SHROOM-CAP: Shared Task on Hallucinations and Related Observable Overgeneration Mistakes in Crosslingual Analyses of Publications

Aman Sinha🍄,🦙 **Federica Gamba**🍄,🦙 **Raúl Vázquez**🍄 **Timothee Mickus**🍄
Ahana Chattopadhyay🍄 **Laura Zanella**🍄 **Binesh Arakkal Remesh**🍄
Yash Kankanampati🍄 **Aryan Chandramania**🍄 **Rohit Agarwal**🍄

🦙These authors have equal contribution.

🍄Université de Lorraine, France; 🍄Charles University, Prague;
🍄University of Helsinki, Finland; 🍄Independent Researcher;
🍄Université Paris Nord; 🍄IIIT Hyderabad, India; 🍄UiT Tromso, Norway

Correspondence: {aman.sinha@univ-lorraine.fr, gamba@ufal.mff.cuni.cz}

Abstract

This paper presents an overview of the SHROOM-CAP Shared Task, which focuses on detecting hallucinations and over-generation errors in cross-lingual analyses of scientific publications. SHROOM-CAP covers nine languages: five high-resource (English, French, Hindi, Italian, and Spanish) and four low-resource (Bengali, Gujarati, Malayalam, and Telugu). The task frames hallucination detection as a binary classification problem, where participants must predict whether a given text contains factual inaccuracies and fluency mistakes. We received 1,571 submissions from 5 participating teams during the test phase over the nine languages. In the paper, we present an analysis of the evaluated systems to assess their performance on the hallucination detection task across languages. Our findings reveal a disparity in system performance between high-resource and low-resource languages. Furthermore, we observe that factuality and fluency tend to be closely aligned in high-resource languages, whereas this correlation is less evident in low-resource languages. Overall, SHROOM-CAP underlines that hallucination detection remains a challenging open problem, particularly in low-resource and domain-specific settings.

Helsinki-NLP/SHROOM-CAP
SHROOM-Series/SHROOM-CAP

1 Introduction

Large Language Models (LLMs) are capable of producing coherent, fluent, and contextually appropriate text across a wide range of domains and languages. However, despite their impressive fluency, they are prone to hallucinations, i.e., generating content that is not supported by the input, or factually incorrect (Ji et al., 2023). Understanding and mitigating such behavior has become a central challenge in the development and deployment of reliable multilingual language technologies. To advance research in this direction, we organized the SHROOM-CAP Shared Task on Hallucinations and Related Observable Overgeneration Mistakes in Crosslingual Analyses of Publications,[1] as part of the CHOMPS Workshop on Confabulation, Hallucinations and Overgeneration in Multilingual and Practical Settings collocated with IJCNLP-AACL 2025 in Mumbai, India.

Figure 1: The SHROOM-CAP logo.

SHROOM-CAP builds on the previous iterations of the series — SHROOM (Mickus et al., 2024) and Mu-SHROOM (Vazquez et al., 2025) — while introducing two key extensions. First, the task targets the scientific domain with ACL anthology publications, encouraging evaluation in a specialized and knowledge-intensive context while previous iterations focused on general domain. Second, previous iteration of shared tasks already addressed multilingual hallucination detection, SHROOM-CAP explores cross-lingual settings covering both high-resource languages (Class[2] 4 to 5) including

[1] https://helsinki-nlp.github.io/shroom/2025a.

[2] We utilize the language taxonomy defined by Joshi et al. (2020), available at https://microsoft.github.io/linguisticdiversity/assets/lang2tax.txt.

Proceedings of the 1st Workshop on Confabulation, Hallucinations and Overgeneration in Multilingual and Practical Settings (CHOMPS 2025), pages 70–80
December 23, 2025 ©2025 Association for Computational Linguistics

English, French, Hindi, Italian, and Spanish; and low-resource languages (ranging from Class 1 to 3) including Bengali, Gujarati, Malayalam, and Telugu, with particular emphasis on the Indic language family. This focus aims to shed light on how language resource availability and linguistic typology affect hallucination behavior and detection performance.

Through SHROOM-CAP, we seek to provide the community with a benchmark for evaluating hallucinations in crosslingual scientific text generation, to foster deeper understanding of the phenomenon as well as the development of methods for producing more factual, fluent, and trustworthy LLM outputs.

2 Related Works

Recent surveys (Ji et al., 2023; Huang et al., 2024) emphasize how hallucinations, i.e., fluent but not factually correct LLMs' output, threaten the reliability of Natural Language Generation (NLG) systems, particularly in knowledge-intensive domains such as scientific writing, where models frequently generate unsupported claims or fabricated citations (George and Stuhlmueller, 2023). Early studies on factuality evaluation proposed benchmarks based on entailment and question-answering proxies (Kryściński et al., 2019; Wang et al., 2020), while later work showed that ensuring factual accuracy often requires domain-specific grounding and evidence retrieval, especially in scientific contexts (Wadden et al., 2022a,b). More recent analyses have revealed that multilingual LLMs display cross-lingual factual inconsistencies, often relying on surface lexical overlap rather than semantically grounded representations (Qi et al., 2023). Together, these findings highlight the need for robust, multilingual benchmarks to study hallucinations beyond English and across diverse generation settings.

The SHROOM series of shared tasks represent major steps toward systematic hallucination evaluation. SHROOM (Mickus et al., 2024) introduced a structured framework and annotated dataset designed to categorize and detect hallucinations across three NLG tasks on monolingual (English) setting. Its successor, Mu-SHROOM (Vazquez et al., 2025), extended the investigation to 14 languages, framing hallucination detection as a span-labeling problem. Building on these efforts, the SHROOM-CAP shared task continues this research line by expanding the scope of analysis to scientific text generation, to promote the development of more reliable and globally robust NLG systems. Together, these shared tasks establish a coherent progression from monolingual to multilingual and domain-specific evaluation of hallucinations complementing earlier factuality benchmarks (Yasunaga et al., 2019; Wadden et al., 2022a) and advancing the field toward trustworthy, evidence-grounded language generation.

3 SHROOM-CAP: Task Definition

Unlike its previous iterations, the SHROOM-CAP shared task presents hallucination as a two-fold problem. The task requires participants to identify two types of errors in LLM-generated scientific texts:

- **Factual mistakes**: content that contains hallucinations i.e., factually incorrect, unsupported, or inconsistent with the source material.
- **Fluency mistakes**: errors affecting linguistic quality, including grammatical inaccuracies, awkward phrasing, or incoherent constructions.

Formally, each error type is addressed as a *binary classification* problem. For each instance, participants are provided with the LLM-generated scientific text in three representations: a string of output text, a list of tokens, and the corresponding token-level logits. Systems are required to predict whether the text contains (a) factual mistakes and (b) fluency errors. The task is conducted in a multilingual setting, with data covering multiple languages and generated by a variety of public-weight LLMs. This design facilitates evaluation across diverse model behaviors and supports systematic comparison of detection performance across languages, model architectures, and resource conditions.

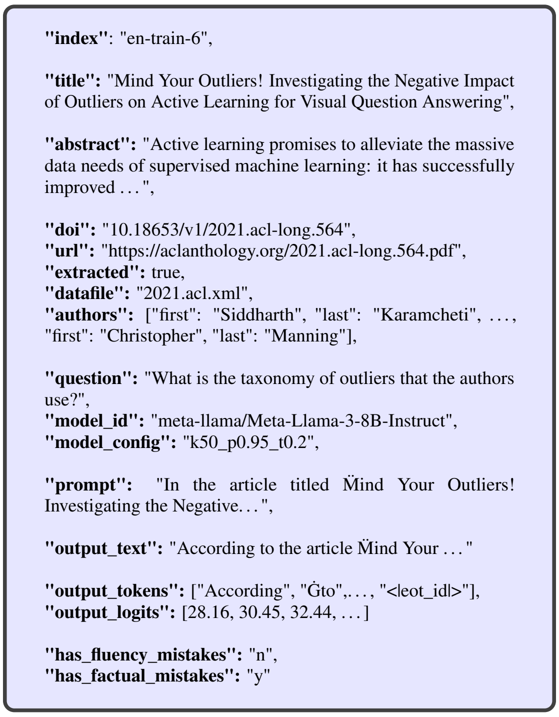
4 The CAP Dataset

We employed Gamba et al., 2025's CAP (Confabulations from ACL Publications) dataset, which is created to study hallucination in scientific text generation. The dataset spans nine languages: five high-resource languages including English, French, Hindi, Italian, and Spanish and four low-resource Indic languages including Bengali, Gujarati, Malayalam, and Telugu. For the high-

	Class	TRAIN	VAL	TEST	TOTAL
en	5	108	240	240	588
es	5	180	240	240	660
fr	5	520	240	240	1000
hi	4	425	240	240	905
it	4	520	240	240	1000
bn	3	/	/	798	798
gu	1	/	/	800	800
ml	1	/	/	788	788
te	1	/	/	800	800
	TOTAL	1233	1200	4386	6819

Table 1: Datapoints across data splits for each of the following languages: English (en), Spanish (es), French (fr), Hindi (hi), Italian (it), Bengali (bn), Gujarati (gu), Malayalam (ml), Telugu (te). *Class* denotes language taxonomy categories for each language proposed by (Joshi et al., 2020).The scale ranges from 1 to 5, with higher numbers indicating more resource-rich languages.

resource languages, to support system development, the dataset was divided into training, validation, and test sets, with training data released in two batches. For the low-resource languages, only test data was provided, and both the languages and their test sets were disclosed only at the start of the test phase, so participants had no prior knowledge of them. In total, each language contains about 100 unique question and 5-10 data points corresponding to each question. Table 1 summarizes the size of each data split, i.e., training, validation, and test sets across all languages. All unique questions are derived from scientific papers selected from a pool of 293 award-winning NLP papers extracted from the ACL Anthology.[3] For a random subset[4] of 100 papers per language, native or highly proficient annotators formulated questions addressing scientific aspects of each paper. These questions were paired with prompts following a standardized template — for example, in English: *"In the article by Author(s) (year), [question]"* — with analogous templates applied in other languages. Following this, 2-3 language specific LLMs per languages were utilized to generate scientific text as response to the curated questions. These LLM-generated answers were then annotated with binary factual and fluency mistakes. More detail on the annotation process can be found in Gamba et al. (2025).



"index": "en-train-6",

"title": "Mind Your Outliers! Investigating the Negative Impact of Outliers on Active Learning for Visual Question Answering",

"abstract": "Active learning promises to alleviate the massive data needs of supervised machine learning: it has successfully improved . . . ",

"doi": "10.18653/v1/2021.acl-long.564",
"url": "https://aclanthology.org/2021.acl-long.564.pdf",
"extracted": true,
"datafile": "2021.acl.xml",
"authors": ["first": "Siddharth", "last": "Karamcheti", . . . , "first": "Christopher", "last": "Manning"],

"question": "What is the taxonomy of outliers that the authors use?",
"model_id": "meta-llama/Meta-Llama-3-8B-Instruct",
"model_config": "k50_p0.95_t0.2",

"prompt": "In the article titled M̈ind Your Outliers! Investigating the Negative. . . ",

"output_text": "According to the article M̈ind Your . . . "

"output_tokens": ["According", "Ġto",. . . , "<|eot_id|>"],
"output_logits": [28.16, 30.45, 32.44, . . .]

"has_fluency_mistakes": "n",
"has_factual_mistakes": "y"

Figure 2: Example datapoint from English split of the CAP dataset (Gamba et al., 2025).

Figure 2 illustrates an example datapoint from the English split of the CAP dataset. Each datapoint in the dataset contains information on the model configuration, generated text, token sequences, logits, and language. Additionally, entries include available metadata such as titles, abstracts, DOIs, URLs, and author names. For the shared task, participants were provided with these dataset entries without the factuality and fluency annotations, which they were required to predict.

5 Evaluation

Metrics Participants systems are evaluated on the binary classification task described in Section 3 using the Macro-F1 score for two criteria: (i) factual errors and (ii) fluency errors.

Baselines We implemented two baseline systems to serve as terms of comparison:

1. RANDOM baseline, which assigns labels randomly for both factuality and fluency, providing a minimal-performance benchmark.
2. SELFCHECKGEMMA baseline inspired by SelfCheckGPT (Manakul et al., 2023) using the `google/gemma-2-9b` model. This halluci-

[3] https://aclanthology.org/.

[4] The set of papers annotated in one language does not necessarily overlap with those used in another. For any paper appearing in multiple languages, questions are not simple translations; rather, they are independently crafted by annotators in each language to reflect language-specific perspectives and nuances.

Team	Languages	Overview	N. Test Phase Submissions
CUET_GOODFELLAS	EN, ES, *rest**	zero-shot prompting with GPT-oss-20B	2
MEDUSA	EN	GPT-5 with RAG	3
NSU-AI	All	Attention mechanism anomalies, fine-tuned Qwen2.5 classifier	50
SCALAR_NITK	HI, *rest**	Retrieval-augmented classification and attention-based DL	2
SMURFCAT	All	Uncertainty estimation, encoder-based classifiers (BERT), decoder-based judges (instruction-tuned LLMs)	1514
AGI**	All*	XLM-RoBERTa-Large fine-tuned on additional training data	n/a

Table 2: Overview of participating teams (listed in alphabetical order). * denotes the languages participated during post-eval phase; ** implies new team submission during post-eval.

nation detection model is *reference-free*, i.e., it operates without external context. The model takes as input only the question-response pair with no context and evaluates its outputs for factuality and fluency, providing an automated assessment without human annotations.

6 Timeline

The shared task followed a four-phase schedule.

Starter Release On July 28, 2025, participants received the task description, data format, and sample data to facilitate early experimentation.

Training Phase Running from July 28 to October 5, 2025, this phase allowed teams to develop and fine-tune their models using the released training data, which was provided in two parts by the organizers: Train-v1 (40%) and Train-v2 (60%).

Testing Phase Held from October 5 to October 16, 2025, participants were asked to submit generated predictions on the hidden test set, which included five seen languages and four unseen surprise languages. Predictions were submitted for official evaluation and final leaderboard ranking.

Post-Evaluation Phase From October 16 to October 25, 2025, upon requests from several teams, the submission platform was reopened, allowing even registered teams[5] that did not participate in the test phase to conduct additional experiments if needed. This phase was also used for analyzing results, submitting system descriptions, and preparing final papers for inclusion in the shared task proceedings.

7 Participants' Systems

Six teams took part in the SHROOM-CAP shared task: five teams submitted a total of 1,571 submissions[6] throughout the test phase, and one participated in the post-evaluation phase. An overview of all teams is provided in Table 2, with detailed descriptions of their systems presented below.

NSU-AI implemented a two-fold framework for hallucination detection. First, a model-aware approach identifies fluency errors by analyzing attention patterns in the model, specifically high attention on the BOS token and low entropy of attention scores. The second, model-agnostic approach, which proved more accurate on average, uses a fine-tuned Qwen2.5 classifier (3B & 5B variant) to detect fluency errors semantic inconsistencies between the answer and the user query and factual errors semantic contradictions with the ground truth. This classifier leverages prompts that integrate the user question, model response, and the most relevant context chunks retrieved via `Alibaba-NLP/gte-multilingual-reranker-base`.

SMURFCAT built their system around the Qwen model, experimenting with different model sizes to optimize performance across languages. The proposed approach fine-tunes decoder-based LLMs (mainly Qwen-based) on translation-augmented training data with retrieved contexts using OpenAI's Vector Store. For comparison, they evaluated uncertainty-based, encoder-based, and proprietary model-based baselines.

CUET_GOODFELLAS relied on a zero-shot prompting approach with the GPT-oss-20B model. They did not incorporate any external data or additional training, leveraging exclusively the provided datasets.

[5] In total, we received 25 unique team registrations. However, not all registered teams took part in the test phase.

[6] We obtained a particularly high number of submissions from one team. See Table 2 for details.

	Language	en	es	fr	hi	it	ml	te	bn	gu
	N. Submissions	183	179	176	179	176	177	167	167	167
	N. Team	4	3	2	3	2	2	2	2	2
Factuality	MAX	0.92	0.76	0.86	0.84	0.87	0.65	0.72	0.69	0.64
	MEAN	0.59	0.55	0.65	0.50	0.69	0.50	0.51	0.37	0.47
	MIN	0.05	0.23	0.10	0.05	0.25	0.31	0.28	0.02	0.33
	Top Team	MEDUSA	NSU-AI	NSU-AI	NSU-AI	NSU-AI	NSU-AI	NSU-AI	NSU-AI	NSU-AI
Fluency	MAX	0.70	0.64	0.85	0.88	0.63	0.74	0.89	0.74	0.67
	MEAN	0.42	0.36	0.51	0.53	0.38	0.46	0.35	0.41	0.36
	MIN	0.15	0.13	0.29	0.19	0.13	0.31	0.06	0.21	0.16
	Top Team	NSU-AI	NSU-AI	NSU-AI	NSU-AI	NSU-AI	NSU-AI	NSU-AI	NSU-AI	NSU-AI

Table 3: Overview of final SHROOM-CAP leaderboard (Macro-F1 scores).

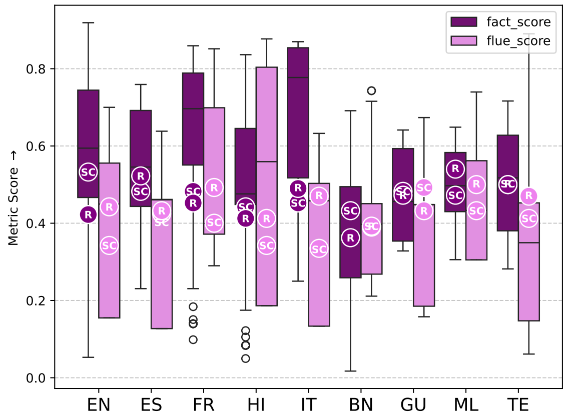


Figure 3: Distribution of systems for Factuality (`fact_score`) and Fluency mistake (`flue_score`) classification. The symbols SC and R are used to denote SelfCheckGemma and Random baselines corresponding to each language.

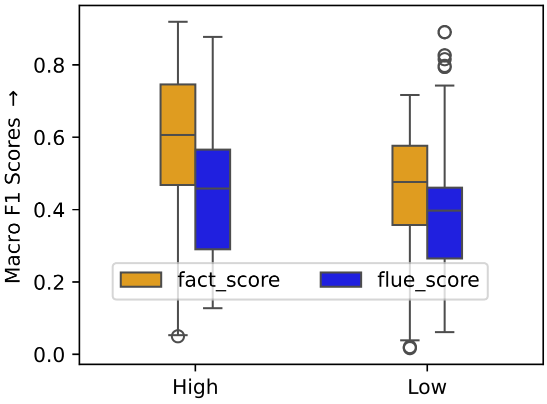


Figure 4: Performance comparison between systems based on high- versus low-resource languages.

MEDUSA experimented with multiple strategies, including the use of a synthetic dataset and a meta-model. Their best results, corresponding to the top submission for English, were achieved using GPT-5-mini with a RAG (Retrieval-Augmented Generation) approach, relying solely on the test data.

SCALAR NITK employed separate non-LLM based pipelines for factual and fluency error detection. For hallucination detection, a multi-step system retrieved relevant reference chunks, extracted similarity, NLI, BM25, and statistical features, and classified them using XGBoost with cross-validation and SMOTE. The fluency pipeline assessed readability through statistical, linguistic, and character features combined with multilingual embeddings, which were fused via an attention mechanism and classified using a multi-layer network with stratified folds and ensemble averaging.

8 Results and Discussion

Table 3 summarizes the final leaderboard, reporting the number of teams and submissions per language, along with the maximum, minimum, and mean Macro-F1 scores for all teams and for the top performer in each language for both Factuality and Fluency during the test phase.[7] Each language had at least two participating teams, with English showing the highest participation (four teams). In the post-evaluation phase, participation increased to at least five teams per language, with English remaining the most represented (six teams).

Tables 4 and 5 in Appendix B present the final rankings with detailed Macro-F1 scores for each team in the test phase, for Factuality and Fluency respectively. Results also include two baselines—Random and SelfCheck-

[7] Results exclude submissions made after the official test-phase deadline.

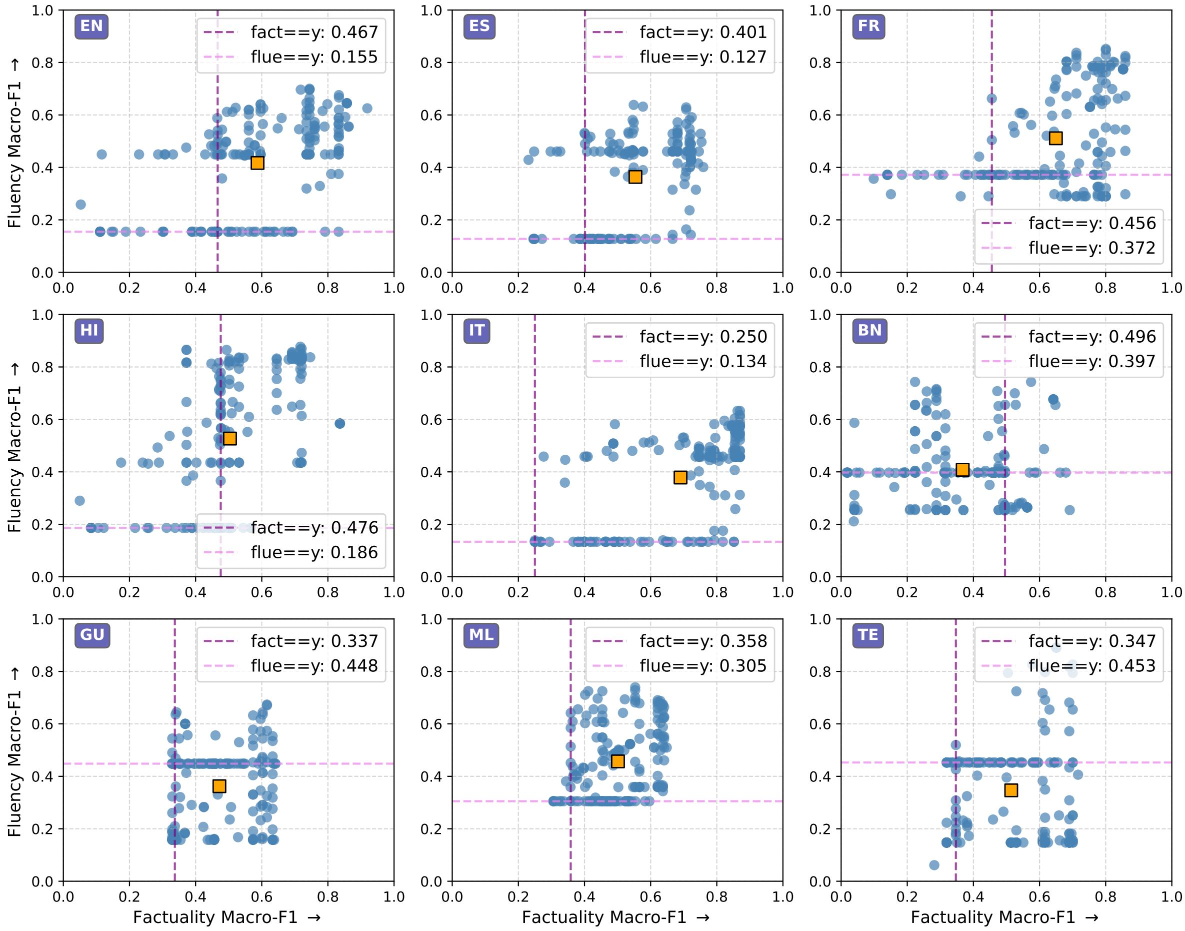


Figure 5: Interaction between fluency and factuality performance across all languages. The vertical and horizontal dashed lines depict system that predict all samples to contain fluency and factuality mistakes. Orange □ marker denotes the mean submission system for each language.

Gemma (Section 5)—for comparison. During the test phase, only two teams (NSU-AI and SMURFCAT) submitted predictions for all nine languages; CEUT_GOODFELLAS submitted for two languages, and MEDUSA and SCALAR_NITK for one language each.

Overall comparison. In Figure 3, we observe that high-resource languages achieved the highest Macro-F1 scores, ranging from 0.76 to 0.92, while submitted systems struggled on low-resource languages, with scores between 0.64 and 0.72 for factuality mistake classification. For fluency mistakes, most languages — except French (FR), Hindi (HI), and Telugu (TE) — showed lower scores, ranging from 0.64 to 0.74 Macro-F1. Overall, all systems outperformed both the random and Self-CheckGemma baselines.

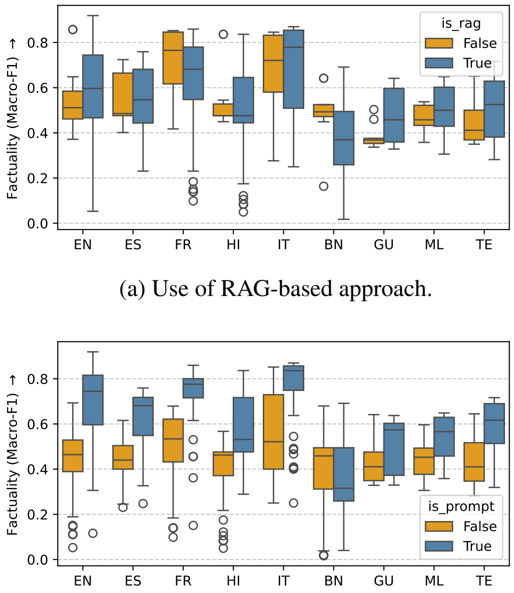
High vs. Low Resource Performance. Figure 4 compares the performance of all submitted systems across high- and low-resource language groups. We observe that models perform better in both factuality and fluency for high-resource languages. To obtain statistical validation of these differences, we conducted Mann–Whitney U tests comparing performance across the two groups. For factuality, the test revealed a significant difference ($U = 433{,}248$, $p < 0.001$) with a medium-to-large effect size ($r = 0.431$). For fluency, the difference was also significant ($U = 346{,}760$, $p < 0.001$), but the effect size was small ($r = 0.145$), indicating a modest gap between the groups. These results suggest that both hallucination detection and fluency verification are more challenging for low-resource languages.

Trade off between Factuality and Fluency Performance. In Figure 5, across languages, the relationship between factuality and fluency in hallucination detection reveals notable variation between high- and low-resource settings. In high-resource languages, factuality and fluency exhibit a moderate positive interaction—systems that produce more fluent text also tend to be more factually accurate, though this alignment is not perfect. English and French demonstrate the most balanced performance, while Italian and Hindi show greater dispersion, indicating less stability across systems. The average submission (denoted by orange square marker) further clarifies these trends, showing that high-resource languages cluster toward higher factuality and fluency regions, reflecting models that perform both accurately and coherently on average. In contrast, low-resource languages display weaker correlations between fluency and factuality, with mean markers positioned toward lower factuality but moderate fluency. This confirms that in resource-scarce settings, models often generate fluent yet factually inconsistent outputs, making fluency a poor proxy for factual reliability. Overall, these findings highlight a consistent performance gap between high- and low-resource languages, where fluency and factuality tend to co-improve with greater resource availability but diverge in low-resource contexts, underscoring the need for tailored approaches to hallucination mitigation across languages.

Diversity in Approaches vs. Performance. With the collected metadata,[8] we observe that out of 1,571 submissions during test phase, ∼94% were RAG-based models. Furthermore, around ∼58% of the submitted systems used a prompt-based approach, with the remainder leveraging fine-tuning-based models.

Figure 6a shows the impact of using a RAG-based approach on factuality performance. While RAG generally leads to strong results, we observe three exceptions — French, Hindi, and Bengali — where RAG-based systems do not outperform non-RAG approaches.

Figure 6b presents the results of our analysis, comparing systems submitted during the test phase based on the use of prompt-based methods versus fine-tuning with respect to factuality performance. Overall, systems employing fine-tuning outperform prompt-based approaches across most datasets, with the exception of Bengali.



(a) Use of RAG-based approach.

(b) Use of prompt-based approach.

Figure 6: Different approaches vs. performance on factuality.

Overall, these results suggest that while fine-tuning and RAG-based architectures generally improve hallucination detection performances, their benefits vary across languages, highlighting the need for language-aware strategies rather than one-size-fits-all approaches.

9 Conclusions

In this paper, we presented an overview of the SHROOM-CAP shared task on Hallucinations and Related Observable Overgeneration Mistakes in Crosslingual Analyses of Publications, organized as part of the First Workshop on Confabulation, Hallucinations, and Overgeneration in Multilingual and Practical Settings (CHOMPS). The shared task focused on detecting hallucinations in scientific texts generated by LLMs, with particular attention to evaluating both factuality and fluency.

The task leveraged the CAP dataset, which comprises nine languages — five high-resource languages and four low-resource Indic languages. During the test phase, we received a total of 1,571 submissions from five participating teams. Most systems employed RAG-based approaches, with roughly equal proportions further incorporating prompting or fine-tuning strategies.

Our analysis revealed a clear distinction in hal-

[8] The metadata was self-reported by participants and may contain minor inconsistencies or inaccuracies.

lucination detection performance between high-resource and low-resource languages. Notably, Bengali emerged as a particularly challenging case, where neither prompting-based nor RAG-based systems achieved substantial improvements.

Despite the encouraging progress observed in recent years, the results highlight that hallucination detection remains a challenging open problem, particularly in low-resource and domain-specific contexts. As the use of LLMs continues to expand, developing robust and generalizable methods for identifying and mitigating hallucinations will be essential to ensuring the reliability and factual integrity of generated content.

While the shared task provided valuable insights, the predominance of submissions from a single team highlights an opportunity to improve the diversity of participation. Given the relatively short preparation timeline, this outcome is understandable; however, broader engagement is essential to strengthen the robustness and generalizability of future findings. To this end, future editions will adopt a longer timeline and include more targeted outreach, particularly toward students and early-career researchers, to encourage wider participation and enhance the overall impact of the shared task.

Acknowledgment

The organization of SHROOM-CAP shared task was supported by the Research Council of Finland through project №353164. "Green NLP – controlling the carbon footprint in sustainable language technology". The work described herein has been using resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure, supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062). FG was partially funded by Charles University Grant UNCE24/SSH/009.

References

Federica Gamba, Aman Sinha, Timothee Mickus, Raul Vazquez, Patanjali Bhamidipati, Claudio Savelli, Ahana Chattopadhyay, Laura A. Zanella, Yash Kankanampati, Binesh Arakkal Remesh, Aryan Ashok Chandramania, Rohit Agarwal, Chuyuan Li, Ioana Buhnila, and Radhika Mamidi. 2025. Confabulations from acl publications (cap): A dataset for scientific hallucination detection. *Preprint*, arXiv:2510.22395.

Charlie George and Andreas Stuhlmueller. 2023. Factored verification: Detecting and reducing hallucination in summaries of academic papers. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 107–116, Bali, Indonesia. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

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

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

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

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.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. *arXiv preprint arXiv:2310.10378*.

Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez 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. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2472–2497, Vienna, Austria. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022a. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022b. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7386–7393.

A SHROOM-CAP Organizers' role

The team of CHOMPS contributors behind this edition of the SHROOM Shared Task is as follows:

Aman Sinha: Conceptualization, overall leadership, advertisement, paper writing, Hindi and English question creation & annotation, annotator recruiting, annotation guidelines, annotator training, baseline preparation, data analysis.

Federica Gamba: Overall leadership, advertisement, paper writing, Italian question creation & annotation, annotator training, answer generation, baseline preparation.

Raúl Vázquez: Conceptualization, Spanish question and data creation, preparation of annotation workflow scripts, advertisement, overall leadership, reviewing process.

Timothee Mickus: Conceptualization, submission platform development, French and English question creation, data analysis, advertisement, overall leadership, reviewing process.

Ahana Chattopadhyay: Bengali and English question creation & annotation.

Laura Zanella: French and Spanish question creation & annotation.

Binesh Arakkal Remesh: Malayalam question creation & annotation.

Yash Kankanampati: Telugu question creation & annotation.

Aryan Chandramania: Gujarati question creation & annotation.

Rohit Agarwal: Hindi question creation & annotation.

B Final Test Phase Leaderboard

Language	Rank	Team	Factuality
EN	1	MEDUSA	0.919
EN	2	SMURFCAT	0.863
EN	3	CUET-GOODFELLAS	0.648
EN	4	baseline (SelfCheck)	0.527
EN	5	NSU-AI	0.512
EN	6	baseline (Random)	0.415
ES	1	SMURFCAT	0.759
ES	2	CUET-GOODFELLAS	0.724
ES	3	NSU-AI	0.535
ES	4	baseline (Random)	0.515
ES	5	baseline(SelfCheck)	0.483
FR	1	SMURFCAT	0.860
FR	2	NSU-AI	0.661
FR	3	baseline (SelfCheck)	0.482
FR	4	baseline (Random)	0.447
HI	1	SMURFCAT	0.836
HI	2	SCALAR_NITK	0.545
HI	3	NSU-AI	0.477
HI	4	baseline (SelfCheck)	0.440
HI	5	baseline (Random)	0.412
IT	1	SMURFCAT	0.870
IT	2	NSU-AI	0.742
IT	3	baseline (Random)	0.486
IT	4	baseline (SelfCheck)	0.453
BN	1	SMURFCAT	0.691
BN	2	NSU-AI	0.525
BN	3	baseline (SelfCheck)	0.432
BN	4	baseline (Random)	0.365
GU	1	SMURFCAT	0.641
GU	2	NSU-AI	0.503
GU	3	baseline (SelfCheck)	0.480
GU	4	baseline (Random)	0.475
ML	1	SMURFCAT	0.649
ML	2	baseline (Random)	0.543
ML	3	NSU-AI	0.522
ML	4	baseline (SelfCheck)	0.465
TE	1	SMURFCAT	0.716
TE	2	baseline (Random)	0.501
TE	3	baseline (SelfCheck)	0.500
TE	4	NSU-AI	0.500

Table 4: Official test leaderboard, all languages, all teams for Factuality.

Language	Rank	Team	Fluency
EN	1	SMURFCAT	0.700
EN	2	MEDUSA	0.625
EN	3	NSU-AI	0.6118
EN	4	CUET-GOODFELLAS	0.549
EN	5	baseline (Random)	0.441
EN	6	baseline (SelfCheck)	0.342
ES	1	SMURFCAT	0.638
ES	2	CUET-GOODFELLAS	0.591
ES	3	NSU-AI	0.528
ES	4	baseline (Random)	0.431
ES	5	baseline (SelfCheck)	0.397
FR	1	SMURFCAT	0.852
FR	2	NSU-AI	0.521
FR	3	baseline (Random)	0.487
FR	4	baseline (SelfCheck)	0.401
HI	1	SMURFCAT	0.877
HI	2	SCALAR_NITK	0.835
HI	3	NSU-AI	0.754
HI	4	baseline (Random)	0.412
HI	5	baseline (SelfCheck)	0.333
IT	1	SMURFCAT	0.633
IT	2	NSU-AI	0.502
IT	3	baseline (Random)	0.466
IT	4	baseline (SelfCheck)	0.334
BN	1	SMURFCAT	0.743
BN	2	NSU-AI	0.708
BN	3	baseline (Random)	0.485
BN	4	baseline (SelfCheck)	0.389
GU	1	SMURFCAT	0.674
GU	2	NSU-AI	0.557
GU	3	baseline (SelfCheck)	0.494
GU	4	baseline (Random)	0.432
ML	1	SMURFCAT	0.740
ML	2	NSU-AI	0.696
ML	3	baseline (Random)	0.498
ML	4	baseline (SelfCheck)	0.430
TE	1	SMURFCAT	0.891
TE	2	baseline (Random)	0.466
TE	3	baseline (SelfCheck)	0.409
TE	4	NSU-AI	0.403

Table 5: Official test leaderboard, all languages, all teams for Fluency.

C Post-Evaluation Leaderboard

Language	Rank	Team	Factuality
EN	1	MEDUSA	0.9191
EN	2	SMURFCAT	0.8627
EN	3	CUET-GOODFELLAS	0.6483
EN	4	AGI	0.5999
EN	5	NSU-AI	0.5333
EN	6	baseline (SelfCheck)	0.5266
EN	7	SCALAR_NITK	0.4667
EN	8	baseline (random)	0.4154
ES	1	SMURFCAT	0.7876
ES	2	CUET-GOODFELLAS	0.7243
ES	3	NSU-AI	0.5354
ES	4	baseline (random)	0.5153
ES	5	AGI	0.4938
ES	6	baseline (SelfCheck)	0.4825
ES	7	SCALAR_NITK	0.4811
FR	1	SMURFCAT	0.8595
FR	2	CUET-GOODFELLAS	0.7769
FR	3	NSU-AI	0.6612
FR	4	SCALAR_NITK	0.5524
FR	5	AGI	0.5401
FR	6	baseline (SelfCheck)	0.4819
FR	7	baseline (random)	0.4468
HI	1	SMURFCAT	0.8364
HI	2	CUET-GOODFELLAS	0.7898
HI	3	SCALAR_NITK	0.6153
HI	4	AGI	0.5344
HI	5	NSU-AI	0.5051
HI	6	baseline (SelfCheck)	0.4401
HI	7	baseline (random)	0.4120
IT	1	SMURFCAT	0.8720
IT	2	NSU-AI	0.8174
IT	3	AGI	0.6234
IT	4	SCALAR_NITK	0.5867
IT	5	CUET-GOODFELLAS	0.5391
IT	6	baseline (random)	0.4861
IT	7	baseline (SelfCheck)	0.4533
BN	1	SMURFCAT	0.7035
BN	2	NSU-AI	0.6546
BN	3	CUET-GOODFELLAS	0.5998
BN	4	AGI	0.5652
BN	5	SCALAR_NITK	0.4933
BN	6	baseline (SelfCheck)	0.4320
BN	7	baseline (random)	0.3645
GU	1	SMURFCAT	0.6413
GU	2	AGI	0.5107
GU	3	NSU-AI	0.5032
GU	4	baseline (SelfCheck)	0.4796
GU	5	baseline (random)	0.4749
GU	6	CUET-GOODFELLAS	0.3852
GU	7	SCALAR_NITK	0.3560
ML	1	SMURFCAT	0.6487
ML	2	CUET-GOODFELLAS	0.5463
ML	3	baseline (random)	0.5428
ML	4	NSU-AI	0.5220
ML	5	AGI	0.4857
ML	6	baseline (SelfCheck)	0.4653
ML	7	SCALAR_NITK	0.3650
TE	1	SMURFCAT	0.7164
TE	2	CUET-GOODFELLAS	0.5704
TE	3	baseline (random)	0.5012
TE	4	baseline (SelfCheck)	0.4999
TE	5	NSU-AI	0.5004
TE	6	AGI	0.4738
TE	7	SCALAR_NITK	0.3529

(Continued from previous page)

Table 6: Official post-evaluation rankings, all languages, all teams for Factuality.

Language	Rank	Team	Fluency
EN	1	NSU-AI	0.627
EN	2	MEDUSA	0.625
EN	3	SMURFCAT	0.556
EN	4	CUET-GOODFELLAS	0.549
EN	5	AGI	0.450
EN	6	SCALAR_NITK	0.450
EN	7	baseline (random)	0.441
EN	8	baseline (SelfCheck)	0.342
ES	1	CUET-GOODFELLAS	0.591
ES	2	SMURFCAT	0.461
ES	3	AGI	0.461
ES	4	SCALAR_NITK	0.461
ES	5	NSU-AI	0.446
ES	6	baseline (random)	0.431
ES	7	baseline (SelfCheck)	0.397
FR	1	SMURFCAT	0.825
FR	2	SCALAR_NITK	0.644
FR	3	baseline (random)	0.487
FR	4	CUET-GOODFELLAS	0.473
FR	5	NSU-AI	0.407
FR	6	baseline (SelfCheck)	0.401
FR	7	AGI	0.290
HI	1	SCALAR_NITK	0.835
HI	2	CUET-GOODFELLAS	0.723
HI	3	NSU-AI	0.695
HI	4	SMURFCAT	0.584
HI	5	baseline (random)	0.412
HI	6	baseline (SelfCheck)	0.333
HI	7	AGI	0.239
IT	1	NSU-AI	0.586
IT	2	CUET-GOODFELLAS	0.546
IT	3	SCALAR_NITK	0.544
IT	4	baseline (random)	0.466
IT	5	SMURFCAT	0.458
IT	6	AGI	0.458
IT	7	baseline (SelfCheck)	0.334
BN	1	CUET-GOODFELLAS	0.550
BN	2	SCALAR_NITK	0.518
BN	3	baseline (random)	0.485
BN	4	SMURFCAT	0.447
BN	5	NSU-AI	0.405
BN	6	baseline (SelfCheck)	0.389
BN	7	AGI	0.254
GU	1	CUET-GOODFELLAS	0.610
GU	2	baseline (SelfCheck)	0.494
GU	3	SMURFCAT	0.448
GU	4	baseline (random)	0.432
GU	5	SCALAR_NITK	0.306
GU	6	NSU-AI	0.235
GU	7	AGI	0.158
ML	1	NSU-AI	0.694
ML	2	CUET-GOODFELLAS	0.637
ML	3	SCALAR_NITK	0.521
ML	4	SMURFCAT	0.510
ML	5	baseline (random)	0.498
ML	6	baseline (SelfCheck)	0.430
ML	7	AGI	0.245
TE	1	CUET-GOODFELLAS	0.716
TE	2	baseline (random)	0.466
TE	3	SCALAR_NITK	0.460
TE	4	baseline (SelfCheck)	0.409
TE	5	SMURFCAT	0.406
TE	6	NSU-AI	0.396
TE	7	AGI	0.147

(*Continued from previous page*)

Table 7: Official post-evaluation rankings, all languages, all teams for Fluency.