SmurfCat at SHROOM-CAP: Factual but Awkward? Fluent but Wrong? Tackling Both in LLM Scientific QA

Timur Ionov[3,5] **Evgenii Nikolaev**[5] **Artem Vazhentsev**[1,2]
Mikhail Chaichuk[1,4] **Anton Korznikov**[1,2,4] **Elena Tutubalina**[1,7]
Alexander Panchenko[2,1] **Vasily Konovalov**[1,2,6] **Elisei Rykov**[2]
[1]AIRI [2]Skoltech [3]MWS AI [4]HSE University
[5]AI Talent Hub, ITMO University, Saint Petersburg, Russia
[6]Moscow Independent Research Institute of Artificial Intelligence
[7]Kazan Federal University
t.ionov@mts.ai elisei.rykov@skol.tech

Abstract

Large Language Models (LLMs) often generate hallucinations, a critical issue in domains like scientific communication where factual accuracy and fluency are essential. The SHROOM-CAP shared task addresses this challenge by evaluating Factual Mistakes and Fluency Mistakes across diverse languages, extending earlier SHROOM editions to the scientific domain. We present Smurfcat, our system for SHROOM-CAP, which integrates three complementary approaches: uncertainty estimation (white-box and black-box signals), encoder-based classifiers (Multilingual Modern BERT), and decoder-based judges (instruction-tuned LLMs with classification heads). Results show that decoder-based judges achieve the strongest overall performance, while uncertainty methods and encoders provide complementary strengths. Our findings highlight the value of combining uncertainty signals with encoder and decoder architectures for robust, multilingual detection of hallucinations and related errors in scientific publications.

1 Introduction

Large Language Models (LLMs) have achieved remarkable success across a wide range of natural language processing (NLP) tasks. However, their tendency to produce hallucinations-outputs containing factually unsupported, unverifiable, or fabricated information-remains a critical barrier to their safe deployment in real-world applications. The risks posed by hallucinations are particularly severe in domains where factual precision is essential, such as scientific communication, healthcare, and legal contexts. Moreover, multilingual and cross-lingual scenarios exacerbate these challenges, as disparities in linguistic resources hinder the development and evaluation of robust factuality assessment systems.

To systematically address these concerns, the SHROOM (Shared-task on Hallucinations and Related Observable Overgeneration Mistakes) series has emerged as the first dedicated benchmark initiative for hallucination detection and mitigation. The inaugural SHROOM 2024 (Mickus et al., 2024) established a foundation by creating multilingual benchmarks and evaluation protocols for hallucination detection in LLMs, with a focus on relatively controlled, general-purpose text settings. Building on this, Mu-SHROOM 2025 (Vazquez et al., 2025) expanded both the scale and scope, introducing broader evaluation methodologies and more linguistically diverse datasets, pushing the community toward developing cross-lingual methods for hallucination analysis.

However, both of these earlier shared tasks-despite their significant contributions–did not fully capture the unique demands of scientific communication. In scientific publications, hallucinations are not merely stylistic or semantic errors but can result in fabricated citations, unsupported claims, or distortions of technical content. Such errors undermine trust and reproducibility, yet existing SHROOM tasks did not explicitly evaluate models in these high-stakes, domain-specific contexts. Furthermore, while multilinguality was central to the earlier SHROOM editions, the emphasis remained on relatively high-resource languages, leaving persistent gaps in evaluating hallucinations in low-resource languages where scientific material is scarce and ground truth is more difficult to establish.

To address these shortcomings, SHROOM-CAP (Shared-task on Hallucinations and Related Observable Overgeneration Mistakes in Crosslingual Analyses of Publications)[1] (Sinha et al., 2025; Gamba et al., 2025) was introduced as the third installment in the SHROOM series. SHROOM-CAP specifically targets the domain of scientific publications and extends the challenge to both high-resource

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

Proceedings of the 1st Workshop on Confabulation, Hallucinations and Overgeneration in Multilingual and Practical Settings (CHOMPS 2025), pages 81–89

and low-resource languages. In addition to hallucinations, SHROOM-CAP introduces a dual focus on evaluating Factual Mistakes (e.g., unsupported claims, fabricated references, and misleading scientific assertions) and Fluency Mistakes (e.g., grammatical errors, disfluencies, and unnatural style that hinders scientific readability). Participants are tasked with detecting and analyzing these errors in LLM outputs conditioned on scientific input material, bridging the methodological advances of prior SHROOM editions with the real-world demands of multilingual scientific communication.

By providing a unified benchmark for hallucination detection in scientific publishing-augmented with explicit evaluation of both factual and fluency mistakes-SHROOM-CAP aims to catalyze research into reliable evaluation metrics and practical mitigation strategies. It places special emphasis on low-resource and linguistically diverse scenarios, thereby encouraging the development of more inclusive, transparent, and trustworthy language technologies. In doing so, SHROOM-CAP not only continues the trajectory established by previous SHROOM competitions, but also addresses critical gaps that remain at the intersection of factuality, fluency, multilingualism, and domain specificity.

2 Related Work

2.1 Factual Mistakes

Hallucinations in scientific discourse. Within the scientific domain, prior work frames factuality as claim verification and reference reliability. Early efforts such as SciFact (Wadden et al., 2020) study whether research claims are supported by evidence from the literature, establishing a foundation for evidence-grounded evaluation over scholarly text and inspiring later open-domain variants; this line underlines the need to ground generations in primary sources when judging factuality in publications.

Uncertainty estimation (UQ) for factuality. Model-centric UQ signals are widely leveraged to detect hallucinations without heavy supervision inluding both white-box and black-box UQ families: probability/entropy-based measures (Sequence Probability, Perplexity, Mean Token Entropy), CCP (Fadeeva et al., 2024) calibration, and RAUQ (Vazhentsev et al., 2025) (uncertainty-aware attention) on the white-box side. In addition, UQ-based methods that increase the faithfulness of generation have been widely used in many applications, including adaptive RAG (Moskvoretskii et al., 2025; Marina et al., 2025) and the development of QA systems across various domains (Aushev et al., 2025; Belikova et al., 2024).

The black-box methods methods provide sequence-level scores that correlate with factual errors amnog them should be mentioned Semantic Entropy (Kuhn et al., 2023), SAR (Duan et al., 2024), KLE, Semantic Density, CoCoA (Vashurin et al., 2025b). The combination of white-box and back-box methods was effective in detecting span-level hallucination in SHROOM-2025 (Rykov et al., 2025a).

Encoder classifiers. Encoder-based models remain a strong baseline for factuality judgments when inputs can be structured. In our setup, a multilingual BERT-family encoder (mmBERT-base[2]) receives concatenated question–answer–context sequences and is fine-tuned with weighted loss for class imbalance; per-language thresholding and macro-F1 selection improve robustness across high- and low-resource languages (Rykov et al., 2025b).

2.2 Fluency Mistakes

Fluency mistakes-grammatical ill-formedness, disfluencies, awkward phrasing, and incoherent structure-degrade readability and can obscure factual content, especially in multilingual scientific writing. SHROOM-CAP evaluates fluency separately from factuality, mirroring editorial practice in scholarly communication.

Instruction-tuned decoder LLMs can be repurposed as fluency judges by prompting them to ignore factuality and return compact decisions (e.g., y/n) (Gu et al., 2024).

Grammatical Error Correction (GEC) pipelines-sequence-to-sequence correctors and grammaticality classifiers (e.g., CoLA-style)-remain complementary: they can produce silver labels for fluency supervision and serve as automatic critics (Qorib et al., 2024).

3 Data

The dataset comprises a total of 7,078 examples, initially split into 1,752 for training, 1,200 for validation, and 4,126 for testing.These examples cover 9 languages: English (EN), Spanish (ES), French (FR), Hindi (HU), Italian (IT), Bengali (BN), Gujarati (GU), Malayalam (ML), and Telugu (TE).

[2] https://hf.co/jhu-clsp/mmBERT-base

Five of these languages (EN, ES, FR, HI, IT) are present in the training and validation sets. The remaining four languages (BN, GU, ML, TE) are exclusively available in the test set, facilitating evaluation in a zero-shot cross-lingual setting.

Each instance in the dataset is represented by the following fields: abstract, link, model_id, model_config, question, prompt, output_text, output_tokens, output_logits.

Furthermore, examples in the training and validation splits are annotated with two binary labels: `has_fluency_mistakes` and `has_factual_mistakes`.

3.1 Retrieval

To augment the data with relevant context from the parsed papers, we used OpenAI's Vector Store[3]. First, we downloaded all PDF files mentioned in the dataset and uploaded them to the Vector Store. Next, to retrieve passages, we performed a search requests to the Vector Store using the question from the dataset. Since each question is followed by the corresponding PDF file, we applied a filter to search for relevant passages within the file, instead searching the entire Vector Store collection.

3.2 Translations

Additionally, we utilized the Yandex Translate API to translate questions and answers into other languages. As a result of this translation, 8,735 examples were added to the training set. The full language distribution of training data is shown in Figure 1.

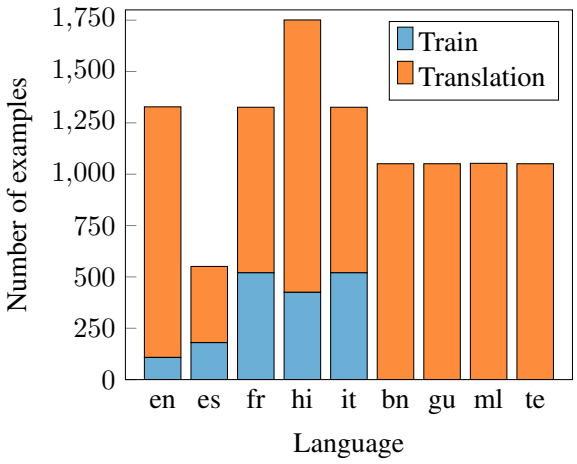

Figure 1: Training data with translation augmentation distribution.

4 Methods

4.1 Baseline

As a baseline, we report the performance of GPT-5 on the test set. As in all subsequent cases, we used contexts retrieved via OpenAI's Vector Store with a specific prompt that asks GPT-5 to analyze an input question, relevant context, paper abstract, and LLM answer, and then identify any factual or fluency errors in the answer. The prompt is shown in Figure 2.

4.2 Uncertainty Quantification

Uncertainty quantification (UQ) (Gal and Ghahramani, 2016; Baan et al., 2023) is a prominent approach for hallucination detection and low-quality output (Malinin and Gales, 2021; Farquhar et al., 2024), particularly in sequence-level tasks, which represent the most standard and suitable settings for UQ (Vashurin et al., 2025a). We consider a variety of state-of-the-art methods from both white-box and black-box categories (Fadeeva et al., 2023).

For the white-box methods, we employ probability-based approaches such as Sequence Probability, Perplexity, Mean Token Entropy (Fomicheva et al., 2020), CCP (Fadeeva et al., 2024), and RAUQ (Vazhentsev et al., 2025). These methods analyze the predicted token-level probability distributions to produce a single sequence-level uncertainty score. Notably, RAUQ combines token probabilities with attention weights from specific "uncertainty-aware" attention heads of the LLM.

We also include sampling-based white-box methods such as Semantic Entropy (Kuhn et al., 2023), SAR (Duan et al., 2024), KLE (Nikitin et al., 2024), Semantic Density (Qiu and Miikkulainen, 2024), and CoCoA (Vashurin et al., 2025b). These techniques assess the diversity among multiple answers generated by an LLM for the same input using an auxiliary Natural Language Inference (NLI) model. Semantic Entropy clusters responses into distinct groups and computes the entropy of the cluster probabilities. SAR, KLE, and Semantic Density reweight sequence probabilities in various ways, while CoCoA simplifies this concept by combining diversity and probability scores multiplicatively.

For the black-box methods, we include Lexical Similarity (Fomicheva et al., 2020), DegMat and Eccentricity (Lin et al., 2024), and LUQ (Zhang et al., 2024). DegMat and Eccentricity model the

[3] https://platform.openai.com/docs/api-reference/vector-stores

set of predictions as a weighted adjacency matrix of a graph to analyze their diversity. Lexical Similarity measures diversity through n-gram similarity scores, whereas LUQ evaluates long-form generation consistency using an NLI model.

4.3 Encoder

We use a multilingual BERT-based encoder approach for binary classification of factual mistakes. Our implementation uses **mmBERT-base** (Marone et al., 2025), which provides strong multilingual capabilities across the different languages in the SHROOM-CAP dataset. Each training example is formatted as a structured sequence that combines question, answer, and context information. We use the template "`[Q] <question>\n[A] <answer>\n[C] <context>`" to help the model understand the relationship between the generated answer and the supporting context.

We fine-tune the model for factual mistake detection using binary classification. We apply weighted binary cross-entropy loss to handle the imbalanced dataset. The `[CLS]` token representation is passed through a classification head to predict the target label. We fine-tune our encoder model on both the original training dataset and the augmented training data that includes translations to increase data diversity. Model selection is based on macro F1-score on the validation set. We also implement per-language threshold optimization to maximize performance for each target language.

4.4 Decoder

We fine-tune large decoder-based language models in a binary classification setup. We leverage 4 different decoders: `Qwen3-Reranker-8B`[4], `Qwen3-14B`[5], `Qwen3-32B`[6], `Qwen3-30A3B`[7], and `sarvamai/sarvam-1`[8], optimized for Indic languages (Bengali, Hindi, Tamil, Telugu, etc.).

For Decoder-based approach, we format each sample as a structured dialog to align with the common decoder instruction-followed format. As inputs, we pass the retrieved context, the original question, and the LLM's answer. To perform classification, we add two MLP heads. For evaluation, per-language thresholds are optimized on the validation set to maximize Macro F1.

[4] https://hf.co/Qwen/Qwen3-Reranker-8B
[5] https://hf.co/Qwen/Qwen3-14B
[6] https://hf.co/Qwen/Qwen3-32B
[7] https://hf.co/Qwen/Qwen3-30B-A3B
[8] https://hf.co/sarvamai/sarvam-1

5 Results

Table 1 shows the overall performance of our methods compared to other top-performing teams at the SHROOM-CAP. The Decoder-based approach is the clear winner in both factuality and fluency metrics, performing well in English and Hindi for factuality and in Telugu for fluency. Although the decoder-based model has a gap in factuality for the English language in the macro F1 score, it demonstrates strong multilingual capabilities.

GPT-5 achieved top results in factuality for Bengali, Spanish, French, and Telugu, as well as the top result for Hindi. In terms of fluency, GPT-5 performed well, achieving the second-best score in English and Hindi and the best score in Telugu. However, it lags behind other teams' approaches in other languages.

Table 2 shows the ablation of the Encoder-based approach. Adding translations significantly improved scores for English, Spanish, Guam, Hindi, Italian, Malayalam, and Telugu, and decreased for French and Bengali. However, compared to other approaches, Encoder-based method yields to other methods methods in most languages, excluding French and Italian for factuality, where Encoder-based method is the third best-performing approach.

Table 3 shows the results on fine-tuning different decoder-based LLMs on SHROOM-CAP train data. Across all languages, the top performer is `Qwen3-32B`, demonstrating the best scores for Bengali, Spanish, French, Gujarati, and Hindi, as well as second-best performance for English, Italian, and Malayalam. Interestingly, `sarvam-1` shows competitive results for English in the factuality metric, while maintaining balanced performance across several other languages. The smaller `Qwen3-Reranker-8B` model also performs surprisingly well, especially in Hindi and Italian, indicating that reranker-style fine-tuning can be beneficial even with reduced model capacity. For fluency, `Qwen3-32B` and `Qwen3-30B-A3B-Instruct` yield the highest scores across most languages, confirming the correlation between model size and linguistic smoothness. Overall, these results suggest that large-scale Qwen3 models are the most effective backbone for multilingual hallucination detection in the decoder-based setup.

Table 6 presents the detailed results obtained using various uncertainty quantification methods. Although the performance of each method varies

Method	Mode	BN	EN	ES	FR	GU	HI	IT	ML	TE
factuality										
Decoder	FT	**0.69**	0.86	**0.75**	**0.86**	**0.82**	0.75	**0.87**	**0.64**	**0.72**
GPT-5	ZS	0.64	0.85	0.72	0.75	0.36	**0.83**	0.48	0.53	0.65
nsu-ai	-	0.52	0.51	0.53	0.66	0.50	0.47	0.74	0.52	0.50
CUET_Goodfellas	-	-	0.64	0.72	-	-	-	-	-	-
medusa	-	-	**0.91**	-	-	-	-	-	-	-
Uncertainty	-	0.5	0.6	0.58	0.56	0.54	0.65	0.71	0.55	0.57
Encoder	FT	0.49	0.57	0.5	0.67	0.45	0.51	0.8	0.5	0.44
fluency										
Decoder	FT	**0.74**	**0.7**	**0.64**	**0.85**	**0.67**	**0.88**	**0.63**	**0.74**	0.83
GPT-5	ZS	0.67	0.64	0.42	0.63	0.60	0.58	0.50	0.52	**0.89**
nsu-ai	-	0.70	0.61	0.52	0.52	0.55	0.75	0.59	0.69	0.40
CUET_Goodfellas	-	0.54	0.59	-	-	-	-	-	-	-
medusa	-	0.62	-	-	-	-	-	-	-	-
Uncertainty	-	0.57	0.35	0.43	0.66	0.57	0.49	0.51	0.60	0.46

Table 1: Comparison of factuality and fluency macro-F1 scores across multilingual settings. Results are reported for our proposed methods and the top three participating teams in the shared task. The highest and second-highest scores for each language are highlighted. Our fine-tuned decoder model achieves state-of-the-art performance in most languages.

Data	BN	EN	ES	FR	GU	HI	IT	ML	TE
train	**0.49**	0.51	0.48	**0.67**	0.34	0.45	0.74	0.36	0.35
+ translations	0.47	**0.57**	**0.50**	0.61	**0.45**	**0.51**	**0.8**	**0.50**	**0.44**

Table 2: Evaluation of the MMBert fine-tuned with and without translated data for factuality test on the SHROOM-CAP. Macro F1 is the evaluation metric. Translations significantly improved the final score for seven languages.

Model	BN	EN	ES	FR	GU	HI	IT	ML	TE
factuality									
Qwen3-Reranker-8B	0.31	0.74	0.72	0.79	0.63	**0.72**	0.86	**0.64**	0.62
Qwen3-14B	**0.70**	0.76	0.71	0.76	0.62	0.65	**0.87**	**0.64**	0.53
Qwen3-30B-A3B-Instruct	0.22	0.83	0.67	0.78	0.60	0.37	0.79	0.45	**0.70**
Qwen3-32B	0.69	0.83	**0.75**	**0.86**	**0.82**	**0.72**	0.86	0.63	0.66
sarvam-1	0.50	**0.86**	0.72	0.76	0.46	0.71	0.86	0.61	0.69
fluency									
Qwen3-Reranker-8B	0.62	0.65	0.58	0.79	0.55	**0.88**	0.55	0.67	0.80
Qwen3-14B	0.59	0.57	0.63	0.79	**0.67**	0.83	0.57	0.66	0.72
Qwen3-30B-A3B-Instruct	**0.74**	0.59	0.53	0.80	0.64	0.87	0.58	0.72	**0.83**
Qwen3-32B	**0.74**	**0.68**	0.53	0.82	0.64	0.87	**0.60**	0.72	**0.83**
sarvam-1	0.60	0.64	**0.64**	**0.84**	0.28	0.83	0.54	**0.74**	0.15

Table 3: Evaluation of the Decoder-based approach with different base models. The training performed on SHROOM-CAP train part. Macro F1 is the evaluation metric.

across languages, sampling-based approaches generally outperform the others, as expected. For instance, the SentenceSAR method performs best for English, while Eccentricity yields the highest performance for Guam, and LUQ performs best for Hindi. However, the DegMat method achieves the best average performance in factuality across all languages.

6 Conclusion

In this work, we present our systems for the SHROOM-CAP shared task. We explore three approaches: decoder-based, encoder-based, and uncertainty quantification. Decoder-based models achieved the strongest overall performance across both factuality and fluency tracks, confirming the advantage of large multilingual decoders when fine-tuned for error detection. Encoder-based models benefited from translation-based augmentation, improving robustness in low-resource settings. Uncertainty-based methods provided efficient, model-agnostic indicators that correlated with factuality errors.

Our findings suggest that reliable hallucination detection in scientific communication requires integrating generative reasoning, multilingual supervision, and uncertainty estimation. Future work may explore large-scale synthetic data augmentation, where the primary challenge lies in generating diverse and realistic negative multilingual samples for factual and fluency errors. This could help improve model robustness and generalization, especially in low-resource languages and domains. Another key area is developing adaptive multilingual models that better handle cross-lingual transfer and zero-shot settings with domain-specific knowledge incorporation.

References

Islam Aushev, Egor Kratkov, Evgenii Nikolaev, Andrei Glinskii, Vasilii Krikunov, Alexander Panchenko, Vasily Konovalov, and Julia Belikova. 2025. RAGulator: Effective RAG for regulatory question answering. In *Proceedings of the 1st Regulatory NLP Workshop (RegNLP 2025)*, pages 114–120, Abu Dhabi, UAE. Association for Computational Linguistics.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *CoRR*, abs/2307.15703.

Julia Belikova, Evegeniy Beliakin, and Vasily Konovalov. 2024. JellyBell at TextGraphs-17 shared task: Fusing large language models with external knowledge for enhanced question answering. In *Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing*, pages 154–160, Bangkok, Thailand. Association for Computational Linguistics.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5050–5063. Association for Computational Linguistics.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9367–9385, Bangkok, Thailand. Association for Computational Linguistics.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. LM-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, Singapore. Association for Computational Linguistics.

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

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1050–1059, New York, USA. PMLR.

Federica Gamba, Aman Sinha, Timothee Mickus, Raúl Vázquez, Patanjali Bhamidipati, Claudio Savelli, Ahana Chattopadhyay, Laura A. Zanella, Yash Kankanampati, Binesh Arakkal Remesh, Aryan Chandramania, Rohit Agarwal, Chuyuan Li, Ioana Buhnila, and Radhika Mamidi. 2025. Confabulations from ACL publications (CAP): A dataset for scientific hallucination detection. *CoRR*, abs/2510.22395.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *CoRR*, abs/2411.15594.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria*.

Maria Marina, Nikolay Ivanov, Sergey Pletenev, Mikhail Salnikov, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Alexander Panchenko, and Viktor Moskvoretskii. 2025. LLM-independent adaptive RAG: Let the question speak for itself. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8708–8720, Suzhou, China. Association for Computational Linguistics.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *arXiv preprint arXiv:2509.06888*.

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.

Viktor Moskvoretskii, Maria Marina, Mikhail Salnikov, Nikolay Ivanov, Sergey Pletenev, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Irina Nikishina, and Alexander Panchenko. 2025. Adaptive retrieval without self-knowledge? bringing uncertainty back home. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6355–6384, Vienna, Austria. Association for Computational Linguistics.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *Advances in Neural Information Processing Systems*, volume 37, pages 8901–8929.

Xin Qiu and Risto Miikkulainen. 2024. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *Advances in Neural Information Processing Systems*, volume 37, pages 134507–134533. Curran Associates, Inc.

Muhammad Reza Qorib, Alham Aji, and Hwee Tou Ng. 2024. Efficient and interpretable grammatical error correction with mixture of experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 17127–17138. Association for Computational Linguistics.

Elisei Rykov, Valerii Olisov, Maksim Savkin, Artem Vazhentsev, Kseniia Titova, Alexander Panchenko, Vasily Konovalov, and Julia Belikova. 2025a. SmurfCat at SemEval-2025 task 3: Bridging external knowledge and model uncertainty for enhanced hallucination detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1034–1045, Vienna, Austria. Association for Computational Linguistics.

Elisei Rykov, Kseniia Petrushina, Maksim Savkin, Valerii Olisov, Artem Vazhentsev, Kseniia Titova, Alexander Panchenko, Vasily Konovalov, and Julia Belikova. 2025b. When models lie, we learn: Multilingual span-level hallucination detection with PsiloQA. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11663–11682, Suzhou, China. Association for Computational Linguistics.

Aman Sinha, Federica Gamba, Ra'ul V'azquez, Timothee Mickus, Ahana Chattopadhyay, Laura Zanella, Binesh Arakkal Remesh, Yash Kankanampati, Aryan Chandramania, and Rohit Agarwal. 2025. SHROOM-CAP: Shared-task on Hallucinations and Related Observable Overgeneration Mistakes in Crosslingual Analyses of Publications. In *Proceedings of the 1st Workshop on Confabulation, Hallucinations & Overgeneration in Multilingual and Practical Settings*, Mumbai, India. Association for Computational Linguistics.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, and 1 others. 2025a. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Roman Vashurin, Maiya Goloburda, Preslav Nakov, Artem Shelmanov, and Maxim Panov. 2025b. Cocoa: A generalized approach to uncertainty quantification by integrating confidence and consistency of LLM outputs. *CoRR*, abs/2502.04964.

Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, and Artem Shelmanov. 2025. Uncertainty-aware attention heads: Efficient unsupervised uncertainty quantification for llms. *CoRR*, abs/2505.20045.

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, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. LUQ: Long-text uncertainty quantification for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262, Miami, Florida, USA. Association for Computational Linguistics.

A Hyperparameters

Hyperparameter	Value
Training max sequence length	4096
Ratio (context / prompt / output)	0.5 / 0.2 / 0.3
Batch size	14
Learning rate	1×10^{-4}
Weight decay	0.1
Optimizer	AdamW
Precision	bfloat16
LoRA rank / alpha	16 / 32
LoRA target	all-linear
Gradient checkpointing	Enabled
Max epochs	3
Validation metric	Macro-F1
Best model selection criterion	Validation loss

Table 4: Decoder training hyperparameters for factuality and fluency classification.

Hyperparameter	Value
Training max sequence length	8,092
Ratio (context / prompt / output)	0.5 / 0.2 / 0.3
Batch size	16
Learning rate	5×10^{-5}
Weight decay	0.1
Optimizer	AdamW
Precision	bfloat16
LoRA rank / alpha	Enabled
LoRA target	Enabled
Gradient checkpointing	Enabled
Max epochs	5
Validation metric	Macro-F1
Best model selection criterion	Macro-F1

Table 5: Encoder training hyperparameters for factual mistake classification.

B GPT-5 prompt

System:
Analyze a question about a scientific paper, the paper's abstract, the context retrieved from the paper, and an LLM answer.
Determine:
1. FACTUAL — whether the LLM answer is factual. If it contains any inconsistency with the abstract or context, mark it as False.
2. FLUENCY — whether the LLM answer has no fluency/language mistakes. If any such mistakes are present, mark it as False.
The abstract and relevant context are in English. The question and the LLM answer may be in any language.
Return the result strictly in this format:
FACTUAL: True|False
FLUENCY: True|False

User:
QUESTION: `<question>`
ABSTRACT: `<abstract>`
CONTEXT: `<context>`
LLM ANSWER: `<llm_answer>`

Figure 2: Prompt template for GPT-5.

C Decoder prompts

System:
You are a multilingual factuality judge. Your task is to determine whether the MODEL ANSWER contains ANY FACTUAL MISTAKES with respect to the provided RETRIEVED CONTEXT.
Factual mistakes = hallucinations, incorrect claims, information not supported or contradicted by the context. Ignore grammar, fluency, or style. Focus ONLY on factual consistency between answer and context.
The text may be in ANY language. Your answer must be language-agnostic.
Reply strictly with: 'y' — if the model answer contains any factual mistakes. 'n' — if the model answer is fully supported by or consistent with the retrieved context.
Do not explain your answer.

User:
Retrieved context: `<context>`
Prompt: `<prompt>`
Model answer to evaluate: `<output_text>`
Remember: reply ONLY with `'y'` or `'n'`.

Figure 3: Prompt template for factual consistency classification.

System:
You are a precise multilingual judge. Your task is to assess ONLY FLUENCY of a given LLM answer. Fluency = grammatical well-formedness, natural phrasing, coherent structure, sensible punctuation, and completeness. Ignore factual correctness and topic relevance entirely.
Reply strictly with: 'y' — if the text contains ANY fluency mistakes. 'n' — if the text has NO fluency mistakes.
Do not explain your answer.

User:
Generated answer to evaluate: `<output_text>`
Prompt for previous generation: `<prompt>`
Remember: reply ONLY with `'y'` or `'n'`.

Figure 4: Prompt template for fluency classification.

D Uncertainty Quantification Methods

Method	BN	EN	ES	FR	GU	HI	IT	ML	TE
				factuality					
SP	0.44	0.50	0.47	**0.56**	0.41	0.52	0.69	0.54	0.49
Perplexity	0.44	0.53	0.45	0.50	0.40	0.53	0.67	0.38	0.52
MTE	0.47	0.51	0.47	0.56	0.34	0.50	**0.71**	0.36	**0.57**
CCP	0.29	0.53	0.43	0.54	0.47	0.59	0.63	**0.55**	0.43
Token SAR	0.42	0.52	0.47	0.49	0.42	0.48	0.67	0.38	0.50
RAUQ	0.34	0.56	0.46	0.55	0.38	0.55	0.60	0.41	0.53
Lexical Similarity	0.42	0.53	0.40	0.49	0.35	0.47	0.56	0.50	0.38
DegMat	0.48	0.56	**0.58**	**0.56**	0.36	0.62	0.59	**0.55**	0.46
Eccentricity	0.34	0.57	0.52	0.52	**0.54**	0.54	0.58	0.41	0.47
LUQ	0.49	0.55	0.55	0.55	0.34	**0.65**	0.62	0.45	0.39
Semantic Entropy	0.49	0.50	0.49	0.55	0.34	0.46	0.67	0.49	0.54
Sentence SAR	0.47	**0.60**	0.44	**0.56**	0.38	0.58	0.60	0.54	0.54
SAR	0.39	0.59	0.48	0.49	0.34	0.62	0.61	0.39	0.56
KLE	**0.50**	0.45	0.44	0.53	0.34	0.61	0.65	0.36	0.35
Semantic Density	0.49	0.56	0.55	0.50	0.37	0.57	0.65	0.38	0.36
CoCoA	0.42	0.54	0.44	0.54	0.40	0.59	0.64	0.52	0.48
				fluency					
SP	0.48	0.18	0.24	0.52	0.56	0.46	0.46	0.59	0.39
Perplexity	0.37	0.30	0.24	0.52	0.50	0.38	0.47	0.33	0.38
MTE	0.38	0.29	0.27	0.49	0.45	0.30	**0.51**	0.31	0.37
CCP	0.45	0.29	0.27	0.51	**0.57**	0.40	0.48	**0.60**	0.27
Token SAR	0.35	0.32	0.27	0.51	0.53	0.24	0.46	0.33	0.36
RAUQ	0.25	0.32	0.25	0.50	0.51	0.38	0.41	0.39	0.39
Lexical Similarity	0.34	0.23	0.26	0.52	0.45	0.35	0.50	0.56	0.45
DegMat	0.43	0.33	0.32	0.46	0.49	0.27	0.45	0.59	0.44
Eccentricity	0.27	0.32	0.25	0.44	0.49	0.29	0.46	0.48	0.32
LUQ	0.44	**0.35**	**0.43**	0.47	0.47	0.34	0.48	0.49	**0.46**
Semantic Entropy	0.41	0.31	0.26	0.55	0.45	0.48	0.48	0.56	0.37
Sentence SAR	**0.57**	0.30	0.20	**0.66**	0.50	**0.49**	0.41	**0.60**	0.38
SAR	0.30	0.32	0.25	0.51	0.45	0.40	0.46	0.35	0.37
KLE	0.40	0.28	0.28	0.53	0.45	0.27	**0.51**	0.31	0.45
Semantic Density	0.41	0.25	0.35	0.44	0.46	0.26	0.45	0.33	**0.46**
CoCoA	0.42	0.32	0.23	0.59	0.56	0.45	0.48	0.57	0.38

Table 6: Detailed evaluation of selected uncertainty quantification methods. The best method is shown in **bold**, and the second-best is shown in underline.