"AGI" Team at SHROOM-CAP: Data-Centric Approach to Multilingual Hallucination Detection using XLM-RoBERTa

Harsh Rathva, Pruthwik Mishra, Shrikant Malviya
Sardar Vallabhbhai National Institute of Technology (SVNIT), Surat, India
{u24ai036,pruthwikmishra}@aid.svnit.ac.in,shrikant@coed.svnit.ac.in

Abstract

The detection of hallucinations in multilingual scientific text generated by Large Language Models (LLMs) presents significant challenges for reliable AI systems. This paper describes our submission to the SHROOM-CAP 2025 shared task on scientific hallucination detection across 9 languages. Unlike most approaches that focus primarily on model architecture, we adopted a data-centric strategy that addressed the critical issue of training data scarcity and imbalance. We unify and balance five existing datasets to create a comprehensive training corpus of 124,821 samples (50% correct, 50% hallucinated), representing a 172x increase over the original SHROOM training data. Our approach fine-tuned XLM-RoBERTa-Large with 560 million parameters on this enhanced dataset, achieves competitive performance across all languages, including **2nd place in Gujarati** (zero-shot language) with Factuality F1 of 0.5107, and rankings between 4th-6th place across the remaining 8 languages. Our results demonstrate that systematic data curation can significantly outperform architectural innovations alone, particularly for low-resource languages in zero-shot settings.

1 Introduction

Hallucinations in LLM-generated scientific text pose serious risks to research integrity and scientific communication, particularly when these systems are deployed in cross-lingual contexts where training data is limited in quanity. The SHROOM-CAP 2025 shared task (Sinha et al., 2025) addresses this critical problem by evaluating hallucination detection systems across 9 languages (5 training languages: English, Spanish, French, Hindi, Italian; 4 zero-shot languages: Bengali, Gujarati, Malayalam, Telugu) in scientific domains.

Most existing approaches to hallucination detection focus on improving model architecture or employing sophisticated prompting techniques with large proprietary models. However, we identify that the fundamental limitation in this task is the severe data imbalance and scarcity in the provided training set (only 724 samples with a 74% correct and 26% hallucinated distribution). Initial experiments reveal that models trained on these limited data exhibited extreme bias, predicting 99-100% of instances as hallucination instead of modeling the decision boundary.

A data-centric approach—systematically collecting, unifying, and balancing diverse hallucination datasets—would provide more substantial performance gains than model architecture modifications alone. This paper makes three primary contributions:

1. Creation of a large-scale, balanced multilingual hallucination detection dataset (124,821 samples) through unification of five existing resources

2. Demonstration that fine-tuning moderately-sized openly available models such as XLM-RoBERTa-Large (Conneau et al., 2020) on carefully curated data achieves competitive performance against larger and more complex systems

3. Analysis of the significant gap between validation and competition performance, highlighting distributional shifts in evaluation benchmarks

To ensure reproducibility and foster further research, we release all code, data processing scripts, and model weights publicly:

- **Code and datasets:** https://github.com/ezylopx5/SHROOM-CAP2025
- **Model weights:** https://huggingface.co/Haxxsh/XLMRHallucinationDetectorSHROOMCAP

2 Related Work

Hallucination Detection Approaches: Previous work on hallucination detection has explored various methodologies. Maynez et al. (2020) employed entailment-based approaches using natural language inference models, while Dhingra et al. (2022) used question-answering frameworks to verify factual consistency. More recent approaches have leveraged large language models with sophisticated prompting strategies (Li et al., 2023), though these often require API access to proprietary models and incur significant computational costs. But they are mostly limited to a uni-language scenario.

Proceedings of the 1st Workshop on Confabulation, Hallucinations and Overgeneration in Multilingual and Practical Settings (CHOMPS 2025), pages 96–100

Multilingual Representation Learning: Cross-lingual transfer learning has been extensively studied, with models like XLM-RoBERTa (Conneau et al., 2020) and mBERT (Devlin et al., 2019) demonstrating remarkable zero-shot capabilities. These models are typically pre-trained on massive multilingual corpora and can be fine-tuned on specific downstream tasks, making them ideal for low-resource language scenarios. But their adaptation to a unified data-centric scenario is largely unexplored.

Data-Centric AI: The recent emphasis on data-centric approaches (Whang et al., 2023) suggests that systematic data improvement often outperforms changes in model architecture. Our work aligns with this perspective, demonstrating that careful data curation and balancing can resolve fundamental model bias issues that architectural modifications cannot address.

Unlike earlier works, our approach does not rely on complex pipelines or proprietary models. Instead, we demonstrate that comprehensive data collection and standard fine-tuning of openly available multilingual models can achieve *competitive results across diverse languages*, including complete *zero-shot transfer to unseen languages*.

3 Dataset

We unify five existing hallucination detection datasets to create our training corpus:

1. **SHROOM TrainSet V1/V2** (Gamba et al., 2025): The official competition training data containing 724 samples across 5 languages (en, es, fr, hi, it) with scientific domain focus.

2. **hallucination_dataset_100k**: To further augment our training corpus, we create a large-scale synthetic dataset of 100,000 samples using AI-generated content. This dataset is constructed through systematic prompt engineering with large language models, following methodologies inspired by Tabular ARGN approaches (Tiwald et al., 2025).

   Generation Methodology: We employ a comprehensive prompt framework that systematically create both hallucinated and correct text samples across multiple domains. The prompt templates are designed to generate diverse hallucination types:

   - **Factual Errors**: Wrong dates, names, locations, and scientific facts
   - **Fabricated Details**: Plausible but entirely fictional information
   - **Mixed Information**: Combining facts from different sources incorrectly
   - **Subtle Hallucinations**: Near-miss dates and plausible but wrong details

   Quality Control: Each generated sample undergoes through multiple validation steps to ensure:

   (a) Clear distinction between hallucinated and correct samples
   (b) Factual accuracy verification for correct examples
   (c) Realistic and plausible hallucination patterns
   (d) Balanced distribution across domains and difficulty levels

3. **LibreEval** (Satya et al., 2024): A multilingual evaluation dataset for detecting various types of model errors, including hallucinations.

4. **FactCHD** (Chen et al., 2024): A fact-checking and hallucination detection dataset with verified annotations.

Preprocessing techniques such as: (1) label normalization to binary classification (correct/hallucinated), (2) language identification and verification, (3) random sampling to achieve perfect 50/50 class balance, and (4) text normalization to handle encoding variations are carried out. This process results in 124,821 high-quality training samples, representing a 172x increase over the original SHROOM training data with optimal class distribution.

4 Approach

4.1 Preprocessing

We model the task as a binary text classification problem. Each input instance consists of the LLM-generated text without additional metadata. We apply minimal text cleaning by stripping white-spaces appearing at the start and end of a text and normalizing unicode characters—while preserving the original linguistic characteristics. The text is tokenized using the XLM-RoBERTa tokenizer with a maximum sequence length of 256 tokens.

4.2 Translation-Based Data Augmentation

To address the challenge of limited training data for Indian languages, we explore two translation-based approaches.

Approach 1: English-to-Indian Language Translation We translate English training sentences into the Indian test languages using Facebook's NLLB-200-3.3B (Costa-Jussà et al., 2022, 2024) model. This creates additional training examples that could improve zero-shot performance by providing synthetic parallel data generated through machine translation.

Approach 2: Multilingual-to-English Translation We translate non-English training data into English using the same NLLB-200-3.3B model to create a larger English-centric training corpus. This approach leverages the abundance of English language models to achieve optimal performance.

Experimental Results: Both translation approaches results are shown in Tables 3 and 4 respectively.

Approach 1 (English-to-Indian) achieves Factuality F1 scores ranging from 0.366-0.595 and Fluency F1

Table 1: Comparison of Hallucination Detection Approaches

Approach	Key Technique	Multilingual Capability	Data Requirements
Entailment-based	Natural Language Inference	Limited	Task-specific data
QA-based	Question Answering	Language-specific	Large QA datasets
LLM Prompting	In-context Learning	Good with multilingual LLMs	Carefully crafted prompts
Our Approach	**Data-centric fine-tuning**	**Excellent (100 languages)**	**Unified multi-dataset**

Table 2: Unified Dataset Statistics

Source	Samples	Domain	Languages	Balance
SHROOM V1/V2	724	Scientific	5	74/26
hallucination_dataset_100k	100,000	General	Multiple	Varied
LibreEval	15,000	Mixed	Multiple	Varied
FactCHD	9,000	Fact-checking	Multiple	Varied
Combined (Ours)	**124,821**	**Mixed**	**100+**	**50/50**

scores from 0.173-0.347 across languages (Table 3). While some languages like Hindi (0.5944) and English (0.5949) show reasonable Factuality performance, the results are inconsistent and fail to match our final data-centric approach.

Approach 2 (Multilingual-to-English) performs even worse, with Factuality F1 scores ranging from 0.257-0.600 across languages (Table 4). Key limitations for both approaches include:

- **Translation Artifacts**: Machine translation introduces linguistic inconsistencies and unnatural phrasing
- **Domain Mismatch**: Scientific terminology translation can often be inaccurate
- **Amplified Bias**: The original dataset imbalance persists through translation
- **Inconsistent Performance**: Results vary significantly across languages without clear patterns

Table 3: Approach 1: English-to-Indian Translation Results

Language	Factuality F1	Fluency F1
Telugu (te)	0.4090	0.2942
Malayalam (ml)	0.4688	0.2996
Gujarati (gu)	0.4564	0.3474
Bengali (bn)	0.5707	0.3199
Italian (it)	0.3659	0.1728
Hindi (hi)	0.5944	0.2941
French (fr)	0.5310	0.2887
Spanish (es)	0.4560	0.1772
English (en)	0.5949	0.2376

Table 4: Approach 2: Multilingual-to-English Translation Results

Language	Factuality F1	Fluency F1
Telugu (te)	0.3689	0.1474
Malayalam (ml)	0.4639	0.3593
Gujarati (gu)	0.4241	0.1579
Bengali (bn)	0.4874	0.2542
Italian (it)	0.2570	0.4582
Hindi (hi)	0.4748	0.4353
French (fr)	0.4818	0.2899
Spanish (es)	0.4000	0.4607
English (en)	0.5999	0.4495

Given these unsatisfactory results from both translation approaches, our final submission utilizes the unified 124,821-sample dataset without translation augmentation. We find that the sheer volume, diversity, and balanced nature of our comprehensive training corpus provided superior coverage across languages, achieving better performance than translation-based approaches. Comparative analysis reveals that systematic data curation consistently outperforms translation-based augmentation for multilingual hallucination detection tasks.

4.3 Model Architecture

We use XLM-RoBERTa-Large (Conneau et al., 2020) as our base model that comprises of 560 million parameters and is pre-trained on 2.5TB of filtered CommonCrawl data [1] across 100 languages. The details about the model architecture are added in Table 5. We add a classification head consisting of a dropout layer

[1] https://github.com/facebookresearch/cc_net

with a 10% dropout rate followed by a linear layer that projected the [CLS] token representation to 2 output classes. The [CLS] token actually encodes the complete dense representation of any input sentence.

Table 5: Model Configuration Details

Parameter	Value
Base Model	XLM-RoBERTa-Large
Parameters	560M
Layers	24
Attention Heads	16
Hidden Dimension	1,024
Sequence Length	256
Classification Head	Dropout (0.1) + Linear

4.4 Training Procedure

We train the model using full fine-tuning (without any parameter-efficient methods) for 3 epochs with a batch size of 32, AdamW (Loshchilov and Hutter, 2017) optimizer (learning rate 2e-5, weight decay 0.01), and linear learning rate warmup over 10% of training steps. We employ a weighted cross-entropy loss with class weights [1.50, 1.00] to further mitigate any residual class imbalance. For training the model, we use an NVIDIA H200 GPU with 141GB VRAM. Model checkpoints are saved every 5,000 steps, and the best model is selected based on the F1 score achieved on the validation set.

5 Results and Discussion

5.1 Performance Evaluation

Our submission achieves competitive results across all 9 languages in the SHROOM-CAP 2025 competition:

Table 6: Official Competition Results

Language	Rank	Factuality F1	Fluency F1
Gujarati (gu)	**2**	**0.5107**	0.1579
Bengali (bn)	4	0.4449	0.2542
Hindi (hi)	4	0.4906	0.4353
Spanish (es)	5	0.4938	0.4607
French (fr)	5	0.4771	0.2899
Telugu (te)	5	0.4738	0.1474
Malayalam (ml)	5	0.4704	0.3593
English (en)	6	0.4246	0.4495
Italian (it)	5	0.3149	0.4582

Notably, our system achieved 2nd place in Gujarati, a zero-shot language, outperforming results in several training languages. This demonstrates the effectiveness of XLM-RoBERTa's cross-lingual representations when combined with sufficient and diverse training data.

5.2 Comparison with Baselines

The competition baseline system utilizes a standard approach without extensive data augmentation. Our method significantly outperforms this baseline in most languages, particularly in Factuality F1 scores. The top-performing team ("smurfcat") employs more complex ensemble methods and potentially larger models, achieving F1 scores between 0.65-0.92 across languages.

5.3 Validation vs. Competition Performance Gap

A notable observation is the substantial gap between our validation performance (macro F1: 0.8510) and competition performance (F1: 0.40-0.51). We identify several potential causes:

1. **Distribution Shift**: The test set likely contains different types of hallucinations or scientific domains not well-represented in the unified training dataset.

2. **Label Definition Misalignment**: Subtle differences in how "hallucination" is defined between the unified datasets and competition test set.

3. **Domain Specificity**: Our training data includes general-domain hallucinations, while the test focuses specifically on scientific text.

5.4 Error Analysis

We manually analyze misclassified examples and identified consistent patterns:

Factual Hallucinations: The model struggles with highly technical scientific claims that requires domain-specific knowledge beyond what is captured during XLM-RoBERTa's pre-training.

Example Error (False Negative):

- Input: "The protein folding mechanism involves quantum tunneling effects at room temperature."
- Model Prediction: Correct (0.62)
- Gold Label: Hallucinated
- Analysis: The model lacks specific biochemical knowledge to identify this as implausible.

Fluency Mistakes: The system performs notably worse on fluency detection (F1: 0.15-0.46) compared to factuality (F1: 0.44-0.51), particularly struggling with grammatical errors that resembles valid stylistic variations.

Cross-lingual Transfer: Surprisingly, zero-shot performance in Gujarati exceeds several training languages, suggesting that the quality and diversity of training data is more important than direct language exposure for this task.

6 Conclusion

We present a data-centric approach to multilingual scientific hallucination detection that achieves competitive results in the SHROOM-CAP 2025 shared task. By systematically unifying and balancing diverse datasets,

we create a robust training corpus that enabled effective fine-tuning of XLM-RoBERTa-Large. Our key finding is that data quantity and quality—particularly class balance—can overcome architectural limitations, with our simple approach achieving 2nd place in Gujarati and competitive rankings across 8 other languages.

Future Directions: Rather than generic suggestions, we propose concrete next steps: (1) investigating domain adaptation techniques specifically for scientific text, (2) developing data augmentation methods that generate scientific-domain hallucinations, (3) creating hybrid systems that combine our data-centric fine-tuning approach with the top team's ensemble strategies, (4) explicitly modeling the distribution shift between validation and test environments through domain generalization techniques, and (5) adding other metadata such as "abstract", "output_logits" to improve the performance of the models.

References

Xiang Chen, Duanzheng Song, Honghao Gui, Chenxi Wang, Ningyu Zhang, Yong Jiang, Fei Huang, Chengfei Lyu, Dan Zhang, and Huajun Chen. 2024. Factchd: benchmarking fact-conflicting hallucination detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Bhuwan Dhingra, Manaal Farina, Xinyi Chen, and Graham Neubig. 2022. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2210.11421*.

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. *arXiv preprint arXiv:2510.22395*.

Junyi Li, Xiaoping Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

Srinivasu Satya, Saad Jon, Gilhuly John, van Nest Nick, Gomes Julia, Khan Aman, Dhinakara Aparna, and Lopatecki Jason. 2024. Arize libreeval 1.0 - an open source dataset for evaluating hallucination in llms. https://github.com/Arize-ai/LibreEval.

Aman Sinha, Federica Gamba, Raúl Vázquez, 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.

Paul Tiwald, Ivona Krchova, Andrey Sidorenko, Mariana Vargas Vieyra, Mario Scriminaci, and Michael Platzer. 2025. Tabularargn: A flexible and efficient auto-regressive framework for generating high-fidelity synthetic data. *Preprint*, arXiv:2501.12012.

Steven E. Whang, Yuji Roh, Hyundong Song, and Jae-Gil Lee. 2023. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*.