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. This iteration builds on the previous SHROOM task with key **changes**:

- **Multilingual Focus**: The task covers multiple languages, including Arabic (Modern Standard), Chinese (Mandarin), English, Finnish, French, German, Hindi, Italian, Spanish, and Swedish.
- **LLM Outputs**: The emphasis is on detecting hallucinations in outputs from instruction-tuned language models (LLMs).
- **Hallucination Detection**: Participants will 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. Each data point is 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 (binarized annotations) indicating hallucination spans
- Soft labels (continuous annotations) 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?",
    "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]
    ]
}
```


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:

1. **Intersection-over-Union**: Measures the overlap between characters marked as hallucinations in the gold reference and those predicted by participants.
2. **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.