SemEval-2025 Task-3 — Mu-SHROOM

1 Introduction

This task is part of SemEval-2025 Task-3—Mu-SHROOM, a multilingual shared-task focused on detecting hallucinations in outputs from instruction-tuned language models (LLMs). It builds on the previous SHROOM task with several key changes: it covers multiple languages, including Arabic, Chinese, English, Finnish, French, German, Hindi, Italian, Spanish, and Swedish; emphasizes LLM outputs; and requires participants to predict the locations of hallucinations within the generated text. More information is available on the official task website.

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.

```
{"id": 1, "lang": "EN", "model_input": "When was the restoration of the Sándor Palace completed?",
"model_output_text": "The restoration of Sándor Palace, also known as the Buda Castle ...",
"model_id": "TheBloke/Mistral-7B-Instruct-v0.2-GGUF",
"soft_labels": [{"start": 33, "prob": 0.3333333333, "end": 53}], "hard_labels": [[53, 64]]}
```

Participants are tasked with detecting hallucinations in the provided text. They will be evaluated based on two character-level metrics: **Intersection-over-Union**, which measures the overlap between characters marked as hallucinations in the gold reference and those predicted by participants, and **Probability Correlation**, which 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.