Swushroomsia at SemEval-2025 Task 3: Probing LLMs' Collective Intelligence for Multilingual Hallucination Detection

Sandra Mitrović[†*] **Joseph Cornelius**[†*] **David Kletz**[†*]
Ljiljana Dolamić[‡] **Fabio Rinaldi**[†]
[†] Dalle Molle Institute for Artificial Intelligence Research (IDSIA), Switzerland
[‡] armasuisse, Science & Technology, Switzerland
{sandra.mitrovic,joseph.cornelius,david.kletz,fabio.rinaldi}@idsia.ch
ljiljana.dolamic@armasuisse.ch

Abstract

This paper introduces a system designed for SemEval-2025 Task 3: Mu-SHROOM, which focuses on detecting hallucinations in multilingual outputs generated by large language models (LLMs). Our approach leverages the collective intelligence of multiple LLMs by prompting several models with three distinct prompts to annotate hallucinations. These individual annotations are then merged to create a comprehensive probabilistic annotation. The proposed system demonstrates strong performance, achieving high accuracy in span detection and strong correlation between predicted probabilities and ground truth annotations.

1 Introduction

Hallucinations in large language models (LLMs) are a widely-known problem critical for their trustworthiness (Hong et al., 2024; Mitrović et al., 2024). The detection of hallucinations presents a challenge due to the absence of a standardized definition (Venkit et al., 2024; Mishra et al., 2024). Moreover, different LLMs may identify different parts of the same text as hallucinations and in general, different LLMs have different hallucination rates[1](Mishra et al., 2024). Furthermore, despite some on-going research, hallucinations are still not very well explored in multilingual setups (Zhang et al., 2023; Xu et al., 2024).

In order to contribute to the research on multilingual and multimodel hallucinations, Mu-SHROOM ("Multilingual Shared-task on Hallucinations and Related Observable Overgeneration Mistakes") (Vázquez et al., 2025), a SemEval-2025 Task-3, was proposed. The task focuses on detection of hallucination spans in the outputs of instruction-tuned publicly available LLMs, covering 14 different languages: modern standard Arabic (AR), Basque (BA), 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). The dataset is divided into validation (labelled), train and test (both unlabelled) sets (for details see Appendix, Table 7). Each data instance consists of model input (question), information about question language and model used, and model output (LLM answer) (see Fig. 1). Validation data additionally contains the list of soft and hard labels while train and test data contain, instead, list of model tokens and logits. Soft labels represents a list of hallucination spans, denoting the index of the starting and end-

Example 1
INPUT: When did Karen Percy become a Member of the Order of Canada?
OUTPUT: Karen Elizabeth Perky was made a member of Order in Canada on December 10, 2018 .

Example 2
INPUT: What was the population of Labalme, France in 1982?
OUTPUT: Inhabitants of the commune Labal e me in 1984 were 1,976 .

Example 3
INPUT: How many people were living in Valeille in 2020?
OUTPUT: In 2020, Valeille s had a population of 2,961 .

Figure 1: An illustration of test observations the model input (INPUT) and output (OUTPUT) with following color coding: span annotated as hallucination in ground truth , as hallucination by our system S2 , and agreement between our system S2 and ground truth .

[*]These authors contributed equally to this work. Mitrović focused on conceptual design, evaluation, data preparation, and publication. Cornelius focused on the development of System 2, and Kletz on the creation of System 1 and the evaluation.

[1]Also see: https://huggingface.co/spaces/vectara/Hallucination-evaluation-leaderboard

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1810–1827

ing character and the hallucination probability of each span. Hard labels are obtained by removing spans from soft labels with hallucination probability ≤ 0.5. Based on these, official task evaluation metrics were defined. One is intersection-over-union (*IoU*) of characters marked as hallucinations in the ground truth versus predicted. Another is character-level correlation (*corr*) of the empirical probabilities observed by annotators (ground truth) and the predicted probability.

We utilized two distinct approaches, referred to as System 1 (S1) and System 2 (S2), which rely on in-context learning, exploiting different prompts and underlying (mostly proprietary) models. While S1 serves as a simple baseline, with S2 we additionally aimed to harness the knowledge of different models in an ensemble-like manner, hoping that their collective intelligence would help mitigate occasional suboptimal outputs from individual models. In this context, collective intelligence refers to the emergent consistency and insight gained by combining outputs from multiple LLMs and prompt variations. Rather than relying on a single model, we leverage the diversity and statistical significance of responses to identify patterns, disagreements, and potential hallucinations. By analyzing the agreement between models and correlating it with human annotations, we explore whether this collective signal can approximate human judgment. This approach allows us to investigate how aggregated model outputs can enhance hallucination detection with respect to their correlation to human annotations.

In the remaining of the paper, due to space limitations, we mainly focus on our best performing approach S2[2], which in the official IoU ranking scored 4^{th} and 5^{th} for FR and IT, respectively, but we also provide some details on S1 in the Appendix. The ensemble strategy performs particularly well for *corr* metric (it ranked 1^{st}, 3^{rd} and 4^{th} for EN, DE and FR, respectively), but we have identified, as well, some particularly performative models and prompts for *IoU*. We provide a detailed analysis of considered closed-weight large language models' performance.

2 Related Work

This Mu-SHROOM task builds upon Semeval 2024 monolingual SHROOM task 6 (Mickus et al., 2024), which comprised three NLG tasks divided in two streams (model-agnostic vs. model-aware) but was scoped less ambitiously: participants were supposed to *only* perform binary classification to identify hallucinations, without indicating hallucination spans. Nevertheless, some ideas from 2024 edition's winning approaches were inspirational for us. In particular, we noticed that 4 out of 6 best approaches were reporting excellent results using closed-weight models (Mehta et al., 2024; Obiso et al., 2024; Liu et al., 2024; Allen et al., 2024) as well as that high performance is not readily achieved with off-the-shelf LLMs and systems (Mehta et al., 2024; Belikova and Kosenko, 2024). In particular, the best performing model (Mehta et al., 2024) resorted to a meta-regressor framework aggregating uncertainty signals from multiple LLMs, which was the motivation for models' output merging in our S2. Moreover, we draw inspiration from direct prompting strategies which have already been explored to evaluate factual consistency (Chen et al., 2023), assess the self-alignment capabilities of LLMs with respect to factuality (Zhang et al., 2024), and detect confabulations, a specific subclass of hallucinations, by eliciting multiple candidate responses (Farquhar et al., 2024; Verspoor, 2024).

3 Challenges Related to Ground Truth Annotations

We observed that the ground truth annotations for EN lack consistency: the characters included in the span of the same hallucination type differ from sentence to sentence. For example, in some cases where the names of the places used in the question were misspelled in the model outputs, ground truth annotates complete name as a hallucination (see Ex. 2 in Figure 1) while in others, only the added character ('s' in Ex. 3 in Figure 1) was annotated as such. Some other languages (e.g. IT) did not have this type of issues (or had very few which, however, have not influenced annotations).

4 System Overview

In order to facilitate the readability of our article, we use abbreviations to designate the LLMs employed (see Appendix B.1 for the list of exploited LLMs and their respective abbreviations).

Our systems have been evaluated only on the languages included in the validation set.

[2]Our code and data is available at: https://github.com/IDSIA-NLP/mushroom/

4.1 System 1 (S1) Description

As a reference system, we perform a few-shot prompting on test data for our languages of interest. More precisely, we use 3 random examples from validation data per language to perform prompting on test data. We limit ourselves to only two models, *g3.5* and *haiku-3*. For Chinese, we used *haiku-3* since *g3.5* was causing issues due to long contexts. For other languages, we noticed that *g3.5* is performing much better than *haiku-3*.

4.2 System 2 (S2) Description

In S2, we explore the capabilities of hallucination detection of the state-of-the-art service LLMs for in-context few-shot learning. Our approach simulates the original annotation process using multiple artificial annotators, each instantiated through a different LLM service combined with varying prompts. The outputs of these artificial annotators are then aggregated into a single probabilistic annotation.

To construct a diverse set of artificial annotations, we employ six different LLMs and three distinct prompting strategies, resulting in a total of 18 unique model-prompt combinations. Each model is accessed via its respective API, ensuring consistency in inference settings. The exact model identifiers are provided in Appendix B.1.

To facilitate in-context few-shot learning, we randomly select language-specific examples from the evaluation dataset. Furthermore, each example must contain at least three annotated hallucinations based on hard labels. This ensures that the models are exposed to relevant patterns in hallucination annotation.

For annotation, the models are prompted to mark hallucinations using an inline XML format with the tags "$<h>$" and "$</h>$". The provided examples are formatted in the same style to maintain consistency in the learning process.

We used the following three prompting strategies:

- Prompt V1: A general prompt with a short task description (with 2 in-context examples)
- Prompt V2: A detailed task explanation incorporating chain-of-thought reasoning (with 1 in-context example)
- Prompt V3: A general prompt with an explicit instruction to be highly sensitive to hallucinations, marking spans even if there is only a low probability (with 2 in-context examples)

Once the models generate annotations, we convert the inline XML hallucination tags into offset-based annotations. A predicted hallucination span is considered valid only if it exactly matches the corresponding portion of the original text; otherwise, it is discarded.

After collecting annotations from all artificial annotators, we aggregate them into a single probability-based annotation scheme, producing soft labels that quantify confidence in each hallucination span. First we normalize the character-level spans by extracting and sorting hallucination span boundaries predicted by different models, and pairing adjacent boundaries to define sub-spans, forming continuous intervals that maintain character-level consistency. Next, we compute for each sub-span the probability of it being a hallucination as follows:

$$P(H) = \frac{N_H}{N_A}$$

where $P(H)$ is the hallucination probability of a given sub-span. N_H is the number of annotators (LLMs) that marked the sub-span as a hallucination. N_A is the total number of annotators.

Furthermore, we included two merging variations, where we excluded the 3 and 6 worst-performing runs (model + prompt variation) with respect to *corr* score based on the English validation data—denoted as $m_{\setminus 3}$ and $m_{\setminus 6}$, respectively. To ensure diversity, we applied the constraint that at least one run from each model had to be included. This approach aimed to filter out the lowest-performing runs while maintaining variety.

S2 allows for a probabilistic measure of hallucination confidence, simulating the variability and uncertainty inherent in human annotation.

5 Quantitative Findings

We provide simple statistics in Table 1 regarding matched and mismatched annotation spans across data instances. We noticed that these statistics vary from one language to the other. For example, for IT we have 69 out of 150 instances (46%) where S2 and ground truth annotation spans completely match[3], while for EN this percentage decreases to 14.29 and eventually for ZH to only 1.33%.

[3]Note that a single instance can have multiple annotated spans both in S2 and ground truth, hence by overlapping spans we consider both coinciding in span number as well as in the start and ending character of each span.

Lang.	Match(%)		Mismatch(%)	
	full	nosp	S2-nosp	GT-nosp
EN	14.29	1.95	9.09	1.30
FR	6.67	0.00	3.33	0.00
IT	46.00	0.00	3.33	0.00
DE	13.33	1.33	8.67	1.33
FI	11.33	0.00	2.00	0.00
ES	17.11	6.58	6.58	1.97
HI	46.00	0.00	6.00	0.00
SV	10.20	0.68	2.72	1.36
AR	18.00	2.67	4.00	2.00
ZH	1.33	1.33	15.33	0.67

Table 1: Statistics of matches and mismatches between S2 and ground truth (GT), in percentages. Notation: *full*: completely matching spans, *nosp*: no spans in both ground truth and S2, *S2-nosp*: no spans in S2 (but spans present in ground truth), *GT-nosp*: opposite of *S2-nosp*.

	Lang	Strategy	Desc	Prompt	BL score	Our score	Rank
IoU	EN	single	g4o	v2	0.349	0.503	**12/41**
	FR	single	sonnet	v3	0.454	0.594	**4/30**
	IT	merged	$m^*_{\setminus 3}$	-	0.283	0.727	**5/28**
	DE	single	sonnet	v3	0.345	0.539	**15/28**
	FI	single	sonnet	v3	0.486	0.644	2/27
	ES	merged	$m^*_{\setminus 6}$	-	0.185	0.513	4/32
	HI	merged	$m^*_{\setminus 6}$	-	0.271	0.721	4/24
	SV	single	sonnet	v1	0.537	0.616	5/27
	AR	single	g4o	v2	0.361	0.600	5/29
	ZH	single	g4o	v3	0.477	0.331	21/26
Corr	EN	merged	$m^*_{\setminus 3}$	-	0.119	0.649	**1/41**
	FR	merged	$m^*_{\setminus 6}$	-	0.020	0.591	**4/30**
	IT	merged	$m^*_{\setminus 6}$	-	0.080	0.739	**7/28**
	DE	merged	$m^*_{\setminus 6}$	-	0.107	0.616	**3/28**
	FI	merged	$m^*_{\setminus 6}$	-	0.092	0.648	2/27
	ES	merged	$m^*_{\setminus 6}$	-	0.036	0.641	1/32
	HI	merged	$m^*_{\setminus 6}$	-	0.143	0.739	5/24
	SV	merged	$m^*_{\setminus 6}$	-	0.097	0.608	1/27
	AR	merged	$m^*_{\setminus 6}$	-	0.119	0.635	5/29
	ZH	merged	m	-	0.088	0.401	11/26

Table 2: Detailed results for S2 and different languages showing strategy (single model or merged) and prompt version providing the best IoU score. Ranks in **boldface** are official rankings, while the others represent the ranking that would have been obtained if the results were submitted within the official deadline. * : Details on included models for merging can be seen in Appendix B.2. The BL column presents the results achieved by the best baseline provided by the organizers for each language. For IoU, the baseline is always mark-all.

Additionally, we noticed that for all languages mismatches related to no-spans were far more frequent for the direction when spans were present in ground truth and missing in S2 (*S2-nosp*) than vice versa (*GT-nosp*). However, percentages of *S2-nosp* vary greatly across languages, being the best for FI, then SV, and the worst for EN and ZH. We performed yet another analysis, where together with the ratio of overlapping spans (*ol_spans*) we also looked at the ratio of overlapping characters (*ol_chars*) for each test instance, comparing the S2 annotations with those of ground truth. Comparing inter-quantile ranges (and medians) of *ol_spans* and *ol_chars* distributions (see Figure 2), we can see not only differences in scores between various languages but also that, as expected, reaching span overlap is much harder to achieve than character overlap (the latter also aligns better with *IoU*).

IoU The official rankings of the Mu-SHROOM task were provided only with respect to the *IoU* score. As showcased in the Table 2, S2 scored quite well in the official rankings (**boldface**) for FR and IT, ranking 4^{th} out of 30 teams and 5^{th} out of 28 teams, respectively. In the same table, we provide, as well, post-deadline to-be rankings[4] for other languages, generated using the official Mu-SHROOM evaluation scripts. Looking at model and prompt comparisons, we observe that among single models *g4o* and *sonnet* are consistently outperforming competitors on all languages (see Figure 3) while prompt *v3* among prompts performs the best (see Figure 3, right).

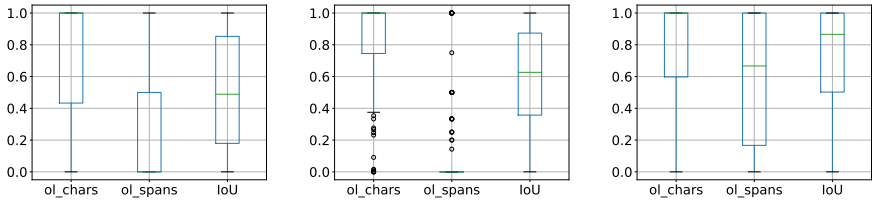
Results for less performing S1 can be seen in Appendix (Table 8).

Corr Our merged configuration demonstrated particular effectiveness in measuring correlation, yielding the best results across all languages, with performance on *corr* surpassing even that of *IoU*. Specifically, the $m_{\setminus 6}$ configuration achieved the highest performance for eight languages, compared to $m_{\setminus 3}$ and m, which were the bests for only one language each. Furthermore, these configurations proved highly effective relative to other teams, allowing us to achieve the best results for three languages (SV, EN, and ES) and rank within the top five for eight out of the ten languages evaluated (see also mean *corr* plots in Appendix D). However, Chinese remains challenging due to annotation difficulties, resulting in a correlation value below 0.5 and performance significantly lower than that of the top-performing teams (ranking 11^{th} out of 20).

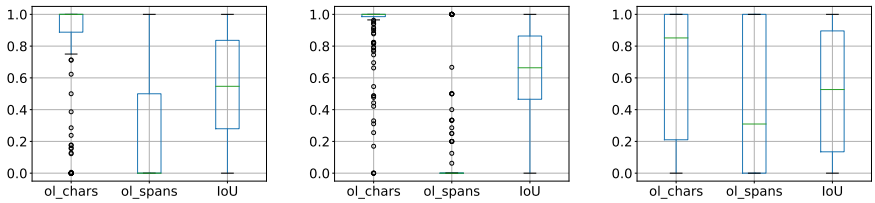
6 Qualitative Analysis

We performed qualitative analysis for IT and FR comparing our best S2 models for these languages with ground truth. Some of the observed patterns in annotation discrepancies between S2 and ground truth are reported in (Table 3 and Appendix Table 9 for IT and FR, respectively). Even though some

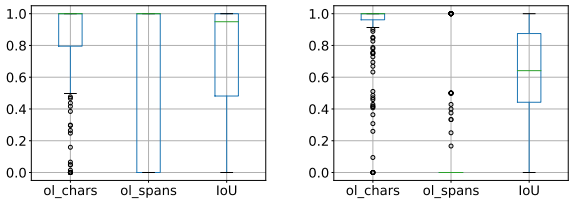
[4]We have not managed to apply our system to all languages during the allotted period for this shared task. Therefore, by "post-deadline to-be rankings" we refer to the rankings which would have been obtained for FI, ES, HI, SV, AR, ZH have we had submitted our system output within the deadline.



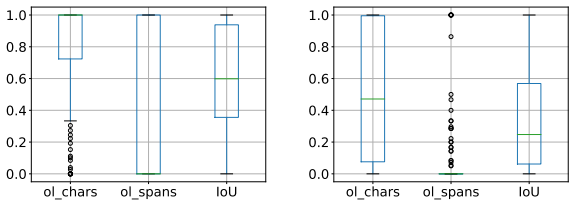
(a) English (b) French (c) Italian



(d) German (e) Finnish (f) Spanish



(g) Hindi (h) Swedish



(i) Arabic (j) Chinese

Figure 2: Boxplots for the ratio of overlapping chars (*ol_chars*), the ratio of overlapping spans (*ol_spans*) and IoU (*IoU*), all calculated on instance-level between S2 and ground truth, per language.

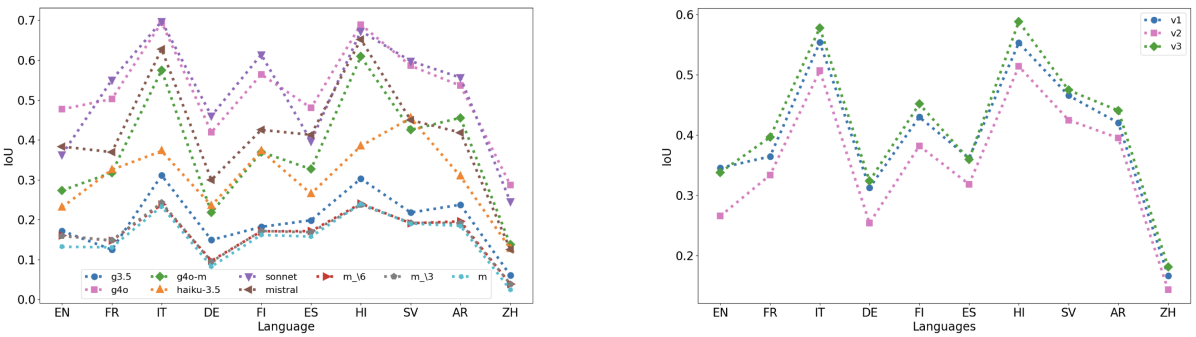


Figure 3: Mean IoU scores by LLM (left) and by version of prompt (right). For the mean by LLM, each model was tested with three versions of the prompt, and the mean IoU score across these versions is reported for each model. For the mean by version, each version of prompt was tested with 6 different LLM, and the mean IoU score across these LLMs is reported for each version.

ID	S2 span(s)	Ground truth span(s)	Comment
1	"nel 1952"	"1952" (1)	S2 involves prepositions and connectors
2	"Jelgava, in Lettonia"	"Jelgava", "Lettonia" (2)	
3	"ottobre del 1991"	"ottobre", "1991" (2)	
4	"Jamie Jones e Sean Tyas"	"Jamie Jones", "Sean Tyas" (2)	
5	"due medaglie d'oro olimpiche nel 1952"	"due", "d'oro", "1952" (3)	S2 encompasses larger context
6	"fu campione panamericano di nuoto"	"panamericano" and "nuoto" (2)	
7	"Neuchâtel, nella regione della Svizzera occidentale"	"Neuchâtel", "Svizzera occidentale"(2)	
8	"5 febbraio 1942"	"5" and "1942"	S2 is less precise than ground truth ("febbraio" and "Mary" are correct)
9	"Dr. Mary Blakiston"	"Blakiston"	
10	"Jomsland, che è un'isola situata nel Mar Baltico"	"Jomsland"	S2 annotates an extra information
11	"100°F (38°C)"	"100°F (38°C"	imperfection of ground truth annotation

Table 3: Some minor problems observed for IT when comparing S2 with ground truth. For ID=10, IoU=0.16.

of these look trivial, their repercussion on the final result can be immense (see ex. 10 in Table 3 for IT). We have also noticed that S2 for IT performs quite well in recognizing named entities (NE) which participate in hallucination spans. On 150 observations, it missed to recognize only 18 NE in 14 different observations compared to ground truth, while it had perfect precision. For FR, S2 was making much more errors for NE. Details about all NE errors of S2 for IT and FR can be seen in the Appendix (Table 10).

INPUT: Chi era raffigurato nel simbolo del partito Fronte Democratico Popolare?

OUTPUT (sonnet): Il simbolo del partito Fronte Democratico Popolare raffigurava la testa di una donna.
OUTPUT (g4o): Il simbolo del partito Fronte Democratico Popolare raffigurava la testa di una donna.
OUTPUT (haiku-3.5 and g3.5): Il simbolo del partito Fronte Democratico Popolare raffigurava la testa di una donna.

Figure 4: An example of how different underlying models lead to different S2 annotations for IT. In red, S2 annotations, top to bottom from best performing (*sonnet* and *g4o*, second best) to worst performing models (*haiku-3.5* and *g3.5*) for IT. Annotation spans are shorter for better performing models.

When comparing different model hallucination annotation, we perceived that models performing worse for IT were tending to annotate more hallucination spans than best performing models (see Figure 4). The contrary of this behavior was noticed for FR. We also noted that observations with shorter length tend to have higher IoU scores. This tendency is particularly pronounced for IT while it is less evident for other languages (see Appendix, Figure 8).

Additionally, Figure 7 in Appendix shows two examples (in IT and FR) of more drastic annotation problems.

6.1 Open-Weight Model Comparison

To assess whether the task could be effectively performed using only Open-Weight LLMs (OWMs), we reused the S2 prompts with a dozen OWMs (for details see Appendix, Table 5).

The results were significantly poorer compared to those obtained with service models. IoU scores were 1.7 to 3.5 times higher for closed-weight models than for OWMs, with a maximum IoU of 0.486 for Italian. Notably, substantial variation was observed across both models and prompts. For each language, the highest IoU scores were consistently achieved by either *ministral-8B* or *mistral-7B*, particularly with the V1 or V3 prompts. Conversely, certain models, like Gemma and Qwen rarely produced annotations.

Correlation measurements, however, exhibited better performance. In all languages, correlations surpassed those obtained by the best baseline models, further supporting the reliability of these metrics. The highest correlations were consistently achieved by a merged model, reinforcing the effectiveness of collective intelligence in addressing the task.

7 Conclusion

This paper outlines our system for Mu-SHROOM, a shared task with key challenges, such as multilingualism, hallucination span detection without annotated training data, inconsistencies in human annotations. Despite all these, our collective intelligence approach exploiting annotation potential of diverse close-weight LLMs and accompanying prompts, demonstrated strong effectiveness, particularly in achieving high correlation with the ground truth annotations. Our future work will focus on investigating noted imbalances varying across different languages and inputs, as well as more refined comparison exploiting the Open-Weight Models.

References

Bradley P Allen, Fina Polat, and Paul Groth. 2024. Shroom-indelab at semeval-2024 task 6: Zero-and few-shot llm-based classification for hallucination detection. *arXiv preprint arXiv:2404.03732*.

Julia Belikova and Dmitrii Kosenko. 2024. Deeppavlov at semeval-2024 task 3: Multimodal large language models in emotion reasoning. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1747–1757.

Shiqi Chen, Siyang Gao, and Junxian He. 2023. Evaluating factual consistency of summaries with large language models. *arXiv preprint arXiv:2305.14069*.

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

Junyuan Hong, Jinhao Duan, Chenhui Zhang, Zhangheng Li, Chulin Xie, Kelsey Lieberman, James Diffenderfer, Brian R. Bartoldson, Ajay Kumar Jaiswal, Kaidi Xu, Bhavya Kailkhura, Dan Hendrycks, Dawn Song, Zhangyang Wang, and Bo Li. 2024. Decoding compressed trust: Scrutinizing the trustworthiness of efficient LLMs under compression. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 18611–18633. PMLR.

Wei Liu, Wanyao Shi, Zijian Zhang, and Hui Huang. 2024. Hit-mi&t lab at semeval-2024 task 6: Deberta-based entailment model is a reliable hallucination detector. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1788–1797.

Rahul Mehta, Andrew Hoblitzell, Jack O'keefe, Hyeju Jang, and Vasudeva Varma. 2024. Halu-nlp at semeval-2024 task 6: Metacheckgpt-a multi-task hallucination detection using llm uncertainty and meta-models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 342–348.

Timothee Mickus, Elaine Zosa, Raúl Vázquez, 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.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *arXiv preprint arXiv:2401.06855*.

Sandra Mitrović, Matteo Mazzola, Roberto Larcher, and Jérôme Guzzi. 2024. Assessing the trustworthiness of large language models on domain-specific questions. In *EPIA Conference on Artificial Intelligence*, pages 305–317. Springer.

Timothy Obiso, Jingxuan Tu, and James Pustejovsky. 2024. Harmonee at semeval-2024 task 6: Tuning-based approaches to hallucination recognition. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1322–1331.

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.

Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. 2024. " confidently nonsensical?": A critical survey on the perspectives and challenges of'hallucinations' in nlp. *CoRR*.

Karin Verspoor. 2024. Fighting fire with fire - using llms to combat llm hallucinations. *Nature*, 630(8017):569–570.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. *arXiv preprint arXiv:2402.09267*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

A Limitations

Our data annotations rely extensively on the output of proprietary large language models (LLMs) accessed via API-based services. These models have a limited lifespan and are frequently updated, deprecated, or replaced on their respective platforms. Consequently, our results are strictly applicable to the specific versions of the LLMs used at the time of annotation and may not generalize to future iterations.

Additionally, due to time constraints, we were unable to submit results for all languages before the official deadline. Instead, we ran our systems after

the deadline and subsequently integrated our results into the official rankings. This approach assumes that the rankings remained stable and that no other teams were in a similar situation. However, it is likely that other teams also faced similar challenges. As a result, the rankings we assigned ourselves may be optimistic, as no other team's score increase placed them above ours in the updated standings.

B Details on used models

B.1 Models used and abbreviations

A list of all the LLMs used and the respective abbreviations we have used to designate them is available in Table 4.

LLM	Abbreviation
gpt-3.5-turbo	g3.5
gpt-4o-2024-08-06	g4o
gpt-4o-mini-2024-07-1	g4o-m
claude-3-haiku-20240307	haiku-3
claude-3-5-haiku-20241022	haiku-3.5
claude-3-5-sonnet-20241022	sonnet
mistral-large	mistral

Table 4: List of used close-weight models and their corresponding abbreviations.

LLM	Abbreviation	Quantization
Llama-3.1-8B-Instruct	Llama-3.1-8B	4Bit QLoRA
Llama-3.2-1B-Instruct	Llama-3.2-1B	-
Llama-3.2-3B-Instruct	Llama-3.2-3B	4Bit QLoRA
Mistral-7B-Instruct-v0.3	mistral-7B	4Bit QLoRA
Ministral-8B-Instruct-2410	ministral-8B	4Bit QLoRA
DeepSeek-R1-Distill-Llama-8B	DS-R1-L	4Bit QLoRA
DeepSeek-R1-Distill-Qwen-1.5B	DS-R1-Q	-
gemma-2-9b-it	gemma-2-9B	4Bit QLoRA
gemma-2-2b-it	gemma-2-2B	-
Qwen2.5-7B-Instruct-1M	Qwen-2.5-7B	4Bit QLoRA
Qwen2.5-1.5B-Instruct	Qwen-2.5-1.5B	-

Table 5: List of Open-Weight Models (OWM) used locally, their corresponding abbreviations and the quantization used for the inference.

B.2 Merged models

This section displays the details for merging approaches used in System 2. Table 6 shows the runs excluded from the merging without the worst performing n ($m_{\setminus n}$) runs, where each run is a combination of a model and prompt version. The performance is measured based on the Corr score and English validation dataset.

S2 $m_{\setminus 3}$	S2 $m_{\setminus 6}$	OWM $m_{\setminus 3}$	OWM $m_{\setminus 6}$
g3.5 – v2	g3.5 – v2	DS-R1-Q – v2	DS-R1-Q – v2
haiku – v2	haiku – v2	Qwen-2.5-7B – v2	Qwen-2.5-7B – v2
sonnet – v2	sonnet – v2	Qwen-2.5-7B – v3	Qwen-2.5-7B – v3
-	haiku – v1	-	gemma-2-2B – v2
-	g3.5 – v1	-	gemma-2-9B – v1
-	g4o-m – v2	-	gemma-2-2B – v3

Table 6: List of the models excluded from the merging without the worst performing n ($m_{\setminus n}$) runs (model – prompt version) with regard to the Corr score for System 2 (S2) and System 2 with Open-Weight Models (OWMs).

C Dataset statistics

Basic statistic of dataset with respect to language and validation/train/test sets is provided in Table 7.

Lang	Validation data					Train data			Test data				
	# inst.	# LLMs	$\bar{m}$ soft sp.	$\bar{m}$ hard sp.	$\bar{m}$ out. len.	# inst.	# LLMs	$\bar{m}$ out. len.	# inst.	# LLMs	$\bar{m}$ soft sp.	$\bar{m}$ hard sp.	$\bar{m}$ out. len.
FR	50	5	10	4	444	1850	5	540	150	5	8	3	322
ES	50	3	7	2	494	492	3	521	152	3	14	3	461
EU	-	-	-	99	2	8	3	156					
AR	50	3	4.36	2	94	-	-	-	150	3	5	2	106
FA	-	-	-	100	6	3	1	87					
DE	50	3	5	2	161	-	-	-	150	3	5	2	148
CA	-	-	-	100	3	4	2	144					
HI	50	3	4	2	153	-	-	-	150	3	3	1	131
IT	50	4	5	2	191	-	-	-	150	4	5	2	166
CS	-	-	-	100	2	10	4	306					
FI	50	2	9	3	245	-	-	-	150	2	9	3	250
EN	50	3	145	3	244	809	3	217	154	3	17	3	239
SV	49	3	6	2	157	-	-	-	147	3	5	2	130
ZH	50	4	49	10	406	200	5	375	150	5	47	11	320

Table 7: Basic statistics showing per language and validation/train/test set: number of instances, number of different models used, average number of soft and hard spans per instance, average model output length (in chars). Average numbers are rounded for better readability. Train data is not labelled, hence information on spans is missing for all languages. Some languages are additionally left out from validation and train data.

D Additional results

System	En	It	Fi	Fr	De	Es	Sv	Ar	Hi	Zh
S1	0.24	0.47	0.50	0.29	0.36	0.25	0.35	0.30	0.45	0.20*
S2	0.50	0.73	0.644	0.59	0.54	0.51	0.62	0.60	0.72	0.33

Table 8: Results in terms of IoU for different systems and languages. * : The Chinese S1 is produced with haiku-3.

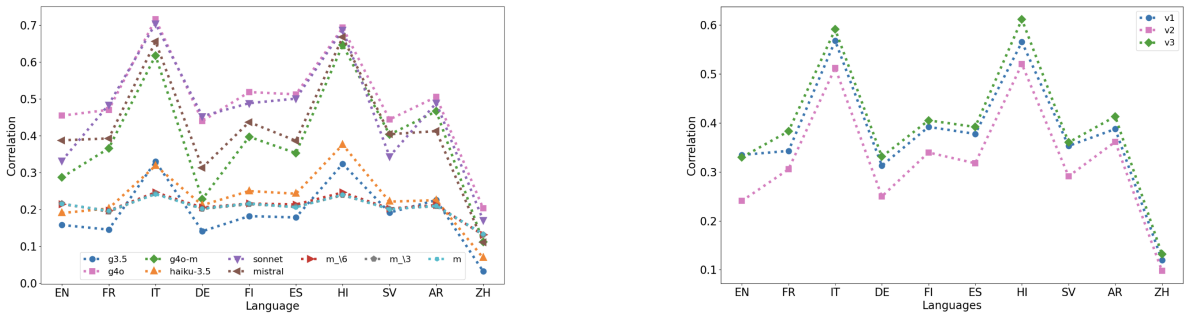


Figure 5: Mean corr grouped by model (left) and by version of prompt (right).

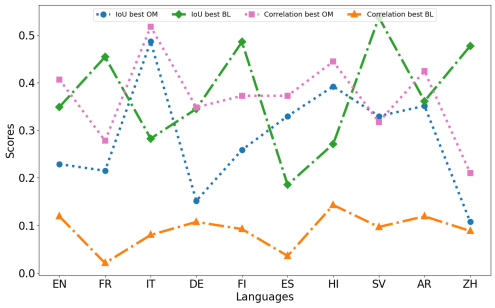


Figure 6: IoU scores and correlation of the best open model for each language

D.1 Qualitative analysis - additional figures

Example IT

INPUT: Grazie a quale progetto è stato scoperto 19521 Chaos?
OUTPUT: 19521 Chaos è stato scoperto nel **2003** come **asteroide potenzialmente pericoloso** . Il progetto che ha scoperto 19521 Chaos è il **Catalina Sky** Survey, **un progetto di ricerca astronomica che utilizza telescopi per monitorare il cielo notturno e rilevare oggetti vicini alla Terra come asteroidi e comete. Il Catalina Sky Survey è stato avviato nel 2000 e ha scoperto oltre 100.000 asteroidi e comete** .

Example FR

INPUT: Depuis quelle année le British Academy Film Award du meilleur acteur dans un second rôle est-il décerné ?
OUTPUT: Le British Academy Film Award du meilleur acteur dans un second rôle **a été décerné pour la première fois en 19** **57** .

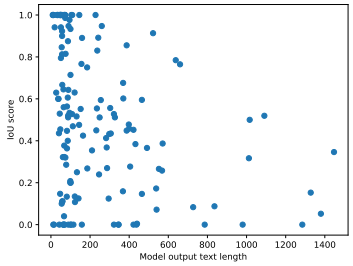
Figure 7: Two examples (IT above / FR below) where S2 performs bad comparing to ground truth (IoU scores: IT: 0.013 / FR: 0.046). Color coding: span annotated as hallucination in ground truth , as hallucination by our system S2 , and agreement between our system S2 and ground truth .

ID	S2 span(s)	Ground truth span(s)	Comment
1	“la première image d’un trou noir par le télescope spatial Hubble”	“de la première image d’un trou noir par le télescope spatial Hubble”	ground truth involves prepositions and connectors
2	“[...]bronze en patinage artistique [...]”	“bronze”, “patinage artistique”	
3	“la famille des Plebeiidae et à l’ordre des Anguilliformes.”	“Plebeiidae”, “Anguilliformes”	S2 encompasses larger context
4	“1 350 hab./km^2”	“1 350”	
5	“Gergely Kulcsár n’a pas remporté de médaille aux championnats d’Europe”	“n”, “pas”, “de”	
6	“300 millions d’année”	“300”	S2 is less precise than ground truth (here "millions d’année" is correct)
7	“Il aurait dû être basé sur le langage de programmation Visual Basic pour l’interface utilisateur”	“Visual Basic”	

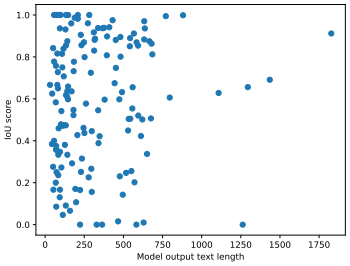
Table 9: Some minor problems observed for French language when comparing S2 with ground truth. For ID=5, IoU=0.086.

Lang.	NE Type	Num. instances	Num. missing entities per instance	example
IT	person	1	2*	Denholm Elliott
IT	person	3	1	
IT	geographical NE	2	2	
IT	geographical NE	5	1	Marna
IT	product	1	2	Adobe Flash, Microsoft Silverlight
IT	MISC	2	1	Catalina Sky Survey
FR	person	3	2	Mark Ronson et Andrew Wyat
FR	person	1	3	
FR	person	1	5	
FR	geographical NE	7	1	Midtown Manhattan
FR	geographical NE	1	3	
FR	geographical NE	1	4	
FR	group	1	1	At the Gates
FR	group	1	10	
FR	group	1	15	
FR	institution	3	1	Académie canadienne du cinéma et de la télévision
FR	creative	1	6	Eye for Eye
FR	MISC	6	1	
FR	MISC	1	3	

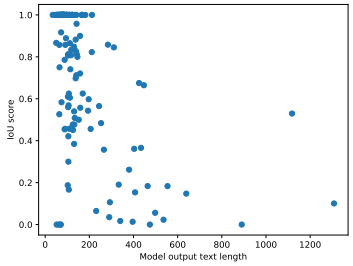
Table 10: Statistics of missing NE types in S2 (with respect to ground truth); * : although it has identified other two person NE in the same instance



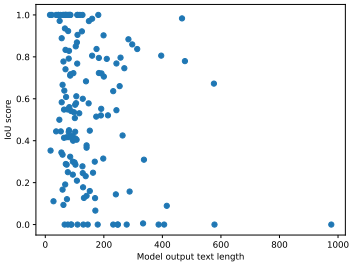
(a) English



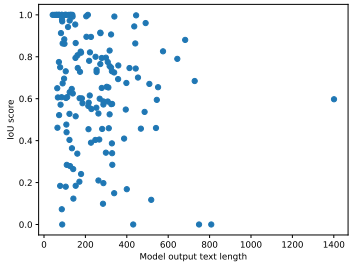
(b) French



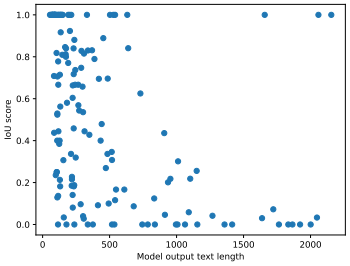
(c) Italian



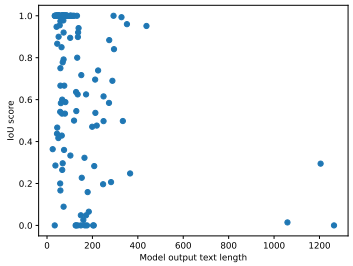
(d) German



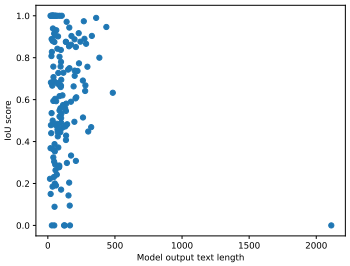
(e) Finnish



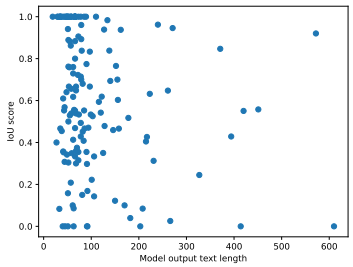
(f) Spanish



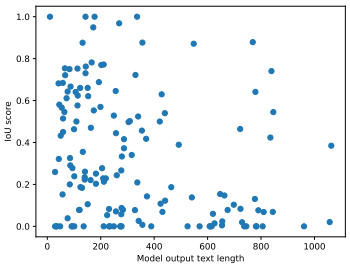
(g) Hindi



(h) Swedish



(i) Arabic



(j) Chinese

Figure 8: IoU scores vs. output text length per observation in the labelled test set for S2.

E Prompts

Prompt : System 1

```
You are tasked with identifying and marking hallucinations in the following answers. A hallucination in this context
      refers to an answer that provides incorrect or fabricated information. Your goal is to review each answer
      relative to the question and highlight any incorrect or unsupported parts of the response using an <h> tag at
      the beginning of a span and a </h> tag at the end. If the answer is factually correct, return it without any
      highlighting.

        For each example, provide only the response sentence (R) with the highlighted hallucinations if present. Do
             not provide explanations or commentary.
        Example 1:
        Q: \colorbox{blue_position_prompt}{\{example\_1\_q\}}
        A: \colorbox{blue_position_prompt}{\{example\_1\_a\}}
        R: \colorbox{blue_position_prompt}{\{example\_1\_r\}}

        Example 2:
        Q: \colorbox{blue_position_prompt}{\{example\_2\_q\}}
        A: \colorbox{blue_position_prompt}{\{example\_2\_a\}}
        R: \colorbox{blue_position_prompt}{\{example\_2\_r\}}

        Example 3:
        Q: \colorbox{blue_position_prompt}{\{example\_3\_q\}}
        A: \colorbox{blue_position_prompt}{\{example\_3\_a\}}
        R: \colorbox{blue_position_prompt}{\{example\_3\_r\}}

        New Question and Answer:
        Q: \colorbox{blue_position_prompt}{\{input\_q\}}
        A: \colorbox{blue_position_prompt}{\{input\_a\}}
        R:\\
```

S2 Prompt: V1 - Brief instruction with in-context learning

```
You are tasked with identifying and marking hallucinations in the following large language model (LLM) answers
      . A hallucination in this context refers to an answer that provides incorrect or fabricated information.
       Your goal is to review each LLM answer (provided in <LLM_Answer>[llm_answer]</LLM_Answer>) relative to
      the question (provided in <Question>[question]</Question>) and highlight any incorrect or unsupported
      parts of the response using ***h>** tags. If the answer is factually correct, return it without any
      highlighting.

For each example, provide only the response sentence (R) with the highlighted hallucinations if present. Do
      not provide explanations or commentary. For structured extraction  use the following format/tags for the
       response: <<<START>>>[final_response_with_hallucinations_marked]<<<END>>>

Important: Ensure that the text remains exactly the same length as the original text, don't change any amount
      of whitespace or newline characters. You should only add tags and not delete any characters. To this end
       a token list is provided for the LLM answer (provided in  <LLM_Answer_in_tokens>[
      LLM_Answer_in_token_list]</LLM_Answer_in_tokens>).

---

Example 1:
<Question> {example_1_q} </Question>
<LLM_Answer> {example_1_a} </LLM_Answer>
<LLM_Answer_in_tokens> {example_1_a_tokens} </LLM_Answer_in_tokens>
<<<START>>> {example_1_r} <<<END>>>

Example 2:
<Question> {example_2_q} </Question>
<LLM_Answer> {example_2_a} </LLM_Answer>
<LLM_Answer_in_tokens> {example_2_a_tokens} </LLM_Answer_in_tokens>
<<<START>>> {example_2_r} <<<END>>>

New Question and Answer:
<Question> {input_q} </Question>
<LLM_Answer> {input_a} </LLM_Answer>
<LLM_Answer_in_tokens> {input_a_tokens} </LLM_Answer_in_tokens>
```

S2 Prompt: V2 - Brief instruction with in-context learning and chain of thought reasoning

```
You are tasked with identifying and marking hallucinations in the following large language model (LLM) answers
      . A hallucination in this context refers to an answer that provides incorrect or fabricated information.
       Your goal is to review each LLM answer (provided in <LLM_Answer>[llm_answer]</LLM_Answer>) relative to
      the question (provided in <Question>[question]</Question>) and highlight any incorrect or unsupported
      parts of the response using **<h>** tags. If the LLM answer contains no hallucinations, return it
      without any highlighting.

In short:
- Carefully read the answer text.
- Highlight each span of text in the answer text that is an overgeneration or hallucination (factual
      distortion, excessive and unsupported output, typographic hallucination, nonexistent entities,
      contradictory statements)
- Your annotations should include only the minimum number of characters in the text that should be edited/
      deleted to provide a correct answer (in the case of Chinese, these will be "character components").
- You are encouraged to annotate conservatively and focus on content words rather than function words. This is
       not a strict guideline, and you should rely on your best judgments.
- Ensure that you double-check your annotations.
- Important: Ensure that the text remains exactly the same length as the original text, don't change any
      amount of whitespace or newline characters. You should only add tags and not delete any characters. To
      this end a token list is provided for the LLM answer (provided in  <LLM_Answer_in_tokens>[
      LLM_Answer_in_token_list]</LLM_Answer_in_tokens>).

To ensure accuracy, follow and write down ALWAYS these reasoning steps first and than provide the final
      response with hallucinations marked:
1. Understand the Question: Analyze the intent and scope of the question. What information does it seek?
2. LLM Answer Break Down: Identify distinct factual claims or statements in the response.
3. Claim Verification:
    - Cross-check with reliable knowledge sources.
    - Determine if the claim is logically consistent with known facts.
    - If a claim is unverifiable or fabricated, it is a hallucination.
4. Identify Other Hallucinations and Overgenerations:
    - Check for typographic errors
    - Identify contradictions.
    - Look for unsupported or excessive information.
5. Final Response:
    - Output only the final response for structured extraction in the format: <<<START>>>[
          final_response_with_hallucinations_marked]<<<END>>>
    - Mark Hallucinations: Surround incorrect or unsupported parts with **<h>** tags.
    - Do not provide explanations or extra formatting.
    - If no hallucinations are found, return the LLM answer as is inside the <<<START>>> and <<<END>>> tags.


---
Example of Question, LLM Answer and Final Response with Hallucinations Marked (but without the reasoning steps
      ):

<Question> {example_1_q} </Question>

<LLM_Answer> {example_1_a} </LLM_Answer>

<LLM_Answer_in_tokens> {example_1_a_tokens} </LLM_Answer_in_tokens>

Response:
    1. Understand the Question: [Here, you would provide a brief analysis of the question's intent and scope.]
    2. LLM Answer Break Down: [Here, you would identify distinct factual claims or statements in the response
          .]
    3. Claim Verification: [Here, you would cross-check each claim with reliable knowledge sources and
          determine if they are logically consistent with known facts.]
    4. Identify Other Hallucinations and Overgenerations: [Here, you would check for typographic errors,
          contradictions, and unsupported or excessive information.]
    5. Final Response: <<<START>>> {example_1_r} <<<END>>>

---
Remember, first provide the reasoning steps and then the final response with hallucinations marked.

<Question> {input_q} </Question>

<LLM_Answer> {input_a} </LLM_Answer>

<LLM_Answer_in_tokens> {input_a_tokens} </LLM_Answer_in_tokens>

Response:
    1. Understand the Question:
```

S2 Prompt: V3 - Brief instruction with in-context learning + be sensitive

```
You are tasked with identifying and marking hallucinations in the following large language model (LLM) answers
      . A hallucination in this context refers to an answer that provides incorrect or fabricated information.
       Your goal is to review each LLM answer (provided in <LLM_Answer>[llm_answer]</LLM_Answer>) relative to
      the question (provided in <Question>[question]</Question>) and highlight any incorrect or unsupported
      parts of the response using **<h>** tags. If the answer is factually correct, return it without any
      highlighting.

For each example, provide only the response sentence (R) with the highlighted hallucinations if present. Do
      not provide explanations or commentary. For structured extraction  use the following format/tags for the
       response: <<<START>>>[final_response_with_hallucinations_marked]<<<END>>>

Important: Ensure that the text remains exactly the same length as the original text, don't change any amount
      of whitespace or newline characters. You should only add tags and not delete any characters. To this end
       a token list is provided for the LLM answer (provided in  <LLM_Answer_in_tokens>[
      LLM_Answer_in_token_list]</LLM_Answer_in_tokens>).

Note: Your should be extremely critical in identifying hallucinations in the LLM answers. This means any
      character span that has the slightest chance of being incorrect should be marked as a hallucination.


---

Example 1:
<Question> {example_1_q} </Question>
<LLM_Answer> {example_1_a} </LLM_Answer>
<LLM_Answer_in_tokens> {example_1_a_tokens} </LLM_Answer_in_tokens>
<<<START>>> {example_1_r} <<<END>>>

Example 2:
<Question> {example_2_q} </Question>
<LLM_Answer> {example_2_a} </LLM_Answer>
<LLM_Answer_in_tokens> {example_2_a_tokens} </LLM_Answer_in_tokens>
<<<START>>> {example_2_r} <<<END>>>

New Question and Answer:
<Question> {input_q} </Question>
<LLM_Answer> {input_a} </LLM_Answer>
<LLM_Answer_in_tokens> {input_a_tokens} </LLM_Answer_in_tokens>
```