MultiWikiQHalluA: A multilingual hallucination benchmark

Freja Thoresen, Dan Saattrup Smart

TrustLLM

Funded by the European Union

LettuceDetect

Kovacs & Recski, 2025

1. **Generation of hallucination dataset** with RagTruth per language

   Niu et. al, 2023

2. **Train token-level binary classifier** per language

3. **Evaluate models** by classifying tokens from QA & summarization tasks

Funded by the European Union

Generating false data

- **RAGFactChecker** will generate hallucinated answers based on the following rules on the hallucination intensity

```
Intensity <= 0.2: Very subtle errors that are hard to detect
Intensity <= 0.4: Moderate errors that are noticeable but plausible
Intensity <= 0.6: Clear errors that are obviously incorrect
Intensity <= 0.8: Strong errors that significantly change meaning
Intensity > 0.8: Extreme errors that completely contradict the original
```

- We use the default error types:

```
Factual: Change specific facts, entities, or claims.
Temporal: Modify dates, time periods, or temporal relationships.
Numerical: Alter numbers, quantities, percentages or measurements.
```


Funded by the European Union

To generate the hallucinated dataset, we need the following:

- Dataset with context, question, ground truth answer
  - **MultiWikiQA**
  - **(Reading comprehension dataset with 306 languages)**
- Hallucination intensity
  - **Drawn from a beta distribution with mean 0.2 and std 0.15**
- Language model to generate the hallucinated answer
  - **OpenAI GPT-5-mini**

Generated hallucination dataset

Question	Answer	Hallucination	Label
For what accolade was Jack B. Sowards put forward in relation to his contribution to Star Trek II: The Wrath of Khan?	Saturn Award	False	[]
For what accolade was Jack B. Sowards put forward in relation to his contribution to Star Trek II: The Wrath of Khan?	Golden Globe Award in 1984	True	[1,1,0,1,1]
On what date did Stagecoach stop running in Norfolk?	April 2018	False	[]
On what date did Stagecoach stop running in Norfolk?	June 2019	True	[1,1]

Funded by the European Union

Conclusion

- **LettuceDetect** classifies tokens(words) as hallucinated or not
- Trained **classifiers for 21 languages, and evaluated on 4 languages**
- Results **(see poster)**
  - Lower-resource languages tend to have a higher hallucination rate
  - Both model architecture and size affect the hallucination rate

- Benchmarks to be released on EuroEval

Funded by
the European Union

Contact us!

- Dr. Freja Thoresen, freja.thoresen@alexandra.dk
- Dr. Dan Saattrup Smart, dan.smart@alexandra.dk

You can also sign up for our newsletter to stay updated with our research!

Funded by the European Union