Scalar_NITK at SHROOM-CAP: Multilingual Factual Hallucination and Fluency Error Detection in Scientific Publications Using Retrieval-Guided Evidence and Attention-Based Feature Fusion

Anjali R and Anand Kumar M
National Institute of Technology Karnataka, India
anjali.247it001@nitk.edu.in

Abstract

One of the key challenges of deploying Large Language Models (LLMs) in multilingual scenarios is maintaining output quality across two conditions: factual correctness and linguistic fluency. LLMs are liable to produce text with factual hallucinations, solid-sounding but false information, and fluency errors that take the form of grammatical mistakes, repetition, or unnatural speech patterns. In this paper, we address a two-framework solution for the end-to-end quality evaluation of LLM-generated text in low-resource languages. (1) For hallucination detection, we introduce a retrieval-augmented classification model that utilizes hybrid document retrieval, along with gradient boosting.(2) For fluency detection, we introduce a deep learning model that combines engineered statistical features with pre-trained semantic embeddings using an attention-based mechanism.

1 Introduction

Natural Language Generation (NLG) under Natural Language Processing (NLP) enables machines to generate human-like text in a wide range of languages and topics. Generating text in multiple languages has become increasingly feasible with the uplift in utilization of LLMs, making applications such as question answering, summarization, and content generation in low-resource languages more feasible. Nevertheless, even with their impressive abilities, LLMs are still vulnerable to two essential types of errors that have a material effect on output quality: factual hallucinations and fluency errors.

Fact-based hallucinations in LLM responses generate text that is semantically consistent and grammatically correct, but factually inaccurate or unsupported by the given context or reference materials. Hallucinations are especially undesirable in knowledge-intensive tasks, such as question answering, where the accuracy of facts is critical. Hallucinated answers can mislead users, weaken their confidence in AI systems, and spread misinformation, particularly in domains such as medical knowledge retrieval, legal document processing, and educational content generation.

Fluency errors, however, take the form of language errors, such as grammatical errors, abnormal repetition patterns, stilted expression, or improper use of language that renders the text unnatural or appears to have been generated by a computer. They heavily compromise the user experience and can indicate fundamental problems with the model's language understanding. Fluency issues are more severe in low-resource languages, where the training dataset is small and models struggle to encode the complexity of morphologically rich scripts.

The task of achieving both factual correctness and linguistic fluency on scientific publications becomes even more significant in multilingual environments, especially for languages that lack the plentiful digital resources available in high-resource languages such as English. Current quality evaluation systems have placed a significant emphasis on either factual confirmation or fluency testing, often in isolation, and primarily for English texts. There is an urgent need for robust systems that can evaluate both aspects of quality in multilingual LLM outputs simultaneously.

Despite the high performance of these current methods, some shortcomings remain. First, the majority of hallucination detection models fail to effectively utilize the rich contextual cues provided by reference documents, instead relying on elementary similarity scores that cannot detect semantic entailment and consistency. Second, fluency detection approaches often overlook crucial linguistic cues, such as lexical diversity metrics and character-level features, which are essential for low-resource languages. Third, there is limited work on creating unified approaches that integrate factual hallucination and fluency detection within a single quality

Proceedings of the 1st Workshop on Confabulation, Hallucinations and Overgeneration in Multilingual and Practical Settings (CHOMPS 2025), pages 90–95

estimation pipeline for multilingual scenarios.

This work fills these voids by introducing a dual-framework methodology for end-to-end quality evaluation of LLM outputs provided by SHROOM-CAP (Sinha et al., 2025)[1]. We present two complementary yet different systems: (1) an augmented retrieval-based classification system for factual hallucination detection merging hybrid retrieval and gradient boosting, and (2) an attention-based neural network for fluency detection merging statistical linguistic features and semantic embeddings.

2 Related Work

Efforts have been made in recent times to formalize the detection of hallucinations in text generated by LLMs. A prominent approach relies on retrieval-augmented methods that anchor generated text against reference documents or knowledge graphs (Gao et al., 2023). The methods utilize information retrieval techniques to extract salient context and calculate similarity or consistency scores between the generated and reference texts. Another task is analyzing model confidence signals, such as output logits and perplexity, as potential indicators of hallucination (Varshney et al., 2023).

Later research has considered self-consistency checking methods. (Manakul et al., 2023) presented SelfCheckGPT, which samples several responses from an LLM given the same prompt and calculates consistency between them, under the hypothesis that hallucinated facts will exhibit greater variance among samples. (Kadavath et al., 2022) illustrated how language models could be prompted to report uncertainty regarding their own responses, and that these self-reported confidence scores correspond with factual accuracy.

Natural Language Inference (NLI) models have also been used for hallucination detection. (Kryściński et al., 2019) introduced FactCC, a BERT-based model trained on synthetic data to predict whether a summary is factually consistent with its source document. (Laban et al., 2022) built on this with SummaC, demonstrating that NLI-based consistency checking can generalize across summarization datasets and domains. (Dziri et al., 2022) investigated attribution-based approaches that require models to cite exact evidence from source documents for every generated claim, enabling more explainable hallucination detection.

For fluency assessment, standard approaches have been grounded in linguistic properties such as part-of-speech patterns, measures of syntactic complexity, and language model perplexity (Higgins et al., 2014). Deep learning approaches that integrate pre-trained embeddings with crafted features have more recently demonstrated potential in addressing both semantic and surface fluency problems (Vajjala and Rama, 2018).

(Kaneko et al., 2022) demonstrated that encoder-decoder models with copy mechanisms can be effectively used to detect and correct grammatical errors. (Bryant et al., 2019) presented ERRANT, an error annotation tool that enables fine-grained analysis of various error types, thereby making fluency assessment more targeted.

The innovation of multilingual language models has opened NLP applications to hundreds of languages. (Conneau et al., 2020) presented XLM-RoBERTa, showing that massively multilingual pre-training allows successful cross-lingual transfer. (Reimers and Gurevych, 2020) generalized this approach to sentence embeddings with multilingual Sentence-BERT, enabling the computation of semantic similarity across languages.

Contemporary information retrieval has increasingly seen the use of hybrid methods merging sparse and dense approaches. (Robertson et al., 2009) set BM25 as the default sparse retrieval benchmark with its efficient term frequency-inverse document frequency weighting. (Karpukhin et al., 2020) presented Dense Passage Retrieval (DPR), demonstrating that dense embeddings from dual-encoder models surpass BM25 on question-answering tasks. Recent research has shown that splicing sparse and dense retrieval performs best. (Formal et al., 2021) presented SPLADE, which connects sparse and dense approaches by learning sparse representations in BERT's vocabulary space.

3 Dataset

Our experiments are conducted on multilingual datasets designed for hallucination detection and fluency error detection in LLM outputs, as provided by the SHROOM-CAP shared task (Gamba et al., 2025). The datasets support various languages, including Hindi (HI), French (FR), Italian (IT), Spanish (ES), and English (EN). Few Indic language based dataset equipped with only test sets are Malayalam (ML), Bengali (BN), Telugu (TE) and Gujarati (GU) enabling both language-specific and cross-lingual analyses. Each dataset includes

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

questions asked to LLM, generated answers along with logits, accompanied by reference documents (abstracts) and human-provided quality labels.

The data is split into three partitions for both languages: training, validation, and test sets as shown in the Table 1. The data is delivered in JSON Lines (JSONL) format. Every dataset contains two parallel files: a data file that contains the inputs and outputs, and a label file that contains the annotations.

Language	Train	Validation	Test
HI	265	240	240
FR	360	240	240
IT	360	240	240
EN	24	240	240
ES	20	240	255
ML*			788
GU*			800
TE*			798
BN*			798

Table 1: Train, Validation, and Test splits across languages.

The data files contain the following notable fields:

- **Index:** A unique index for every sample, allowing for alignment between data and labels.
- **Question:** The input question or prompt given to the language model.
- **Output_text:** The text output by the LLM in reply to the question.
- **Output_logits:** List of confidence scores (log probabilities) per token in the produced output, informative about model uncertainty.
- **abstract:** Source document or gold-standard text that has factually accurate information in relation to the question.

The label files provide annotations for two error types:

- **has_factual_mistakes:** Binary label as to whether the produced output includes hallucinations or factually inaccurate information.
- **has_fluency_mistakes:** Binary tag showing whether the output contains grammatical errors, linguistic errors, or unnatural patterns of language.

* -Train and Validation sets not provided for these languages in the dataset.

4 System Overview

We proposed two-framework solution for quality evaluation of LLM outputs: (1) a retrieval-augmented classifier for hallucination detection, and (2) an attention neural network for fluency detection.

4.1 Factual Hallucination Detection System

This system verifies the factual consistency between generated responses and reference documents through retrieval-augmented classification. Figure 1 illustrates our hallucination detection pipeline, which includes document chunking, hybrid retrieval, feature extraction, and XGBoost classification.

1. **Document Chunking:** The reference abstracts are divided into overlapping chunks using NLTK sentence tokenization to facilitate fine-grained retrieval while maintaining context.

2. **Hybrid Retrieval:** We use a three-stage retrieval pipeline. Stage one, utilizes BM25 sparse retrieval to select primary candidates based on lexical matching. Stage two, utilizes E5-large dense embeddings to rerank documents based on semantic similarity, achieved through a weighted aggregation of BM25 and cosine scores. Stage three, uses MiniLMv2 cross-encoder for final reranking, yielding the most relevant chunks. This hybrid model demonstrates lexical precision, semantic comprehension, and refined cross-attention.

3. **Feature Extraction:** We extract features in four categories from the returned chunks.
   - Similarity features for mean, max, and min cosine similarity between question/answer and chunks.
   - NLI features for entailment scores.
   - BM25 features for relevance scores.
   - Statistical features for length ratios, word overlaps, and text statistics.

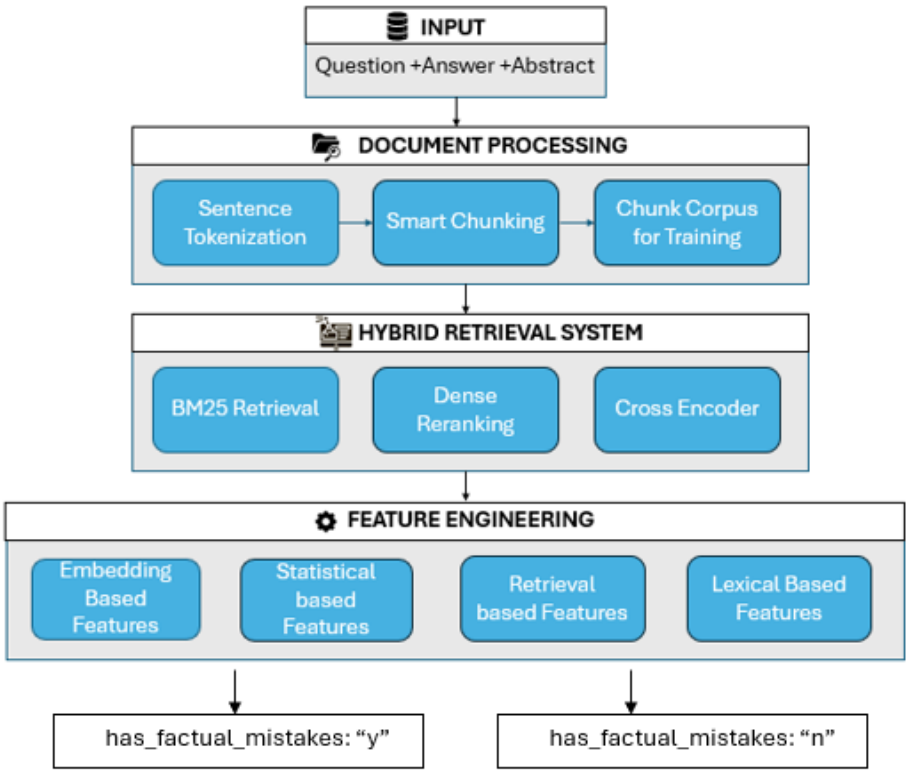

Figure 1: Architecture of Factual Hallucination Detection System

4. **Training:** We employ stratified cross-validation with XGBoost. SMOTE oversampling addresses class imbalance, and StandardScaler scales the features. Model predictions are combined across folds.

4.2 Fluency Detection System

Our fluency detection system integrates statistical linguistic features with semantic embeddings using an attention-based fusion network. Figure 2 presents our architecture, which processes text via parallel feature extraction paths and combines them through learned attention.

1. **Feature Engineering:** We derive statistical features grouped into three types,
   - **Logit Features:** Model confidence indicators such as mean, min, max, and standard deviation of output logits, log perplexity, low confidence ratio, and sequence length.
   - **LIWC-Style Linguistic Features:** Linguistic features such as pronoun, question, negation, conjunction, and quantifier ratios; frequencies of punctuation; word and sentence counts; repetition measures.
   - **Character Features:** Unicode-level statistics such as ASCII statistics, character type proportions for alphabetic, digit, space, punctuation, and uppercase characters; Devanagari script proportion; character variety; and special character patterns.

   Semantic Embeddings: Multilingual Sentence-BERT embeds output text in the form of dense semantic vectors that extract contextual meaning.

2. **Dual-Branch Processing:** Statistical features and embeddings are separately processed in two branches. Both branches feature linear layers with batch normalization, ReLU activation, and dropout as regularization techniques.

3. **Attention Fusion:** Concatenated branch outputs pass through an attention module that learns softmax weights to determine the optimal combination of statistical and semantic representations.

4. **Classifier:** The attention-fused representation is used as input to a multi-layer classifier with batch normalization, activation functions, and

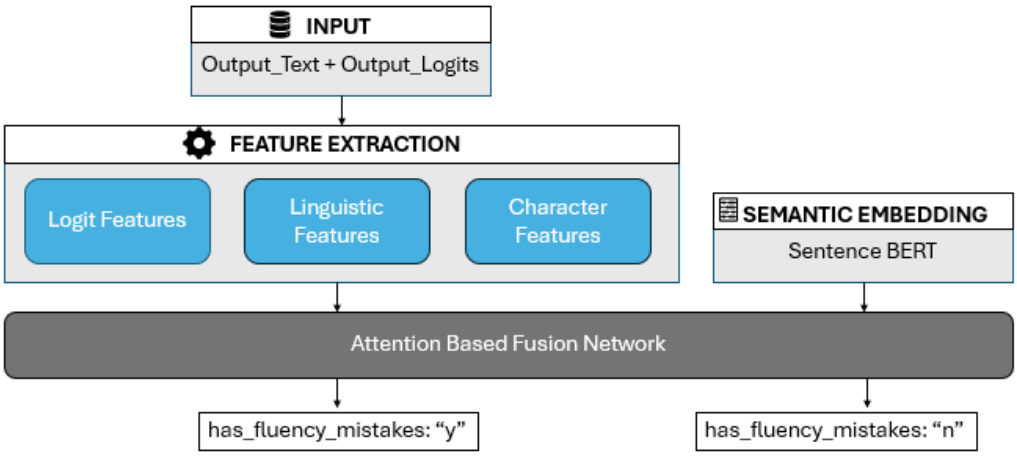

Figure 2: Architecture of Fluency Hallucination Detection System

dropout to yield final binary predictions.

5. **Training Strategy:** We employ stratified cross-validation with the AdamW optimizer, Focal Loss to address class imbalance, learning rate scheduling, validation-based F1 early stopping, and gradient clipping. Models of all folds are averaged at prediction.

5 Experimental Details

All experiments were conducted on NVIDIA GPU systems using PyTorch for neural networks and scikit-learn for machine learning models. We employ three pre-trained models without fine-tuning: intfloat/multilingual-e5-large for dense embeddings, cross-encoder/mmarco-mMiniLMv2 for reranking, and sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 for semantic embeddings. Random seeds are fixed for reproducibility.

Language	Validation		Test	
	Factual	Fluency	Factual	Fluency
HI	0.5980	0.86	0.6153	0.8359
FR	0.5663	0.6761	0.5524	0.6436
IT	0.6255	0.5071	0.5867	0.5442
EN	0.4432	0.4366	0.4667	0.4495
ES	0.6066	0.3443	0.4811	0.4607
ML*			0.3650	0.5209
GU*			0.3560	0.3060
TE*			0.3529	0.4597
BN*			0.4933	0.5182

Table 2: Validation and Test Macro-F1 Scores for Factual and Fluency Metrics across Languages.

6 Results

We evaluate our dual-framework approach on all the languages across both factual hallucination and fluency detection tasks. All reported results are averaged across 5-fold cross-validation, with final evaluation on held-out test sets. We prioritize macro F1-score as our primary metric due to significant class imbalance in both tasks. The results are as shown in the Table 2

7 Conclusion and Future Work

This paper presents a dual-framework approach for quality assessment of LLM-generated outputs, addressing factual hallucinations and fluency errors in multilingual contexts, with a particular focus on low-resource languages. We propose two complementary systems: a retrieval-augmented classifier that leverages hybrid BM25-dense-cross-encoder retrieval for hallucination detection, and an attention-based neural architecture that fuses statistical linguistic features with semantic embeddings for fluency assessment.

As part of our future work plan, we consider enhancing neural architectures with multi-head attention, hierarchical fusion, and Transformer-based encoders, and extending to span-level error localization and developing unified joint models for simultaneous hallucination and fluency detection.

References

Christopher Bryant, Mariano Felice, Øistein E Andersen, and Ted Briscoe. 2019. The bea-2019 shared task on grammatical error correction. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pages 52–75.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating attribution in dialogue systems: The begin benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

Federica Gamba, Aman Sinha, Timothee Mickus, Raul Vazquez, Patanjali Bhamidipati, Claudio Savelli, Ahana Chattopadhyay, Laura A Zanella, Yash Kankanampati, Binesh Arakkal Remesh, and 1 others. 2025. Confabulations from acl publications (cap): A dataset for scientific hallucination detection. *arXiv preprint arXiv:2510.22395*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Derrick Higgins, Chris Brew, Michael Heilman, Ramon Ziai, Lei Chen, Aoife Cahill, Michael Flor, Nitin Madnani, Joel Tetreault, Daniel Blanchard, and 1 others. 2014. Is getting the right answer just about choosing the right words? the role of syntactically-informed features in short answer scoring. *arXiv preprint arXiv:1403.0801*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Masahiro Kaneko, Sho Takase, Ayana Niwa, and Naoaki Okazaki. 2022. Interpretability for language learners using example-based grammatical error correction. *arXiv preprint arXiv:2203.07085*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

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

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

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

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Aman Sinha, Federica Gamba, Ra'ul V'azquez, Timothee Mickus, Laura Zanella, Binesh Arakkal Remesh, Yash Kankanampati, Ahana Chattopadhyay, 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.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal cefr classification. *arXiv preprint arXiv:1804.06636*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.