MultiWikiQHalluA: A multilingual hallucination benchmark

Freja Thoresen, Dan Saattrup Smart
Alexandra Institute
Rued Langgaards Vej 7, 2300 København
freja.thoresen@alexandra.dk, dan.smart@alexandra.dk

Summary

- Multilingual hallucination benchmark with synthetic datasets for 306 languages and token-level classifiers for 30 European languages
- LettuceDetect framework with GPT-5-mini to generate hallucinated answers across factual, temporal, and numerical error types
- mmBERT-small (140m) outperforms EuroBERT and Ettin models, achieving ~91% accuracy on Danish and German hallucination detection
- Evaluated 5 models on 4 languages (English, Danish, German, Icelandic), with lower-resource languages showing significantly higher hallucination rates
- Model size alone doesn't predict hallucination rates; architecture and multilingual training data composition matter as much as parameter count

Hallucination data generation

For the LettuceDetect framework to generate a hallucinated dataset, it needs the following inputs:
- **Dataset** consisting of context, question, ground truth answer
- **Hallucination intensity** & **error types**
- **Language model** to generate the hallucinated answer

We provide the MultiWikiQA **dataset** separately for each language, the **hallucination intensity** is drawn for each sample (context, question, answer) from a beta distribution with mean of 0.2 and standard deviation of 0.15, and we use GPT-5-mini **language model** from OpenAI as the language model.

RAGFactChecker will generate hallucinated answers based on the following rules on the **hallucination intensity**:
- ≤ 0.2: Very subtle errors that are hard to detect
- ≤ 0.4: Moderate errors that are noticeable but plausible
- ≤ 0.6: Clear errors that are obviously incorrect
- ≤ 0.8: Strong errors that significantly change meaning
- > 0.8: Extreme errors that completely contradict the original

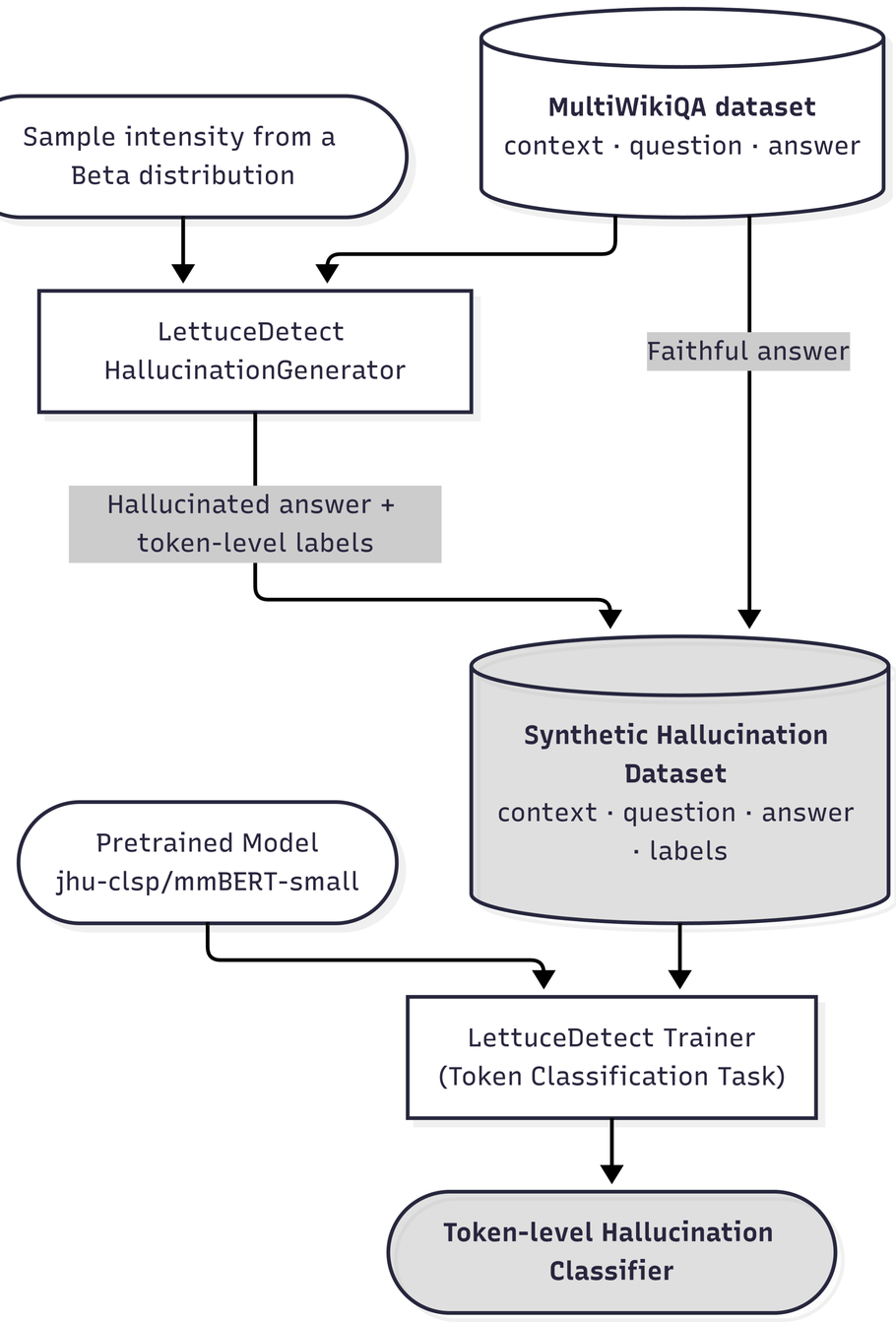
We use the default **error types** in LettuceDetect (and RAGFactChecker), which are the following:
- **Factual**: Change specific facts, entities, or claims.
- **Temporal**: Modify dates, time periods, or temporal relationships.
- **Numerical**: Alter numbers, quantities, percentages or measurements.

Concretely, RAGFactChecker instructs a language model to rewrite the reference answer according to the sampled intensity and error types, and to return
(i) the rewritten answer text and
(ii) a list of character-span pairs marking every modified portion.

These span annotations are projected onto the answer's subword tokens: any token whose character range overlaps with at least one hallucinated span is labelled 1 (unsupported); all remaining tokens are labelled 0 (supported). The generation results in a dataset with 5000 samples (4000 for training and 1000 for testing) with entirely hallucinated answers, for each language.

Classifier training

For each language, we use both the MultiWikiQA dataset with correct answers, and the hallucinated answer dataset generated with LettuceDetect. Hence, for each sample there is a "true" sample and a "hallucinated" sample, and both samples are used for training purposes, with binary labels assigned per token.



Evaluation of models

Each model was run on the MultiWikiQA test set (1,000 samples per language), and the fine-tuned mmBERT-small token-level hallucination classifier labeled every output token as either supported or hallucinated. The two largest models achieved the lowest or tied-lowest hallucination rates on three of four languages. The smallest model was substantially worse across the board.

Metric	Language	Qwen3-0.6B	Qwen3-14B	Gemma-3-12B-IT	Cogito-Qwen-32B	Cogito-Llama-70B
Hallucination rate	DA	17%	8%	8%	**7%**	**7%**
	DE	9%	**3%**	5%	5%	5%
	EN	3%	**1%**	2%	**1%**	2%
	IS	36%	17%	20%	18%	**15%**
Answer-level rate	DA	52%	12%	13%	9%	8%
	DE	17%	**4%**	6%	6%	6%
	EN	7%	2%	3%	**1%**	3%
	IS	60%	26%	27%	**18%**	19%

Resources

- **Dataset**: The synthetic hallucination dataset for 306 languages is available on HuggingFace.
- **Models**: The finetuned mmBERT-small hallucination classifiers for 30 European languages are available as a HuggingFace model collection.
- **Code**: The code for data generation, training, and evaluation is available on GitHub.
- **Evaluation**: Presented on EuroEval.

TrustLLM
ALEXANDRA INSTITUTTET
Funded by the European Union