detection methods classify text at the sentence or document level, lacking the precision to identify the exact spans of text containing hallucinations, termed hallucination spans. We propose a methodology that generates supplementary context and processes it alongside the evaluated text through an LLM, extracting the internal weights (features) per token from various layers. These extracted features serve as input for a neural network classifier designed to perform token-level binary classification of hallucinations. Finally, we identify hallucination spans by mapping token-level predictions to character-level predictions. Our hallucination detection model ranked top-ten in 13 of 14 languages and first in French, evaluated on the Mu-SHROOM dataset within the SemEval: Multilingual Shared-task on Hallucinations and Related Observable Overgeneration Mistakes (Mu-SHROOM).

1 Introduction

The domain of Natural Language Generation (NLG) is witnessing a remarkable transformation with the emergence of Large Language Models (LLMs) (OpenAI, 2024; Manyika and Hsiao, 2023; Dubey et al., 2024). LLMs have been shown to outperform traditional Natural Language Processing (NLP) approaches across a wide range of applications (Kung et al., 2023; Mousavi et al., 2023). Despite the rapid advancements in LLMs, a concerning trend has emerged where these models generate hallucinations (Bang et al., 2023; Ji et al., 2023a), resulting in content that appears plausible but is factually unsupported. Hallucinations can be categorized into extrinsic errors, where claims conflict with external facts, and intrinsic errors, where claims are not fully grounded in the source material. This issue is particularly critical in sensitive domains such as healthcare, finance, and legal services, where the accuracy of generated content is paramount. Hence, the automatic detection of hallucinated content has become an active area of research, aiming to enhance the reliability and trustworthiness of LLM-generated content (Zhang et al., 2023b; Bai et al., 2024).

Recent studies have explored different methodologies for hallucination detection, including natural language inference (NLI) and factual consistency checking (Zha et al., 2023; Chandler et al., 2024; Tang et al., 2024), as well as textual entailment techniques (Sankararaman et al., 2024; Fan et al., 2024). Additionally, approaches like reference-free (Zero Context) hallucination detection have been investigated (Manakul et al., 2023; Hu et al., 2024a; Li et al., 2024b), alongside evidence retrieval methods utilizing Retrieval-Augmented Generation (RAG) or Web Search (Zimmerman et al., 2024; Tian et al., 2024; Li et al., 2024a).

However, fact-checking models often demonstrate inconsistent performance when evaluating text across different languages. Vu et al. (2024) highlights that even state-of-the-art (SOTA) LLMs, when used for fact-checking, struggle with text in low and medium-resource languages. Moreover, many popular hallucination detection methods classify hallucinations at the sentence or document level, which limits their ability to precisely identify and correct the specific text responsible for these errors.

To address this limitation, Liu et al. (2022) introduced the HaDes dataset (HAllucination DEtection dataSet), enabling fine-grained, reference-free hallucination classification at the token level. Uncertainty-based and consistency-based methods were proposed to detect token level hallucinations (Ji et al., 2023b). For example, Mitchell

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1089–1097
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

et al. (2023) utilized log-probability curve detection, while Kuhn et al. (2023) estimated semantic likelihoods by clustering generated sequences. Furthermore, Zhang et al. (2023a) explored token type and frequency for detecting hallucinations based on uncertainty. Building on these ideas, Ma and Wang (2024) developed metrics assessing token cohesiveness through successive rounds of random token deletion and measuring semantic differences.

Recent advancements have shown promise in using LLM internal states for detecting token-level hallucinations. For instance, Hu et al. (2024b) focused on identifying hallucinations by analyzing embeddings and gradients to gauge probability distribution differences, while Sun et al. (2025) applied mechanistic interpretability within RAG scenarios. Many of these solutions, however, encounter common challenges. They often rely heavily on large amounts of labeled training data and necessitate multiple inference calls for each sentence, which can be resource-intensive. They also frequently fall short in testing across low and medium-resource languages, or in conducting comprehensive multilingual evaluations.

To boost this area of research further, the SemEval[1] organizers introduced the Mu-SHROOM task. This task focuses on detecting hallucination spans in the outputs of instruction-tuned LLMs in a multilingual context. The contribution of this study are:

- We employed web search to incorporate supplementary contextual information into the model.
- We develop a binary multi-lingual token-level hallucination detection classifier, where the internal weights of LLM are used as a feature vectors. The resulting token-level predictions are then converted into character-level predictions, allowing for the precise identification of hallucinated spans within the text.
- Our model ranks within the top 10 for 13 out of 14 languages on the Mu-SHROOM dataset and secured first place in French.

2 Mu-SHROOM Dataset

The Mu-SHROOM dataset is a multilingual benchmark dataset for detecting hallucination spans in outputs generated by LLM. The dataset encompasses a diverse set of 14 languages: Arabic-Modern Standard (AR), Basque (EU), Catalan (CA), Mandarin Chinese (ZH), Czech (CS), English (EN), Farsi (FA), Finnish (FI), French (FR), German (DE), Hindi (HI), Italian (IT), Spanish (ES), and Swedish (SV).

Language	**Training Samples**	**Test Samples**
Arabic (AR)	50	150
Basque (EU)	0	99
Catalan (CA)	0	100
Chinese (ZH)	50	150
Czech (CS)	0	100
English (EN)	53	154
Farsi (FA)	0	100
Finnish (FI)	50	150
French (FR)	52	150
German (DE)	50	150
Hindi (HI)	50	150
Italian (IT)	50	150
Spanish (ES)	53	152
Swedish (SV)	49	147
Total samples	507	1902

Table 1: Sample distribution across languages in training and test sets of the Mu-SHROOM dataset.

The Mu-SHROOM dataset contains the following columns:

- **id**: a unique datapoint identifier
- **lang**: the language of the question and output text
- **model_input**: the input passed to the models for generation
- **model_id**: denoting the HuggingFace identifier of the corresponding model
- **model_output_text**: the output generated by a LLM when provided the aforementioned input
- **model_logits** : the logits from the model
- **model_tokens** : the tokens created by model
- **soft_labels**: provided as a list of dictionary objects, where each dictionary objects contains the following keys:

[1] https://helsinki-nlp.github.io/shroom/

- 'start', indicating the start of the hallucination span
- 'end', indicating the end of the hallucination span
- 'prob', the empirical probabilty (proportion of annotators) marking the span as a hallucination

- **hard_labels**: provided as a list of pairs, where each pair corresponds to the start (included) and end (excluded) of a hallucination

Table 1 provides a detailed breakdown of sample distribution within the training and testing sets across the various languages represented in the dataset. The dataset comprises 507 samples for training and 1902 samples for testing. For evaluation purpose, the shared task organizers assessed the performance of the submissions on a test set of 1902 samples. The test set labels were not disclosed to participants during the submission phase. Additional details about the task and dataset are available at (Vázquez et al., 2025).

3 Proposed Approach

In this section, we describe our proposed approach for detecting hallucination spans as depicted in Figure 1. The approach encompasses three primary components: **1)** context generation, **2)** extracting token-level internal weights from LLM, and **3)** constructing a binary classifier to produce token-level predictions, which are subsequently transformed into character-level predictions.

3.1 Context Generation

To enhance the model's understanding, we retrieve additional contextual information relevant to the model_output_text. Following Chen et al. (2022); Ousidhoum et al. (2022), we systematically decompose the model_output_text into a structured list of claims using `GPT-4o-mini`, as this decomposition allows for us to increase the recall of needed facts. Subsequently, we input this list of claims into `GPT-4o-mini` to generate queries for each claim for the purpose of fact-checking. The prompts employed for both claim decomposition and query generation are detailed in Appendix Section A. These queries are submitted to the DuckDuckGo search engine to retrieve titles, relevant text snippets, and URLs for each query. Finally, we concatenate all the search results and refer to this aggregated output as the Context.

3.2 Extracting Token-level Features

Once the context has been prepared, we format the input to include a instruction, context, model_input, and model_output_text, structured as follows:

Instruction: "Answer the following question in the language of the question and then compare your answer with the given output."
Context: context
Question: model_input
Output: model_output_text

This above input is processed by the `Llama-3.2-1B/3B-Instruct` models, and the token-level internal weights are extracted from selected layers of the LLM. The specific layers utilized include hook_attn_out, hook_resid_post, and hook_scale (hook_ln1_scale, hook_ln2_scale, ln_final_hook_scale). The motivation for leveraging these internal weights lies in their capacity to capture intricate patterns in language representation, enabling a deeper understanding of the model's decision-making processes.

The hook_attn_out feature reflects the final output of the attention mechanism, which is critical for understanding the relationships and contextual relevance between tokens. The hook_resid_post provides information about the residual connections after the normalization and attention layers in each block, ensuring the retention of critical features throughout the model. Finally, hook_scale represents the scaling factors of the layer normalization in each transformer block. Improper scaling at this stage could cause attention mechanisms to overemphasize or disregard certain tokens, potentially leading to hallucinations. More details on the dimension of each token feature vector are provided in Table 3 in Appendix Section B.

By analyzing these internal weights, we can access the model's learned knowledge and contextual embeddings, which are crucial for accurately detecting hallucination spans. This approach facilitates a more granular analysis of the interactions between tokens, enhancing the predictive performance of our hallucination detection classifier and allowing for more precise assessments of generated content.

3.3 Binary Classifier

The token-level features are subsequently provided as input to a linear classifier consisting of two

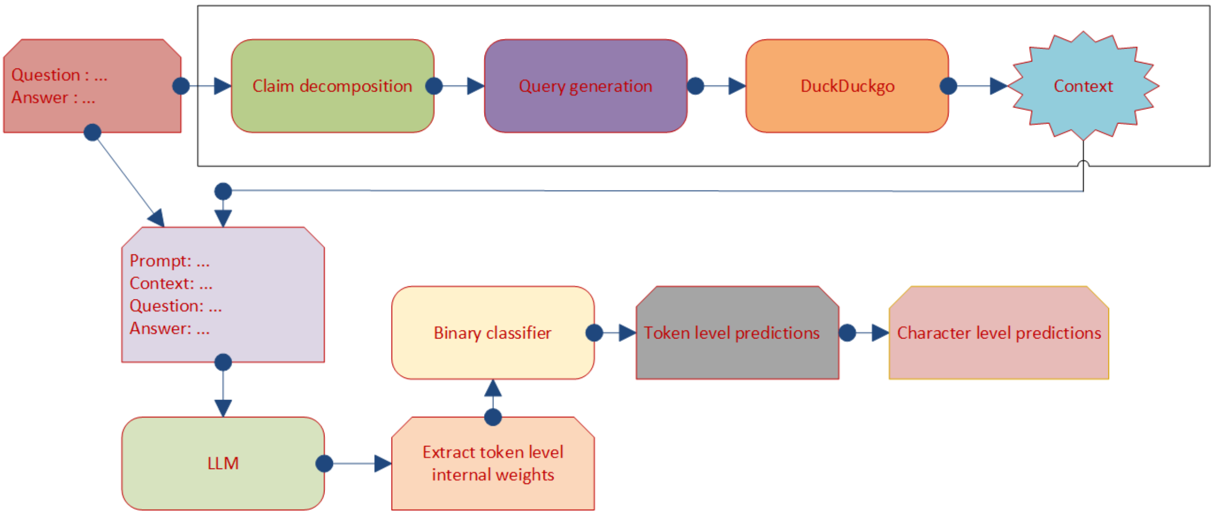

Figure 1: Our End-to-End Pipeline

fully connected layers activated by ReLU, which facilitates the generation of token-level predictions. These token-level outputs are then transformed into character-level features by performing a substring search, aligning each token produced by the language model with its corresponding character-level indices in the model_output_text. From this binary array of character-level predictions, we extract continuous sequences of hallucination spans.

4 Experimemts

This section details the experimental evaluation of our approach. To assess the effectiveness of our method, we employed two established character-level metrics such as Intersection-over-union (IoU) and Spearman correlation (S.Corr). The IoU is calculated as the ratio of the number of characters identified as hallucinations by our model to the total number of unique characters in both the predicted and actual sets. Conversely, Spearman correlation is calculated by comparing the probabilities assigned by the model that indicate a character's classification as part of a hallucination against the empirical probabilities observed from human annotators.

4.1 Results

The performance of our pipeline on the test dataset, evaluated externally, is summarized in Table 2. We report IoU scores, Spearman Correlation, and rank performance based IoU on the Mu-SHROOM Eval Leaderboard.[2] Our pipeline demonstrates competitive performance across all languages except for Chinese, securing first position in French based on IoU metrics.

4.2 Language Model Selection

We experiment with `Llama-3.2-1B-Instruct` and `Llama-3.2-3B-Instruct` models to process input and extract internal attention weights. Table 4 in Appendix Section B shows that the larger model, `Llama-3.2-3B-Instruct`, outperforms `Llama-3.2-1B-Instruct` in 7 out of 14 languages based on IoU scores and 13 out of 14 languages according to S.Corr scores. We limit our study to these lightweight models due to the significant memory overhead required by the TransformerLens[3] library to track and store internal attention weights during inference with longer texts.

4.3 Impact of Including Web Search Results

We conducted experiments to evaluate the impact of enabling versus disabling the search component within our pipeline on hallucination span detection performance. As shown in Table 5 in Appendix Section B, the integration of web search results into the LLM prompt yielded a marginal improvement in performance, with improved performance observed in 12 out of 14 languages as measured by Intersection over Union (IoU) and Spearman correlation. These findings suggest that the inclusion of search results enhances performance detection.

5 Conclusion

In this study, we tackles the critical challenge of hallucination phenomenon observed in LLMs. By

[2]https://helsinki-nlp.github.io/shroom/iou_rankings

[3]https://transformerlensorg.github.io/TransformerLens/

Language	Performance Metrics				
	IoU (Ours)	S.Corr (Ours)	Rank	SOTA Team	Baseline Neural
Arabic	0.604 (1B)	0.605 (1B)	4/32	NotMSA (0.670)	0.042
Basque	0.522 (3B)	0.516 (3B)	8/26	NotMSA (0.613)	0.021
Catalan	0.530 (1B)	0.557 (1B)	9/24	UCSC (0.721)	0.052
Chinese	0.460 (3B)	0.299 (3B)	13/29	YNU-HPCC (0.554)	0.024
Czech	0.443 (1B)	0.481 (1B)	7/26	AILSNTUA (0.543)	0.096
English	0.523 (1B)	0.561 (1B)	10/44	iai_MSU (0.651)	0.031
Farsi	0.575 (1B)	0.519 (1B)	9/26	AILSNTUA (0.711)	0.000
Finnish	0.631 (1B)	0.636 (1B)	4/30	UCSC (0.648)	0.004
French	**0.647 (3B)**	0.619 (3B)	**1/33**	**Deloitte (0.647)**	0.002
German	0.566 (3B)	0.549 (3B)	6/31	UCSC (0.624)	0.032
Hindi	0.632 (3B)	0.639 (3B)	10/27	ccnu (0.747)	0.003
Italian	0.706 (1B)	0.614 (1B)	8/31	UCSC (0.787)	0.010
Spanish	0.407 (3B)	0.585 (3B)	10/35	ATLANTIS (0.531)	0.072
Swedish	0.622 (3B)	0.537 (3B)	3/30	UCSC (0.642)	0.031

Table 2: Uncertainty-based and consistency-based results for languages and teams in the Mu-SHROOM shared task challenge. Values show IoU scores (with Llama-3.2-1B/3B), Spearman correlation, ranking position (out of total participants for that language), SOTA team with their IoU score in parentheses, and the neural baseline performance.

employing a neural network classifier that utilizes features extracted from various layers of an LLM, we enable precise identification of hallucination spans within generated text. Our model achieved a top ten ranking across 13 languages achieving first place specifically in French.

6 Limitations and Future Work

Our methodology necessitates direct access to the internal states of Large Language Models (LLMs), which restricts its deployment to systems that facilitate such access. Utilizing a limited dataset of approximately 507 labeled training examples, we implemented a rudimentary linear classifier for hallucination prediction. By treating each token independently, we potentially lose important signals for hallucination detection. Although our search-based verification process enhances performance metrics, it introduces increased latency and computational demands.

As part of future work, we plan to investigate advanced sequence modeling architectures capable of leveraging the complete sequential relationships inherent in LLM internal states, thereby integrating both layer-wise and token-wise dependencies. Additional features, such as internal gradients and other attention patterns, may yield more informative signals for detection purposes. With more training data, these architectures could better capture the complex dynamics of hallucination generation. With an expanded dataset, these architectures could potentially capture the intricate dynamics associated with hallucination generation more effectively. Furthermore, assessing the efficacy of our method in identifying intrinsic hallucinations constitutes a significant avenue for continued research.

References

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Alex Chandler, Devesh Surve, and Hui Su. 2024. Detecting errors through ensembling prompts (deep): An end-to-end llm framework for detecting factual errors. *Preprint*, arXiv:2406.13009.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yue Fan, Hu Zhang, Ru Li, YuJie Wang, Hongye Tan, and Jiye Liang. 2024. FRVA: Fact-retrieval and verification augmented entailment tree generation for explainable question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9111–9128, Bangkok, Thailand. Association for Computational Linguistics.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024a. Knowledge-centric hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6953–6975, Miami, Florida, USA. Association for Computational Linguistics.

Xiaomeng Hu, Yiming Zhang, Ru Peng, Haozhe Zhang, Chenwei Wu, Gang Chen, and Junbo Zhao. 2024b. Embedding and gradient say wrong: A white-box method for hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1959, Miami, Florida, USA. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023b. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Preprint*, arXiv:2302.09664.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

Haonan Li, Xudong Han, Hao Wang, Yuxia Wang, Minghan Wang, Rui Xing, Yilin Geng, Zenan Zhai, Preslav Nakov, and Timothy Baldwin. 2024a. Loki: An open-source tool for fact verification. *Preprint*, arXiv:2410.01794.

Qing Li, Jiahui Geng, Chenyang Lyu, Derui Zhu, Maxim Panov, and Fakhri Karray. 2024b. Reference-free hallucination detection for large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4542–4551, Miami, Florida, USA. 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. *Preprint*, arXiv:2104.08704.

Shixuan Ma and Quan Wang. 2024. Zero-shot detection of llm-generated text using token cohesiveness. *Preprint*, arXiv:2409.16914.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *Preprint*, arXiv:2303.08896.

James Manyika and Sissie Hsiao. 2023. An overview of bard: an early experiment with generative ai. *AI. Google Static Documents*, 2.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *Preprint*, arXiv:2301.11305.

Seyed Mahed Mousavi, Simone Caldarella, and Giuseppe Riccardi. 2023. Response generation in longitudinal dialogues: Which knowledge representation helps? *Preprint*, arXiv:2305.15908.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 2024-07-18.

Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. Varifocal question generation for fact-checking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2532–2544, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hithesh Sankararaman, Mohammed Nasheed Yasin, Tanner Sorensen, Alessandro Di Bari, and Andreas Stolcke. 2024. Provenance: A light-weight fact-checker for retrieval augmented llm generation output. *Preprint*, arXiv:2411.01022.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *Preprint*, arXiv:2410.11414.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. *Preprint*, arXiv:2404.10774.

Jacob-Junqi Tian, Hao Yu, Yury Orlovskiy, Tyler Vergho, Mauricio Rivera, Mayank Goel, Zachary Yang, Jean-Francois Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Web retrieval agents for evidence-based misinformation detection. *Preprint*, arXiv:2409.00009.

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.

Kim Trong Vu, Michael Krumdick, Varshini Reddy, Franck Dernoncourt, and Viet Dac Lai. 2024. An analysis of multilingual factscore. *Preprint*, arXiv:2406.19415.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *Preprint*, arXiv:2305.16739.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023a. Enhancing uncertainty-based hallucination detection with stronger focus. *Preprint*, arXiv:2311.13230.

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. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Ilana Zimmerman, Jadin Tredup, Ethan Selfridge, and Joseph Bradley. 2024. Two-tiered encoder-based hallucination detection for retrieval-augmented generation in the wild. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 8–22, Miami, Florida, US. Association for Computational Linguistics.

A Appendix:A

Claim Decomposition Prompt

Role:

Your task is to decompose the following output into standalone, decontextualized sentences while retaining all original information. Each sentence should be individually verifiable, free from implied connections or dependencies on other sentences. Avoid introducing information not explicitly stated. If the output cannot be meaningfully decomposed, return it unchanged.

Guidelines:
1. Each decomposed sentence must be standalone, without relying on other sentences for context or meaning.
2. Avoid making assumptions or inferring connections not explicitly stated in the output.
- Ensure that all information from the original output is preserved and split into its most granular, decontextualized form.

Query Generation Prompt

Objective:
Your task is to generate a query for each fact provided. Each query must be concise, specific, and designed to retrieve or verify the exact information presented in the fact. Use the format provided in the example, separating each query with a new line and a dash.

Guidelines:
1. Each query must be standalone, without relying on other facts for context or meaning.
2. Avoid introducing additional information

or rephrasing the fact unnecessarily.
3. Ensure each query is precise enough to verify the specific fact it corresponds to.

B Appendix:B

Feature Type	Dimension	# Blocks	Total Features
hook_attn_out	3,072	28	86,016
hook_resid_post	3,072	28	86,016
ln1.hook_scale*	1	28	28
ln2.hook_scale*	1	28	28
ln_final.hook_scale*	1	1	1
		Total Features:	**172,089**

Table 3: Feature set composition for `Llama-3.2-3B-Instruct` classifier. *Including final layer normalization.

Language	Llama-3.2-1B-Instruct		Llama-3.2-3B-Instruct	
	IoU	S.Corr.	IoU	S.Corr.
Arabic	**0.60**	0.60	0.59	**0.64**
Basque	0.47	0.50	**0.52**	**0.52**
Catalan	**0.53**	0.56	0.50	**0.62**
Chinese	0.45	**0.32**	**0.46**	0.30
Czech	**0.44**	0.48	0.37	**0.50**
English	**0.52**	0.56	0.51	**0.58**
Farsi	**0.58**	0.52	0.51	**0.54**
Finnish	**0.63**	0.64	0.63	**0.64**
French	0.57	0.60	**0.65**	**0.62**
German	0.55	0.53	**0.57**	**0.55**
Hindi	0.61	0.62	**0.63**	**0.64**
Italian	**0.71**	0.61	0.63	**0.65**
Spanish	0.40	0.56	**0.41**	**0.59**
Swedish	0.61	0.51	**0.62**	**0.54**

Table 4: Hallucination Span detection performance by language over the test dataset for `Llama-3.2-1B-Instruct` and `Llama-3.2-3B-Instruct`. Values are rounded to two decimals. Bold values indicate the best performance for IoU and Spearman correlation for each language.

Language	With Search		Without Search	
	IoU	S.Corr.	IoU	S.Corr.
Arabic	**0.59**	**0.64**	0.55	0.58
Basque	**0.52**	**0.52**	0.50	0.51
Catalan	**0.50**	**0.62**	0.48	0.55
Chinese	0.46	0.30	**0.49**	**0.51**
Czech	0.37	**0.50**	**0.41**	0.46
English	**0.51**	**0.58**	0.50	0.54
Farsi	**0.51**	**0.54**	0.49	0.50
Finnish	**0.63**	**0.64**	0.61	0.62
French	**0.65**	**0.62**	0.61	0.57
German	**0.57**	**0.55**	0.51	0.50
Hindi	**0.63**	0.64	0.62	**0.65**
Italian	**0.63**	**0.65**	0.61	0.62
Spanish	**0.41**	**0.59**	0.35	0.53
Swedish	**0.62**	**0.54**	0.51	0.53

Table 5: Hallucination Span detection performance by language over the test dataset with and without search for `Llama-3.2-3B-Instruct`. Values are rounded to two decimals. Bold values indicate the best performance for IoU and Spearman correlation for each language.