AILS-NTUA at SemEval-2025 Task 3: Leveraging Large Language Models and Translation Strategies for Multilingual Hallucination Detection

Dimitra Karkani, Maria Lymperaiou, Giorgos Filandrianos, Nikolaos Spanos, Athanasios Voulodimos, Giorgos Stamou
School of Electrical and Computer Engineering, AILS Laboratory
National Technical University of Athens
dkarkanh@gmail.com, {marialymp, geofila, nspanos}@ails.ece.ntua.gr,
thanosv@mail.ntua.gr, gstam@cs.ntua.gr

Abstract

Multilingual hallucination detection stands as an underexplored challenge, which the Mu-SHROOM shared task seeks to address. In this work, we propose an efficient, training-free LLM prompting strategy that enhances detection by translating multilingual text spans into English. Our approach achieves competitive rankings across multiple languages, securing two first positions in low-resource languages. The consistency of our results highlights the effectiveness of our translation strategy for hallucination detection, demonstrating its applicability regardless of the source language.

1 Introduction

Hallucinations in Large Language Models (LLMs) pose a significant challenge, as they can generate fluent yet factually incorrect or misleading content (Zhang et al., 2023; Huang et al., 2025). Several detection techniques have been developed, either by accessing the LLM's internals (Azaria and Mitchell, 2023; Sriramanan et al., 2024; Chen et al., 2024), probing generation inconsistencies (Manakul et al., 2023a; Mündler et al., 2023) or exploiting token-based classification (Quevedo et al., 2024).

While hallucination detection has been widely studied in monolingual contexts, multilingual settings are severely underexplored and accompanied by additional complexities. Variations in linguistic structure, resource availability, and training data distribution can lead to uneven model reliability across languages. Low-resource languages, in particular, are more susceptible to hallucinations due to limited high-quality training data, making factual consistency a critical issue. Recent works in multilingual hallucinations detection verify faithfulness shortcomings (Qiu et al., 2023a; Shen et al., 2024), shift the focus on data rather than model capacity (Guerreiro et al., 2023a) while also pinpointing evaluation concerns (Kang et al., 2024).

To this end, the Mu-SHROOM shared task is proposed in order to fill this gap by providing high-quality data on 14 languages, including low-resource languages such as Farsi, Czech, Finnish, Swedish and Basque. This effort is accompanied by annotations on hallucinated data spans within given sentences, which participants have to automatically detect.

In our work, we leverage LLM prompting and translation strategies to address hallucination detection without designing independent systems per language. Particularly, we combine two LLMs, Llama 3 (Grattafiori et al., 2024) and Claude (AI), prompting them in a few-shot manner to detect hallucinatory spans. Our language-adaptive system is proven efficient and successful in both high- and low-resource languages *without any training or fine-tuning*. As a result, we achieve first position in the low-resource Farsi and Czech languages, second position in the high-resource Italian language and among the top 15% positions in English, Spanish, German, Hindi and Basque.

2 Background

2.1 Task description

Mu-SHROOM (Vázquez et al., 2025) is a multilingual extension of SHROOM (Mickus et al., 2024) comprising 14 languages: Arabic (Modern standard)-*AR*, Basque-*EU*, Catalan-*CA*, Chinese (Mandarin)-*ZH*, Czech-*CS*, English-*EN*, Farsi-*FA*, Finnish-*FI*, French-*FR*, German-*DE*, Hindi-*HI*, Italian-*IT*, Spanish-*ES*, and Swedish-*SV*. Participants are tasked to detect hallucinatory spans within generated text as accurately as possible, so that if the span was omitted the hallucination would be removed.

2.2 Related work

Multilingual NLP Hallucinations. While hallucination detection has been actively researched in

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1289–1305

monolingual contexts across various fields (Dhuliawala et al., 2023; Manakul et al., 2023b; Min et al., 2023; Fabbri et al., 2022; Maynez et al., 2020; Scialom et al., 2021), its multi-lingual counterpart has primarily focused on identifying hallucinations in machine translation, where such hallucinations are defined as translations containing information completely unrelated to the input (Guerreiro et al., 2023b). Machine translation also has established benchmarks (Dale et al., 2023), as well as metrics (Kang et al., 2024), for evaluating model performance in cross-lingual generation and transfer. Multilingual hallucinations are also extensively studied in text summarization applications. In low-resource languages, summaries are often translated into high-resource languages, such as English, to utilize more reliable evaluation metrics (Qiu et al., 2023b). Mu-SHROOM (Vázquez et al., 2025), based on last year's SHROOM challenge (Mickus et al., 2024), is the inaugural benchmark for multilingual hallucination detection, addressing a significant gap in low-resource language research due to the absence of established benchmarks.

3 System Overview

We focus on LLM prompting and translation strategies to tackle hallucination detection challenges in a language-agnostic manner. Due to the prompt-heavy nature of our approach, no further training is required to attain high scores in either high- or low-resource languages. Specifically, we combine two LLMs, Llama 3.1 405B[1] and Claude 3.5 Sonnet[2], prompting them in a few-shot (FS) way to detect hallucination spans. To improve detection performance in low-resource languages, we also experiment with incorporating a translation tool to translate original input-output data to English. Finally, given the inputs of each MuSHROOM instance, we instruct Llama and Claude to generate the corresponding output and incorporate it as a *hypothesis* to facilitate hallucination detection.

Based on the results of our preliminary experiments presented in App.C, our final system consists of three components, i.e. three experiments:

Component 1 We prompt Claude to detect hallucination spans given the input text, the output text in the original language and their translations in English, as well as the outputs produced by Llama as hypothesis.

[1]meta.llama3-1-405b-instruct-v1:0

[2]anthropic.claude-3-5-sonnet-20241022-v2:0

Component 2 We prompt Llama to detect hallucination spans given the input text, the output text in the original language and their translations in English, as well as the outputs produced by Claude as hypothesis.

Component 3 We prompt Llama to detect hallucination spans given the input text, the output text in the original language and their translations in English *without* providing extra generated answers as hypothesis. We adopt this approach since generated hypotheses can themselves contain hallucinations, resulting in misleading outcomes. Moreover, the LLMs sometimes place undue emphasis on the provided generated hypothesis rather than relying on their internal knowledge, causing them to miss hallucinatory spans in the outputs.

Each component produces a list of hallucination spans and then the three lists are combined as follows: for each produced span, the assigned probability is calculated as the ratio of the experiments that characterize it as hallucination over the total number of experiments (three).

4 Methods

Hallucination categorization To initiate the hallucination detection process, we begin by defining what constitutes a hallucination. This initial categorization allows us to effectively handle data from various tasks and domains. Drawing on Huang et al. (2025) and our exploratory analysis of both the task's sample and validation data, we identify four distinct types of hallucinations:

(1) **Input-Output inconsistency**: The produced output is inconsistent with the input, i.e. it does not satisfy the input query or is irrelevant to it.

(2) **Factual inconsistency**: The output contains information that is factually inconsistent in a sense that it cannot be associated with verifiable real-world facts.

(3) **Internal output inconsistency**: The output contains contradictory facts, i.e.c—c- in the generated text span there is inconsistent information.

(4) **Misspellings**: The output contains misspelled words.

Output Format After defining hallucinations, we specify the expected output format to effectively process LLM outputs. For that purpose we attempt two approaches: In our first approach, in order to make the procedure simpler, we split the output text in parts and prompt the LLMs given the input, the output and a specific part of the output to

decide whether that specific part contains a hallucination. This technique is not successful as shown in Tables 1, 2 in the preliminary column. In the second approach, which we ultimately adopt, we prompt the LLM with both the input and output to detect hallucination spans while also encouraging chain-of-thought (CoT) reasoning.

Prompting strategies After initializing the process of hallucination detection by providing the hallucination definition and the expected output format, we experiment in both a **zero-shot (ZS)** and a **few-shot(FS)** way. In the FS scenario we present an example for each aforementioned category, together with the expected output format. For the main process, we adopt the system/user prompt. After consolidating the system prompts, we experiment with the user prompts and specifically the data given as input to the LLMs. For the simplest approach, we just provide the input-output pair to the LLM. Then, we attempt to enhance performance by also demonstrating the *translations* of inputs and outputs. To deploy our final system, as explained above, we also supply a *hypothesis* generated by the LLMs. To show how each of the steps of the procedure we propose improves the final results we present the following experiments:

1. Standalone prompting We prompt Claude or Llama at a time in either a ZS or a FS manner without translating in English or leveraging hypotheses.

2. Prompting + Translation We prompt Llama and Claude to detect hallucination spans using the input and output texts in the original language, as well as their English translations, without providing additional generated hypotheses; then, we combine the two lists of produced hallucination spans.

3. Prompting + Translation + Hypothesis We prompt Claude to detect hallucination spans using the input and output text in the original language, their English translations, and the outputs produced by Llama as hypotheses. Conversely, we prompt Llama with the same inputs but use Claude's outputs as hypotheses. Additionally, we incorporate the results from Llama's previous experiment, combining the three produced hallucination span lists.

Translation Given the disparity in linguistic resources across different languages and our objective of developing a system that performs robustly across multilingual settings, we investigate various strategies to address this challenge. Specifically, we examine the following key questions in the context of multilingual hallucination detection: *"Is it more effective to provide both the input and output in their original languages and allow the LLM to detect hallucinations, or should we instead supply their English translations?"* Furthermore, *"If input-output pairs are provided in their original language, should the prompt also be in the same language, or is it preferable to present it in English?"*. Conversely, *"If we supplement the detection process with translated input-output pairs, is it more effective for the LLM itself to handle the translation before the detection step, or should an external translation system be employed?"*

To explore these questions, we conduct the following experiments. Firstly, we experimented with the impact of the language of the prompt given the input-output pairs in their original languages. In this direction, the following ***Original Input-Output Pairs*** experiments are conducted.

No translator The simplest approach is to provide the definition of hallucination, along with the examples per category and instructions for the output format in English, and then present the input-output pairs in their original language.

External translator - original language Given the input-output pairs in their original language, we translate the prompts—which include the hallucination definition and output format instructions—into the original language. To achieve this, we use the Google Translate API for Python[3].

For the second part of our experiments, we translate the input-output pairs into English, the highest-resource language, and then use the translated pairs to detect hallucinations, while the prompt remains in English. The LLM is exposed to both the original language (before translation), as well as with its English version to ensure fairness. The experiments we conduct belong to the ***Translated Input-Output Pairs*** category.

External translator - English We translate the input-output pairs into English using Google Translate and then prompt the LLMs (providing *both* the original and English versions of data) to generate a CoT for hallucination detection in English.

LLM as the translator We prompt the LLMs to perform the analysis in two steps: Firstly to translate input-output pairs into English when the text is in another language, and then based on that to detect hallucinations in the same chat, thus ensuring that the LLM is exposed in both languages before concluding to the hallucination spans identified.

[3] https://pypi.org/project/googletrans/

Language (id)	Baseline	Preliminary	ZS	FS	FS + Translation	FS + Translation + Hypothesis
Arabic (ar)	0.04/0.36/0.05	0.223	0.379	0.425	0.527	**0.584**
Catalan (ca)	0.05/0.24/0.08	0.273	0.482	0.540	0.675	**0.703**
Czech (cs)	0.10/0.26/0.13	0.301	0.388	0.448	0.556	**0.587**
German (de)	0.03/0.35/0.03	0.199	0.531	0.564	0.578	**0.587**
English (en)	0.03/0.35/0.03	0.223	0.425	0.487	-	**0.555**
Spanish (es)	0.07/0.19/0.09	0.239	0.385	0.454	0.468	**0.500**
Basque (eu)	0.02/0.37/0.01	0.299	0.431	0.458	0.518	**0.571**
Farsi (fa)	0.00/0.20/0.00	0.202	0.492	0.558	0.687	**0.753**
Finnish (fi)	0.01/0.49/0.00	0.210	0.464	0.529	0.635	**0.683**
French (fr)	0.00/0.45/0.00	0.251	0.447	0.499	0.535	**0.617**
Hindi (hi)	0.00/0.27/0.00	0.189	0.581	0.624	0.709	**0.726**
Italian (it)	0.01/0.28/0.00	0.267	0.597	0.657	0.774	**0.802**
Swedish (sv)	0.03/0.53/0.02	0.276	0.492	0.537	0.585	**0.601**
Chinese (zh)	0.02/0.47/0.02	0.200	0.212	0.304	0.378	**0.419**

Table 1: Prompting scenarios comparison – IoU metric. The three baselines are: neural/ mark-all/ mark-none. The best-performing method per language is in **bold**. This Table considers the best translation strategy.

Language (id)	Baseline	Preliminary	ZS	FS	FS + Translation	FS + Translation+hypothesis
Arabic (ar)	0.11/0.01/0.01	0.190	0.484	0.636	0.601	**0.612**
Catalan (ca)	0.06/0.06/0.06	0.397	0.610	0.564	0.700	**0.709**
Czech (cs)	0.05/0.10/0.10	0.368	0.480	0.419	0.557	**0.590**
German (de)	0.11/0.01/0.01	0.333	0.583	0.466	0.614	**0.629**
English (en)	0.11/0.00/0.00	0.357	0.511	**0.635**	-	0.628
Spanish (es)	0.04/0.01/0.01	0.456	0.464	0.547	0.537	**0.565**
Basque (eu)	0.10/0.00/0.00	0.401	0.530	0.555	0.524	**0.566**
Farsi (fa)	0.11/0.01/0.01	0.378	0.583	0.547	0.684	**0.737**
Finnish (fi)	0.09/0.00/0.00	0.478	0.552	0.584	**0.666**	0.652
French (fr)	0.02/0.00/0.00	0.254	0.564	**0.617**	0.609	0.614
Hindi (hi)	0.14/0.00/0.00	0.565	0.666	0.676	0.754	**0.760**
Italian (it)	0.08/0.00/0.00	0.526	0.692	0.765	0.757	**0.817**
Swedish (sv)	0.10/0.01/0.01	0.194	0.502	0.525	0.535	**0.562**
Chinese (zh)	0.08/0.00/0.00	0.264	0.317	0.401	**0.487**	0.464

Table 2: Prompting scenario comparison – Correlation. The three baselines are: neural/ mark-all/ mark-none.The best-performing method per language is in **bold**. This Table considers the best translation strategy.

5 Experimental setup

Dataset For the results presented, the test set provided by the task organizers is used. The test set is presented in detail in Appendix A.

Baselines presented by the organizers comprise a neural-based model, and the edge cases of mark-all and a mark-none.

Evaluation comprises two character-level metrics: first, Intersection-over-Union **(IoU)** of characters marked as hallucinations in the gold reference vs. characters predicted as such; second, the **correlation** between the hallucination probabilities occurring from the detection system and the gold reference probabilities provided by the annotators.

Computational resources All our experiments are executed in Amazon Bedrock using Google Colab platform for the API calls.

6 Results

The results of our experiments are shown in detail in Tables 1, 2 for prompting experiments (IoU and correlation metrics respectively) and in Table 3 regarding translation experiments. In the prompting experiments, the FS approach for both hallucination definition and expected output format significantly improves the results: The detected hallucination spans are more accurate and the output format is strictly followed, which is fundamental in order to automatically handle the answers and extract the feedback provided by the LLMs. Furthermore, incorporating the English translation of the texts appears to enhance the LLM's performance. A similar effect is observed when integrating the hypothesis from the other LLM. In this case, the LLM is able to compare the hypothesis with the actual output provided to determine more effectively the presence of hallucinations and identify their respective spans. Notably, these patterns remain consistent across all languages, *regardless of whether they are low-resource or high-resource*. However, the addition of translations and hypotheses has *a more pronounced impact on low-resource* languages compared to high-resource ones. Regarding the translation experiments, interesting insights emerge. On one hand, in the simplest approach—where prompts are given in English while pairs remain in their original language—the English language score is not the highest. This suggests that translation aids in hallucination detec-

Language (id)	Original Input-Output Pairs		Translated Input-Output Pairs	
	No Translation	External transl. - Original	LLM Translator	External Transl. English
Arabic (ar)	0.47/0.55	0.32/0.40	0.61/0.51	**0.58/0.61**
Catalan (ca)	0.46/0.58	0.50/0.62	0.49/0.59	**0.70/0.71**
Czech (cs)	0.39/0.42	0.37/0.43	0.42/0.43	**0.59/0.59**
German (de)	0.50/0.51	0.47/0.56	0.48/0.54	**0.59/0.63**
English (en)	**0.55/0.63**	-	-	-
Spanish (es)	0.49/0.49	0.31/0.477	0.42/0.480	**0.50/0.56**
Basque (eu)	0.35/0.46	0.34/0.44	0.37/0.49	**0.57/0.57**
Farsi (fa)	0.50/0.61	0.49/0.58	0.52/0.63	**0.75/0.74**
Finnish (fi)	0.54/0.57	0.54/0.56	0.53/0.58	**0.68/0.65**
French (fr)	0.49/0.530	0.43/0.450	0.45/0.46	**0.617/0.614**
Hindi (hi)	0.65/0.67	0.66/0.68	0.70/0.710	**0.73/0.760**
Italian (it)	0.62/0.620	0.604/0.679	0.730/0.680	**0.802/0.817**
Swedish (sv)	0.53/0.550	0.555/0.567	0.570/0.540	**0.601/0.562**
Chinese (zh)	0.343/0.399	0.399/0.388	0.379/0.333	**0.419/0.464**

Table 3: Translation performance comparison - IoU/Correlation metrics respectively. The best-performing method per language is in **bold**. The best translation strategy is used in the results presented in Tables 1, 2.

id	Sentence
ca	El municipi de Yushu es troba a 4.500 metres sobre el nivell del mar.
cs	Řeka Labe (německy Elbe) pramení v Českém lese, konkrétně v okrese Jičín, v nadmořské výšce 816 metrů. Pramení v údolí mezi vrcholy Kozákov (744 metrů) a Říp (459 metrů).
de	Mario Bola ti wechselt im Jahr 1998 zum Verein AC Mailand.
en	Mouthier is located in the department of Haute-Loire.
eu	Hiru espezie bakarrik daude.
fi	Folorunsho Alakija on nigerialainen kirjailija ja aktivisti. Hän on kirjoittanut useita kirjoja, muun muassa "The Slave Girl" ja "The Slave Girl's Daughter", jotka käsittelevät naisten sortoa ja orjuutta.
fr	L'espèce Pseudomugil gertrudae appartient à la famille des Poeciliidae, qui est une famille d'espèces de poissons d'eau douce et d'eau salée. Elle est également connue sous le nom de poisson-chat de Gertrude ou de poisson-chat de Gertrude.
it	Il produttore dell'album "Plastic Letters" di Blondie fu Mike Chapman.
sv	David Sandbergs födelseort är New York.
zh	新缬草原产于欧洲，特别是地中海沿岸地区，包括西班牙、葡萄牙、法国南部、意大利和希腊等地。它在这些地区的自然环境中广泛分布，并且在园艺上也被引种到其他地区。由于其美丽的花朵和耐旱的特性，新 Valerie 在全球各地都有一定的栽培和观赏价值

Table 4: Qualitative results in various languages. The detected hallucination span is highlighted.

tion and partially addresses the challenges of low-resource languages. However, the reverse process, translating into the language of the pairs, does not appear to offer the same benefits. Additionally, the use of the Google Translator is more effective compared to the end-to-end system where the LLMs are prompted to translate the input and output texts themselves. Thus, the most effective approach for identifying multilingual hallucinations is to provide *both the prompt and the input-output pairs in English*, using an *external translation system* rather than incorporating translation as a step within the LLM pipeline. This finding holds consistently across all languages in the dataset.

The results tables also show that for high-resource languages such as Spanish, Chinese, and German, the FS scenario and the incorporation of the generated hypothesis contribute the most towards performance improvements. In contrast, *for low-resource languages, translation is a crucial component* in achieving similar results. The prompts are detailed in App. D.

As a demonstration of our system, in Table 4 we present some qualitative results from our system which achieve a perfect score of IoU=1.

7 Conclusion

In this work, we detect multilingual hallucination spans in the outputs from the SemEval 2025-Task 3 MuSHROOM dataset using LLM prompting and translation techniques. We showcase the merits of converting all data in English, especially for low-resource languages, achieving highly-ranked results in a language-agnostic manner overall.

References

Anthropic AI. The claude 3 model family: Opus, sonnet, haiku.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. *ArXiv*, abs/2402.03744.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loïc Barrault, and Marta R. Costa-jussà. 2023. Halomi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. *Preprint*, arXiv:2305.11746.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *Preprint*, arXiv:2309.11495.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, 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, Danny Wyatt, 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, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, 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, Jack Zhang, 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, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, 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, Maria Tsimpoukelli, 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, Ning Zhang, 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, Rohan Maheswari, 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, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, 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, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, 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, Ce Liu, Changhan Wang,

Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, 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, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, 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, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, 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, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, 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, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, 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, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, 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, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, 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, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Natalia Grigoriadou, Maria Lymperaiou, George Filandrianos, and Giorgos Stamou. 2024. AILS-NTUA at SemEval-2024 task 6: Efficient model tuning for hallucination detection and analysis. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1549–1560, Mexico City, Mexico. Association for Computational Linguistics.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André Martins. 2023a. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023b. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. 2024. Comparing hallucination detection metrics for multilingual generation. *Preprint*, arXiv:2402.10496.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023a. 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.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023b. Selfcheckgpt: Zero-resource black-box hal-

lucination detection for generative large language models. *Preprint*, arXiv:2303.08896.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

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.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *Preprint*, arXiv:2305.14251.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin T. Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *ArXiv*, abs/2305.15852.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo Ponti, and Shay Cohen. 2023a. Detecting and mitigating hallucinations in multilingual summarisation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8914–8932, Singapore. Association for Computational Linguistics.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo M. Ponti, and Shay B. Cohen. 2023b. Detecting and mitigating hallucinations in multilingual summarisation. *Preprint*, arXiv:2305.13632.

Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomás Cerný. 2024. Detecting hallucinations in large language model generation: A token probability approach. *ArXiv*, abs/2405.19648.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiaming Shen, Tianqi Liu, Jialu Liu, Zhen Qin, Jay Pavagadhi, Simon Baumgartner, and Michael Bendersky. 2024. Multilingual fine-grained news headline hallucination detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7862–7875, Miami, Florida, USA. Association for Computational Linguistics.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. LLM-check: Investigating detection of hallucinations in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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.

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. *Preprint*, arXiv:2309.01219.

A Exploratory data analysis

1.1 Sample set

In the preliminary phase of the competition, a small sample set is released to allow participants to acclimate to the task. The sample set consists of a total of 8 samples in 3 different languages (3 samples in English, 3 samples in Spanish and 2 samples in French). The features of each samples are:

- 'id': a unique number of the sample.
- 'lang': the id of the language of the participating text.
- 'model_input': the input prompt given to the model.
- 'model_output_text: the output text the model generated.
- 'model_id ': the id of the model that produced the output.
- 'soft_labels': spans that include a start and end character number, together with an assigned probability to the span constrained by these two characters.
- 'hard_labels': spans for which the assigned soft label probability is more than 0.5.

Models producing hallucinations to be detected are the following: The different models used are:

- Model id: TheBloke/Mistral-7B-Instruct-v0.2-GGUF (4 annotated samples).

- Model id:meta-llama/Meta-Llama-3-8B-Instruct (1 annotated sample).
- Model id:Iker/Llama-3-Instruct-Neurona-8b-v2 (3 annotated samples).

Since we propose a model-agnostic approach, information regarding the model from which hallucination occurs is overlooked in practice.

1.2 Validation set

The validation set consists of 10 subsets of 50 samples each, separated by language, comprising 500 samples in total. The different languages are: arabic, german, english, spanish, finnish, french, hindi, italian, swedish and chinese. The features of the samples were the same as the ones of the sample set in version 1, with the extension of:

- model_output_tokens: the tokens of the output of the model
- model_output_logits: the logits of the output of the model

in version 2 that was released after the evaluation phase.

Since the data were of different tasks and domains, the exploratory analysis contained an effort to categorize the hallucinations found in the data.

Based on the definition of hallucinations and overgeneration mistakes, we distinguish the following types of hallucinations:

A hallucination is the production of fluent but incorrect output of an LLM. The definition of incorrect output falls in four categories:

1. The *output is inconsistent with the input*, so the produced answer does not answer the input query or is irrelevant to it.

2. The *output contains a factual inconsistency*, so contains something that is not a validated fact or is wrong.

3. The *output contains contradictory facts*, so in the output there are things that cannot be true at the same time.

4. The *output contains misspelled words*.

Based on these categories, in an attempt to understand better the annotation procedure, we conducted manually a categorization of the hallucinations marked in the outputs:

- Input-Output Inconsistency: 10
- Factual Inconsistency: 31
- Output conflicts: 6
- Misspelled words: 8

1.3 Test set

The number of annotated samples per language in the test set are presented in Figure 1. Most languages contain around 150 samples, with the exception of Basque (EU), Farsi (FA), Catalan (CA) and Czech (CS) that contain around 100 samples each.

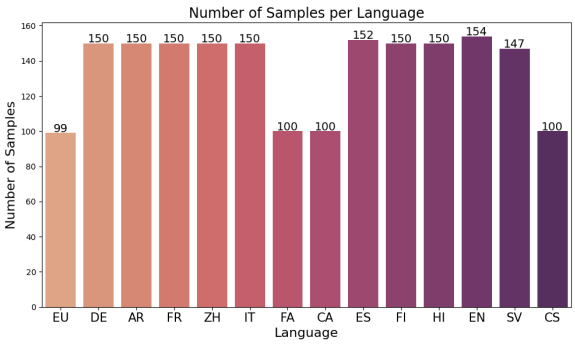

Figure 1: Number of annotated test data samples per language.

Input & Output sentence length In the following Figures, we present the length distribution regarding the input and output sentences per language on the test set.

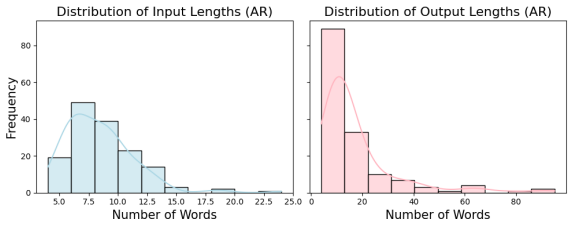

Figure 2: Input/Output length distribution for Arabic.

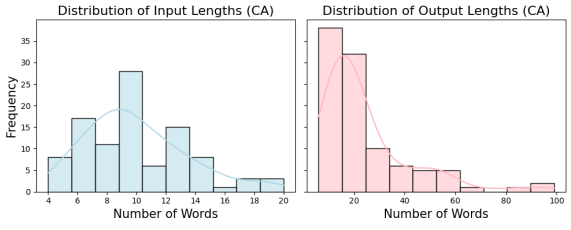

Figure 3: Input/Output length distribution for Catalan.

B Examples

We present some input/output examples from respective LLMs in Table 5.

Lang	Input	Output
EU	Zenbat Pentastomida espezie daude?	Hiru espezie bakarrik daude.
DE	Welche Filme produzierte David Sandberg im Jahr 2014?	David Sandbergs Filme im Jahre 2014 waren "The Amazing Spider-Man 2: Rise of Electro" (2014), "Die Tribute von Panem - Mockingjay Teil 2" ("The Hunger Games: MOCKINGJAY - Part 2", 2014), und "Der Hobbit: Die Schlacht der fünf Heere" "(The Hobit: The Battle of the Five Armies", 2015).
FR	À la suite de quel grand événement a été créee la ville de La Bellière ?	La ville de La Bellière a été créée en 1975 après la fusion de cinq communes voisines : Bellière, Bois-Jérôme, Bretteville-l'Orgueilleuse, Saint-Martin-d'Écublei et Saint-Michel-de-la-Pierre. Le motif de la fusion était la volonté des communes de coopérer pour améliorer leurs infrastructures et offrir un meilleur niveau de services aux résidents. Ce n'y a pas d'événement particulier qui a déclenché la création de la ville, mais plutôt une décision concertée des conseils municipaux de ces cinq communes pour s'unir pour des raisons économiques et administratives.
ZH	大卫·桑德柏格的导演处女作是什么?	大卫桑德的导演处子秀是1993年的电影《性、谎言和录像带》。这部电影获得了戛纳电影节金棕榈奖，并成为当时独立电影的里程碑之一。
IT	Da chi fu coniato il termine "paradosso francese"?	Il termine "paradosso francese" fu coniato dal nutrizionista americano Ancel Keys nel 1952.
CA	A la vora de quin mar es troba la ciutat russa Kaspíisk?	Kaspíisk es una ciutat situada a la costa del mar de Azov, a la província de Krasnodar, a la Rússia.
ES	¿En qué equipos de la NHL jugó Bert Olmstead durante su carrera?	Bert Olmstead jugó para los Montreal Canadiens, Toronto Maple Leafs, New York Rangers y Los Angeles Kings durante su carrera en la NHL.
FI	Millä nimellä David Sandberg tunnetaan YouTubessa?	David Sandberg tunnetaan paremmin nimellä "AtomicDave" YouTubessa. Hän on tunnettu sisällöntuottaja, joka julkaisee usein videoita, joissa hän pelaa pelejä ja tekee haasteita. Hänellä on yli miljoona tilaajaa YouTubessa, ja hän on tunnettu energisestä ja viihdyttävästä tyylistään.
EN	Did Alberto Fouillioux ever play in a world cup championship?	No, Albero Foulois was not in any of the FIFA World Cup finals.
SV	I vilken stad föddes regissören David Sandberg?	David Sandbergs födelseort är New York.
CS	Jaká oblast České republiky je označována slovem sudety?	Sudety je označení pro horské pásmo v České republice a Polsku, které se táhne od severozápadu na jihovýchod. V České republice se nacházejí Sudetské hory, které se dělí na Krkonoše, Jeseníky, Orlické hory, Adršpach-Teplické skály, Rychlebské hory, Hrubý Jeseník, Javoří hory, Zlaté hory a Králický Sněžník. Tyto hory byly důležitým místem během druhé světové války, kdy se zde odehrávaly bitvy mezi německými a československými vojsky.

Table 5: Examples (inputs/LLM outputs) in MuSHROOM languages.

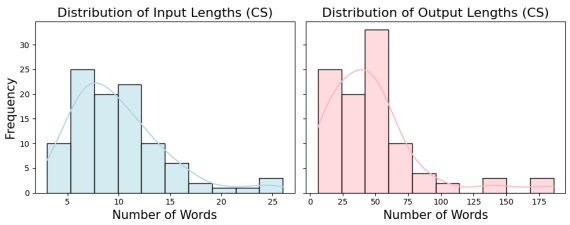

Figure 4: Input/Output length distribution for Czech.

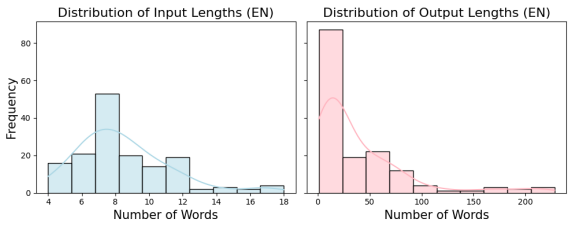

Figure 6: Input/Output length distribution for English.

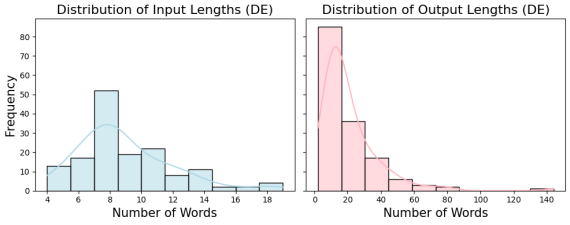

Figure 5: Input/Output length distribution for German.

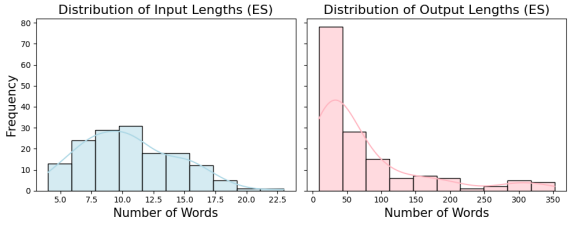

Figure 7: Input/Output length distribution for Spanish.

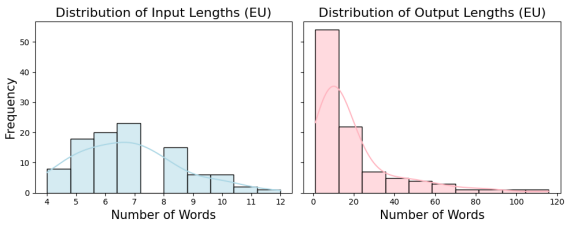

Figure 8: Input/Output length distribution for Basque.

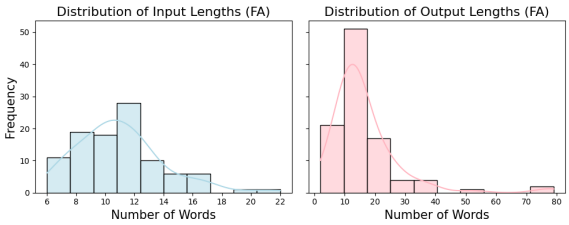

Figure 9: Input/Output length distribution for Farsi.

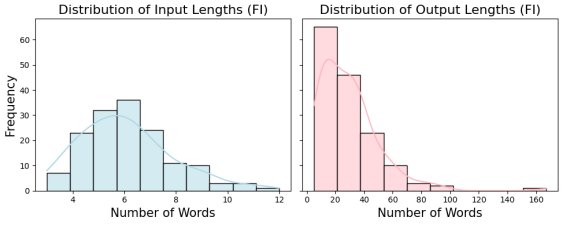

Figure 10: Input/Output length distribution for Finnish.

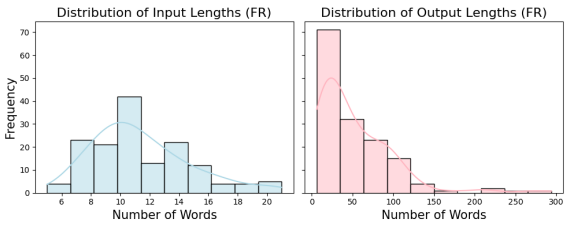

Figure 11: Input/Output length distribution for French.

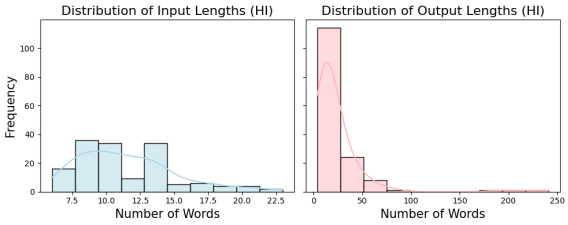

Figure 12: Input/Output length distribution for Hindi.

C Preliminary experiments

In order to deploy our LLM-based system, we conduct some exploratory experiments with Llama and Claude. For that purpose we initially employ the labeled test set from the SHROOM-shared task of 2024 (Mickus et al., 2024), due to its larger

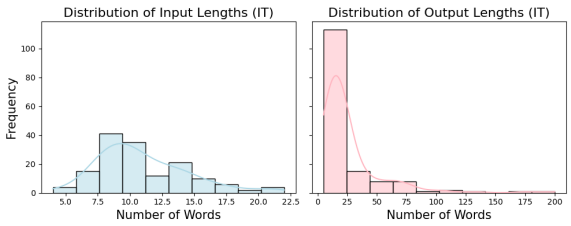

Figure 13: Input/Output length distribution for Italian.

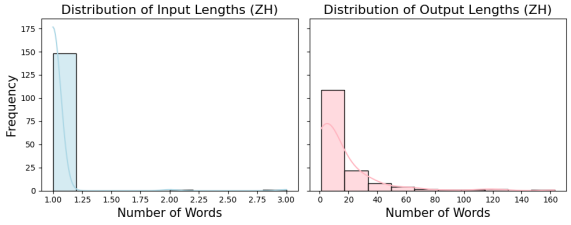

Figure 14: Input/Output length distribution for Chinese.

amount of labeled examples. This dataset contains instances with the following features: *Source - src* is the input given to a model, *hypothesis - hyp* is the output generated by the model, *target - tgt* comprises the ground truth output for this specific model, *reference - ref* indicates whether target, source or both of these fields contain the semantic information necessary to establish whether a datapoint is a hallucination, *task* refers to the task being solved and *model* to the model being used (in the model-agnostic case the model entry remains empty).

In this task, the participants were asked to classify the output of the LLM as hallucination or not based on the meaning of the target output that the LLM should have produced. Each instance also contains the tag *Hallucination/Not Hallucination* and a *probability* expressing the ratio of the annotators that marked the output as Hallucination over all the annotator that participated.

In order to explore the capabilities of Llama and Claude in hallucination detection, we manipulate the SHROOM-2024 dataset by bringing it to the format of this year's dataset. To perform that we create the feature 'model input' by combining the source and the task feature that has three distinct values (MT - Machine Translation, PG - Paraphrase Generation, D - Definition Modeling) as described in the Table 6.

Preliminary experiments involve probing the hallucination detection capabilities of Llama and Claude models. More specifically we conduct the following four experiments:

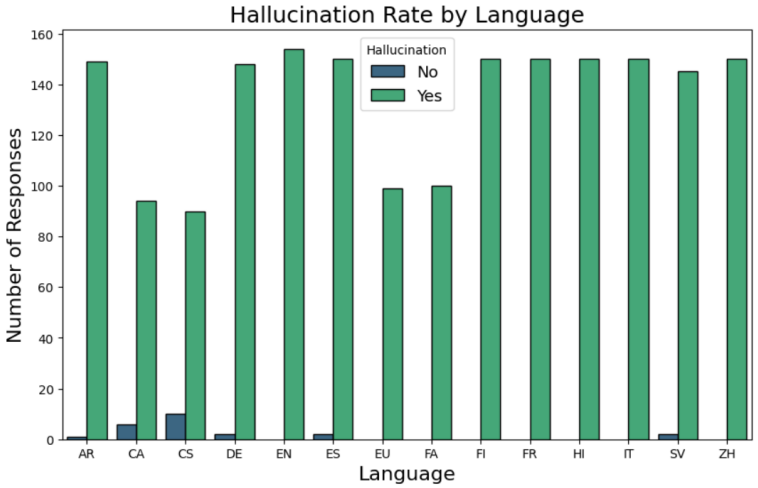

Figure 15: Hallucination rate per language according to 'soft label' annotations.

Task	Source	Model Input
Machine Translation (MT)	src	Translate the following sentence in English : {src}
Definition Modelling (DM)	src	{src}
Paraphrase Generation (PG)	src	Paraphrase the following sentence : {src}

Table 6: Combination of task identifier and source text to create model input feature

1. Zero-shot (ZS) We use a simple zero-shot prompt that questions in the form: *"Given the {input} and the {output}, is the output a hallucination? Reply with Yes/No"*.

2. ZS + Hypothesis To further boost this, we also leverage the *hypothesis* provided, transforming the prompt as: *"Given the input, the output and the hypothesis, is the output a hallucination? Reply with Yes/No"*.

3. CoT + Hypothesis Since the results' accuracy was close to randomly assigning a Hallucination/not Hallucination tag, instead of prompting the LLMs to simply respond with a binary Yes/No label, we prompt them to develop their thoughts; then, based on this and the hypothesis, we generate the final Yes/No label. The prompt used in this case is: *"Given the {input}, {output} and the {hypothesis} is the output a hallucination? Firstly, explain your thought and in the end write the word Yes or No if it is or it is not a hallucination respectively."*

4. Hypothesis generation and similarity Since the Mu-SHROOM dataset does not contain ground truth hypotheses, we prompt the LLMs to generate those by replying to the input query, so that we can assess the hypothesis similarity with the generated output using Natural Language Inference (NLI) models, similar to Grigoriadou et al. (2024). The results are presented in Table 7.

After these experiments, we manually observe the false negatives, and conclude the following:

1. Even though Claude performs better than Llama in general, there are cases where Llama is able to detect some hallucinations that Claude could not.

2. Even though providing a hypothesis improves the results, there are cases that the LLM (either of the two) focuses more on the hypothesis than its internal knowledge and therefore it fails to recognize a hallucinated part.

On the second part of the preliminary experiments, we tried combining different components in order to benefit from the extra information that the combinations provide and then choose the best strategy for the development of our final system. Those experiments were conducted with the validation set in order to gain some insight on the performance of the models with the multilingual texts and are the following:

1. No Hypothesis In the first experiment we prompt Claude and Llama to detect hallucination spans without providing a hypothesis.

2. Model + Hypothesis from the same model In the second experiment we prompt Claude and Llama to generate answers for the input texts, and then prompt them to detect hallucination spans on the output text given the answers each model produced, i.e. prompt Llama given as hypothesis the answers that Llama produced and Claude given as

Model	ZS	ZS + Hypothesis	CoT + Hypothesis	Hypothesis Generation
Llama	0.52	0.55	0.62	0.79
Claude	0.51	0.55	0.63	0.83

Table 7: Results of preliminary experiments: Columns ZS, ZS + Hypothesis, CoT + Hypothesis refer to the accuracy of the classification task (Hallucination/Not Hallucination) and the Hypothesis Generation refer to the similarity of the produced answers with the ground truth provided

Metric	**C**	**L**	**C, CH**	**C, LH**	**L, LH**	**L, CH**
IoU	0.465	0.477	0.478	0.521	0.491	0.543
Cor	0.525	0.433	0.587	0.525	0.522	0.582

Table 8: IoU and Correlation scores for the english: C: Claude, No Hypothesis, L: Llama, No Hypothesis, C,CH: Claude + Claude Hypothesis, C,LH: Claude + Llama Hypothesis , L, LH: Llama +Llama Hypothesis , L,CH: Llama + Claude Hypothesis

	Metrics		Metrics		Metrics
C+L	IoU:0.42 Cor: 0.57	**L + C,LH**	IoU: **0.52 Cor: 0.61**	**L, LH +L, CH**	IoU: 0.41 Cor: 0.57
C+ C,CH	IoU: 0.50 Cor: 0.57	**L + L, LH**	IoU: 0.39 Cor: 0.51	**C, CH+ L,LH**	IoU: 0.46 Cor: 0.61
C + C, LH	**IoU: 0.53 Cor: 0.59**	**L + L, CH**	IoU: 0.51 Cor: 0.62	**C, CH + L, CH**	IoU: 0.46 Cor: 0.55
C + L, LH	IoU: 0.44 Cor: 0.59	**L+ C, CH**	IoU: 0.44 Cor: 0.59	**C, LH + L, LH**	IoU: 0.44 Cor: 0.612
C +L, CH	**IoU: 0.53 Cor: 0.59**	**C, CH +C, LH**	IoU: 0.50 Cor: 0.62	**C, LH +L, CH**	**IoU: 0.54 Cor: 0.64**

Table 9: IoU and Correlation scores for the combination of the results from different experiments: C: Claude, No Hypothesis, L: Llama, No Hypothesis, C,CH: Claude + Claude Hypothesis, C,LH: Claude + Llama Hypothesis , L, LH: Llama +Llama Hypothesis , L,CH: Llama + Claude Hypothesis

hypothesis the answers that Claude produced.

3. Model + Hypothesis from the other Model In the third experiment we prompt Claude and Llama to generate answers for the input texts, and then prompt them to detect hallucination spans on the output text given the answers the other model produced, i.e. prompt Llama given as hypothesis the answers that Claude produced and Claude given as hypothesis the answers that Llama produced.

4. Combinations of two of the above For the fourth experiment we examine how the combinations of the results of two of the components above improve the performance.

To measure these results we used the evaluation metrics of the task (IoU and Cor) but we prioritized the IoU to choose the dominant components.

In Tables 8 , 9 we present the results for each experiment in english in detail and then we present the results for the top-3 strategies for every other language.

For the last experiment, we calculated the IoU of the predictions that occurred from the different experiments and we came to the conclusion that even though the IoU between predictions with and without hypothesis are lower than these of prediction with hypothesis from different sources, the combination of predictions with and without hypothesis reached better scores. After carefully inspecting the results, we found that reasonable because the lack of hypothesis led the LLM to emphasize on more parts that might contain hallucinations rather than the factual inconsistencies that were more often inspected when the hypothesis was provided. An example is provided for English in Table 11.

Language id	**Experiment, IoU**
ar	Claude, Llama Hypothesis: 0.53 Llama, Claude Hypothesis: 0.52 Claude, no Hypothesis: 0.49
de	Llama, Claude Hypothesis: 0.56 Claude, Llama Hypothesis: 0.55 Llama, no Hypothesis: 0.54
es	Llama, Claude Hypothesis: 0.42 Llama, no Hypothesis: 0.41 Claude, Llama Hypothesis: 0.49
fi	Llama, Claude Hypothesis: 0.61 Claude, Llama Hypothesis: 0.59 Llama, no Hypothesis: 0.56
fr	Claude, Claude Hypothesis: 0.58 Claude, Llama Hypothesis: 0.57 Claude, no Hypothesis: 0.56
hi	Llama, Claude Hypothesis: 0.57 Claude, Llama Hypothesis: 0.55 Llama, no Hypothesis: 0.52
it	Llama, Claude Hypothesis: 0.60 Claude, Llama Hypothesis: 0.59 Llama, Llama Hypothesis: 0.59
sv	Claude, Llama Hypothesis: 0.51 Llama, Claude Hypothesis: 0.47 Llama, Llama Hypothesis: 0.44
zh	Claude, Llama Hypothesis: 0.35 Claude, Claude Hypothesis: 0.33 Llama, Llama Hypothesis: 0.30

Table 10: Best Components for each language

Based on these findings showing how the LLMs benefit from the answers provided by the other LLM, we design the prompting experiments presented in the main paper (Section 3 - Prompting

Components	IoU1	IoU2	IoU3	IoU4
Claude, no Hypothesis + Claude, Llama Hypothesis	0.47	0.52	0.465	**0.53**
Claude, Llama Hypothesis + Claude, Claude Hypothesis	0.52	0.48	0.87	0.50

Table 11: For combinations of components, we calculate the IoU of each separate component with the reference values (IoU1, IoU2), the IoU between the predictions of different components (IoU3) and the IoU between the predictions of the combined components and the reference values (IoU4)

strategies).

D Prompts

The prompts to initialize the hallucination detection and the answer generation processes are presented in Tables 12, 13. The system and user prompts designed for our approaches are presented in Tables 14, 15.

Description	Prompt
	Zero-Shot Scenario
Hallucination Definition	You are a hallucination detector. A hallucination is the production of fluent but incorrect output of an LLM.The definition of incorrect output falls in four categories: a. The output is inconsistent with the input, so the produced answer does not answer the input query or is irrelevant to it., b. The output contains a factual inconsistency, so contains something that is not a validated fact or is wrong. c. The output contains contradictory facts so in the output there are things that cannot be true at the same time. d. The output contains mispelled words
Output Format Instruction	I will provide some examples for you to find hallucinations based on the given definition of hallucination. Although there might be several parts where hallucinations of probably different types occur, the answer should only end with the phrase 'So the hallucinations are: ' followed by the hallucinations exactly as they are written in the sentence given, inside "" and separated by commas
	Few-Shot Scenario
One example for each hallucination type	**Example 1(input-output conflict)**:The input is: Where did the Olympic Games of 2004 take place? The output is: The Olympic Games of 2020 took place in London.As a hallucination detector you should point out that there is a hallucination here because the output replies the answer where did the Olympic Games of 2020 take place. So the hallucinations are:"2020". **Example 2 (factual inconsistency)**:The input is: Where did the Olympic Games of 2004 take place? The output is: The Olympic Games of 2004 took place in Florida.As a hallucination detector you should point out that there is a hallucination here because the Olympic Games of 2004 took place in Athens, so there is a factual inconsistency in the word "Florida".So the hallucinations are: "Florida" **Example 3(internal output conflict)**:The input is: Where did the Olympic Games of 2004 take place and what was the biggest Stadium used? The output is: The Olympic Games of 2004 took place in Athens, Greece. All stadiums were designed for that purpose but the biggest was Olympic Stadium of Athens "Spyros Louis" that was built in 1982.As a hallucination detector you should point out that there is a hallucination here because the output states that all stadiums were created for that purpose but the Olympic Stadium of Athens "Spyros Louis" was built in 1982 so much earlier than the Olympic Games, So the hallucinations are: "All stadiums were designed for that purpose".

Table 12: Prompts used to initialize the hallucination detection or the answer generation process.

Description	Prompt
Few-Shot Scenario	
One example for each hallucination type	**Example 4(misspelling)**:The input is: Where did the Olympic Games of 2004 take place? The output is: The OLympooooc Games of 2004 took place in Athens, Greece. As a hallucination detector you should point out that there is a hallucination here because the OLympooooooc Games are a misspelling of the Olympic Games So the hallucinations are: "OLympooooc".
One example for the expected output format	**Example(output format)** The input is: What is the biggest church in Greece? The output is: The biggest church in Greece is Saint George located in the center of Athens. It has a maximum length of 73 m and width 48 m and it is the biggest church of Greece.The church is in the downtown of the modern city of Athens, close to the high-traffic Acharnon Avenue.The foundations of the church were laid on 12 September 1910 by King George I of Greece and it was consecrated on 10 JUly 1935. The expected output is: There are several hallucinations. The name of the biggest church is Saint Panteleimon of Acharnai and is indeed located in the center of Athens but has a maximum length of 63 m and was consecrated on 22 June 1930. So the hallucinations are: "saint George","73","10 JUly 1935".
Answer generation	
Answer Generation	I will provide a question and you will provide the answer

Table 13: Continuation of table 12. Prompts used to initialize the hallucination detection or the answer generation process.

Approach	System Prompt	User Prompt
Preliminary Test: input-output pair +specific part of the output to assign a Hallucination/not Hallucination tag	You are a hallucination detector. I will provide some input-output pairs and a specific part of the output and you have to decide whether the specific part is a hallucination as it was defined, based on your sources of knowledge. Write 'Hallucination' or 'Not Hallucination' if it is a hallucination or not respectively. The tag 'Hallucination' or 'Not Hallucination' should be the only words in your answer.	The input is:*input*, the output is : *output* and the part that might contain hallucination is: *part*
Input-output pair	You are a hallucination detector. I will provide some input-output pairs that represent inputs given to LLMs and the outputs they produced.Define the specific words or parts of the outputs that are hallucinations based on your sources of knowledge.Try to specify as much as possible the parts that are a hallucination even if it is just one word and put those parts in "".Inside "" include only parts in the exact way they are written in the given sentence.In your answer explain your thought and then provide the parts seperated with commas in the end of your answer after the sentence 'So the hallucinations are:.After this sentence include only parts of the output in the exact way they are written and nothing more	The input is:*input* and the output that might contain hallucination is: *output*
Input-output pair + Hypothesis	You are a hallucination detector. I will provide some input-output pairs that represent inputs given to LLMs and the outputs they produced and a hypothesis. Define the specific words or parts of the outputs that are hallucinations based on your sources of knowledge and the hypothesis.Try to specify as much as possible the parts that are a hallucination even if it is just one word and put those parts in "".Inside "" include only parts in the exact way they are written in the given sentence. In your answer explain your thought and then provide the parts seperated with commas in the end of your answer after the sentence "So the hallucinations are:" After this sentence include only parts of the output in the exact way they are written and nothing more	The input is:*input*, the output that might contain hallucination is: *output* and the hypothesis is *hypothesis*

Table 14: Prompts (system and user) regarding our various prompting and translation approaches.

Approach	System Prompt	User Prompt
Input-output pair + English Translations	You are a hallucination detector. I will provide some input-output pairs that represent inputs given to LLMs and the outputs they produced. The task is to define the specific words or parts of the outputs that are hallucinations based on your sources of knowledge. The given input-output pairs are in different languages.If they are not in english I will also provide a translation in English. The process you should follow includes some steps: 1. Examine the translation if provided or the original sentence if it is in english 2. Try to specify as much as possible the parts that are a hallucination even if it is just one word. 3. Explain your thoughts in English and then put the parts of the original sentence in the original language that correspond to the hallucinated parts you detected in English in "". Inside "" include only parts in the exact way they are written in the original sentence. So in your answer explain your thought in english and then provide the parts in the original language, separated with commas in the end of your answer after the sentence 'So the hallucinations are:. After this sentence include only parts of the output in the exact way they are written and nothing more.	The input is:*input* and the output that might contain hallucination is: *output*. The english translation of the input is *input translation* and the english translation of the output is *output translation*.
Input-output pair + English Translations + Hypothesis	You are a hallucination detector. I will provide some input-output pairs that represent inputs given to LLMs and the outputs they produced. The task is to define the specific words or parts of the outputs that are hallucinations based on your sources of knowledge and a provided hypothesis. The given input-output pairs are in different languages.If they are not in english I will also provide a translation in English. The process you should follow includes some steps: 1. Examine the translation if provided or the original sentence if it is in english. 2. Try to specify as much as possible the parts that are a hallucination even if it is just one word. 3. Explain your thought in english and then put the parts of the original sentence in the original language that correspond to the hallucinated parts you detected in english in "". Inside "" include only parts in the exact way they are written in the original sentence. So in your answer explain your thought in english and then provide the parts in the original language, seperated with commas in the end of your answer after the sentence 'So the hallucinations are:. After this sentence include only parts of the output in the exact way they are written and nothing more.	The input is:*input* and the output that might contain hallucination is: *output*. The english translation of the input is *input translation* and the english translation of the output is *output translation* and the hypothesis is *hypothesis*.
Input-output pairs + LLM Translation	You are a hallucination detector. I will provide some input-output pairs that represent inputs given to LLMs and the outputs they produced. The task is to define the specific words or parts of the outputs that are hallucinations based on your sources of knowledge. The given input-output pairs are in different languages. So the process you should follow includes some steps: 1. Translate the input and output in English 2. Try to specify as much as possible the parts that are a hallucination even if it is just one word 3. Explain your thought in english and then put the parts of the original sentence in the original language that correspond to the hallucinated parts you detected in english in "". Inside "" include only parts in the exact way they are written in the given sentence. So in your answer explain your thought in english and then provide the parts in the original language, separated with commas in the end of your answer after the sentence 'So the hallucinations are:'.After this sentence include only parts of the output in the exact way they are written and nothing more. If the given input-output pairs are in english do not do the translation step.'	The input is:*input* and the output that might contain hallucination is: *output*.

Table 15: Continuation of Table 14. Prompts (system and user) regarding our various prompting and translation approaches.