SemEval-2025 Task-3 — Mu-SHROOM

1 Introduction

This task is part of SemEval-2025 Task-3 — Mu-SHROOM, the Multilingual Shared-task on Hallucinations and Related Observable Overgeneration Mistakes, which builds on the previous SHROOM task with key changes: it covers multiple languages including Arabic (Modern Standard), Chinese (Mandarin), English, Finnish, French, German, Hindi, Italian, Spanish, and Swedish; emphasizes detecting hallucinations in outputs from instruction-tuned language models (LLMs); and requires participants to predict the locations of hallucinations within the generated text.

Mu-SHROOM focuses on identifying spans of text that represent hallucinations in LLM outputs, working within a multilingual and multi-model context. More information is available on the official task website.

2 Dataset Details

The datasets provided include a sample set, validation set, and unlabeled train set, with each data point formatted as a JSON object containing a unique identifier (`id`), language (`lang`), model input question (`model_input`), model identifier (`model_id`), generated output (`model_output_text`), hard labels indicating hallucination spans, and soft labels with start, end, and empirical probability of hallucination.

The hard labels will be used to assess intersection-over-union accuracy, while soft labels will measure correlation. Participants will reconstruct soft labels during evaluation by providing the required keys for detected spans.

2.1 Sample Dataset Illustration

{"id": 1, "lang": "EN", "model_input": "When was the restoration of the Sándor Palace completed?", "mode

3 Tasks

Participants are tasked with detecting hallucinations in the provided text.

3.1 Evaluation Metrics

Participants will be evaluated based on two character-level metrics: **Intersection-over-Union** measures the overlap between characters marked as hallucinations in the gold reference and those predicted by participants, and **Probability Correlation** assesses how well the probabilities assigned by participants align with those observed by annotators.

This is a preliminary overview of our proposal; further refinements will be made as the task develops.