A multilingual hallucination benchmark: MultiWikiQHalluA

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

Abstract

Most hallucination evaluations focus on English, leaving it unclear whether findings transfer to lower-resource languages. We investigate faithfulness hallucinations, defined as model-generated content that is fluent and plausible but diverges from the provided input or is internally inconsistent. Leveraging the multilingual MultiWikiQA dataset, we utilize the LettuceDetect framework to create synthetic hallucination datasets for 306 languages, from which we train token-level hallucination classifiers for 30 European languages. In this work, we present evaluations of model hallucinations on a selection of languages: English, Danish, German, and Icelandic. Using these classifiers, we evaluate the hallucination rates for Qwen3-0.6B, Qwen3-14B, Gemma-3-12B-IT, cogito-v1-preview-qwen-32B, and cogito-v1-preview-llama-70B. Our classifiers reveal notably higher hallucination rates for Qwen3-0.6B (up to 60% of answers containing at least one hallucination, peaking in Icelandic) and generally lower rates for larger models, with cogito-v1-preview-qwen-32B and cogito-v1-preview-llama-70B performing best on most languages. Hallucination rates are consistently higher for lower-resource languages, particularly Icelandic.

Keywords: hallucination detection, multilingual natural language processing, token-level classification

1. Introduction

Large Language Models (LLMs) are prone to generating fluent yet false outputs, which is known as hallucinations. We adopt the definition of faithfulness hallucinations as proposed by Huang et al. (2025): a language model generates fluent and plausible content that diverges from the given input/prompt, or is internally inconsistent. For example, if a model is asked to summarise a passage about climate change and introduces a claim not present in the source text. This is distinct from factuality hallucinations, which involve factual errors with respect to real-world knowledge regardless of what input was provided, for example, a model stating that the Eiffel Tower is located in London. Accordingly, the evaluation frameworks in this work focus on internally inconsistent or ungrounded model behaviour rather than external factual correctness.

Studies assessing language models' factuality or evaluating whether the methods are effective to mitigate model hallucinations use different datasets and metrics. This makes it difficult to compare, in the same conditions, the factuality of different models as well as to compare the effectiveness of hallucination detection approaches. In this work, we use the same dataset, the open multilingual MultiWikiQA dataset by Smart (2025), to evaluate models in the different languages.

Most hallucination evaluations are conducted in English, leaving it unclear whether findings transfer to lower-resource languages. English, German, Danish, and Icelandic span a spectrum from highly to minimally represented in LLM pretraining corpora, providing a natural setting to study how language resource availability affects hallucination behaviour. We release an open source synthetic hallucination dataset covering 306 languages and train token-level classifiers for 30 European languages; in this paper we report evaluation results for four of those languages (English, Danish, German, and Icelandic).

In summary, our contributions are:

- Release a synthetic hallucination dataset for 306 languages (covering the full language support of MultiWikiQA).
- Release token-level hallucination classifiers for 30 European languages (a subset of the dataset languages for which we fine-tune models).
- Evaluate hallucination rates for five language models on four languages (English, Danish, German, and Icelandic).

2. Related Work

Hallucinations in language model outputs are commonly categorised into two types: factuality and faithfulness (Huang et al., 2025). Factuality hallucinations involve claims that contradict established world knowledge (e.g. stating that the Eiffel Tower is in London). Faithfulness hallucinations occur when generated text diverges from a provided source context, such as introducing unsupported claims when summarising a passage.

Factuality benchmarks assess a model's parametric knowledge. FEVER (Thorne et al., 2018) verifies claims against evidence corpora; FActScore (Min et al., 2023) evaluates atomic factual precision in long-form generations; TruthfulQA (Lin et al.,

Proceedings of The Fourth Workshop on the Role of Resources in the Age of Large Language Models (RESOURCEFUL 2026), pages 187–192 May 11, 2026.

2022) probes susceptibility to common misconceptions; HaluEval (Li et al., 2023) benchmarks hallucination detection across QA, summarisation, and dialogue; HalluLens (Bang et al., 2025) provides a broad multi-task evaluation of LLM hallucinations; and SimpleQA (Wei et al., 2024) measures short-form factual accuracy. These approaches primarily test world knowledge and may miss context-grounded errors.

Faithfulness evaluation targets settings where generation should be grounded in a provided context, such as reading comprehension or Retrieval-Augmented Generation (RAG). NLI-based methods recast faithfulness verification as textual entailment: TRUE (Honovich et al., 2022) shows that off-the-shelf NLI classifiers can serve as strong factual-consistency detectors. Other approaches include similarity-based metrics such as BERTScore (Zhang et al., 2019), model-based judges such as Halu-J (Wang et al., 2024), and stochastic self-consistency methods such as SelfCheckGPT (Manakul et al., 2023). Diagnostic frameworks such as RAGChecker (Ru et al., 2024) further motivate evaluation beyond coarse answer-level correctness. Most recently, LettuceDetect (Kovacs and Recski, 2025) moves from answer-level to token-level hallucination detection, enabling precise localisation of unfaithful spans.

Notably, the benchmarks and detection methods described above focus predominantly on English, leaving it unclear whether findings transfer to lower-resource languages. We adopt the LettuceDetect approach for its token-level precision and extend it to a multilingual QA setting using MultiWikiQA (Smart, 2025), training hallucination detection models for 30 European languages spanning a range of resource levels.

3. Methods

LettuceDetect (Kovacs and Recski, 2025) is a tool for detecting hallucinations in Retrieval-Augmented Generation (RAG) systems. It generates a hallucination dataset based on the dataset RagTruth (Niu et al., 2024) and then trains a binary token-level classifier on it. This trained model can then be used to detect hallucinations in LLM-generated text in a reading comprehension context. LettuceDetect has multilingual support (7 languages) using EuroBERT from (Boizard et al., 2025) and implementations with small Ettin models from (Weller et al., 2025). As a new addition to EuroBERT and Ettin models, we also train the mmBERT model from (Marone et al., 2025), and we introduce two new languages (Icelandic and Danish) which were not previously supported by LettuceDetect.

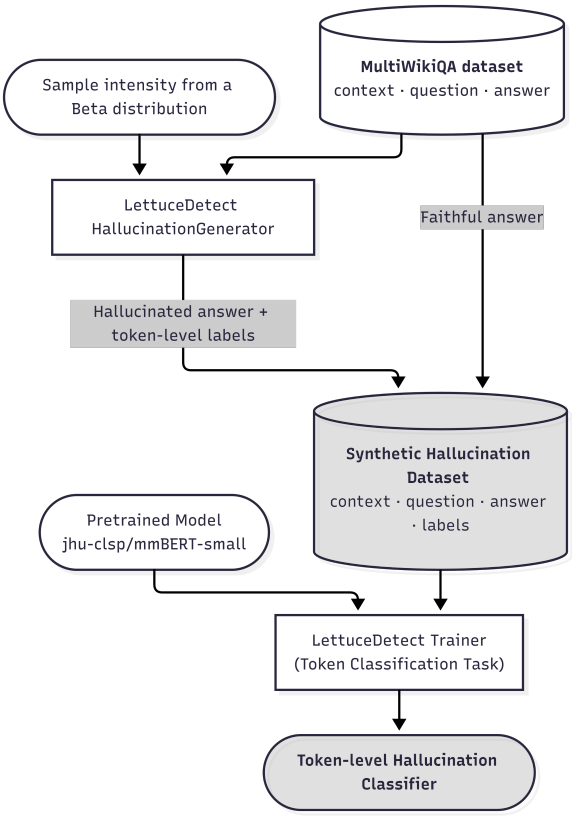

Figure 1: Overview of the two-stage methodology: synthetic hallucination data generation pipeline, where MultiWikiQA contexts, questions, and ground-truth answers are passed to the LettuceDetect framework, which uses a language model to produce token-labelled hallucinated answers; and fine-tuning of the mmBERT-small token-level hallucination classifier on the resulting dataset. The grey highlights the two deliverables: The synthetic hallucination dataset, and the token-level classifiers.

3.1. Datasets

We use the open multilingual dataset MultiWikiQA (Smart, 2025) as the foundation for all subsequent steps. The MultiWikiQA dataset supports 306 languages and contains context from Wikipedia articles, with questions, where the answers appear verbatim in the Wikipedia articles. In this study we evaluate on English, Danish, German, and Icelandic, while releasing resources for a broader set of 30 European languages. The training split includes 4000 context-question-answer triples and the test split contains 1000. In all subsequent experiments, models are evaluated on the exact same test set.

3.2. Hallucination data generation

The data generation works by supplying a dataset with context, questions, and answers, and the Let-

tuceDetect framework will then generate hallucinated answers using a language model. Instead of using the RagTruth dataset, as originally used in the (Kovacs and Recski, 2025), we use the MultiWikiQA dataset. We provide the LettuceDetect framework with the contexts, questions and answers from MultiWikiQA, and the framework then creates a false but plausible answer for each question. Hence, the result is a dataset with hallucinated answers, which can be used to train a classifier.

For the LettuceDetect framework to generate a hallucinated dataset, it needs the following inputs:

- Dataset consisting of context, question, ground truth answer
- Hallucination intensity
- 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 from OpenAI (OpenAI, 2024) as the language model. The beta distribution parameters were chosen to produce a distribution of hallucination intensities skewed toward subtle errors. The LettuceDetect framework then uses RAGFactChecker from Ru et al. (Ru et al., 2024) to generate the hallucinated answer. 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

RAGFactChecker can also create hallucinations on different error types. 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 $[(s_1, e_1), (s_2, e_2), \ldots]$ 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.

3.3. 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. To select the best base model for our classifier, we finetuned the token-level classifiers on the models in Table 1 using Danish and German. We chose these two languages because German is a high-resource language and Danish is a lower-resource language, allowing us to assess model performance across different resource levels. The F1-scores and accuracies are reported in Table 1. The mmBERT-small model performed best in both Danish and German, and therefore we use the mmBERT-small model as the model to finetune for hallucination detection for European languages.

3.4. Model Evaluation

When evaluating the models, we run model inference on the test set from the MultiWikiQA dataset. Then, we classify with the mmBERT-small finetuned classifier for each token if it was hallucinated or not. We evaluate Qwen3-0.6B and Qwen3-14B from Yang et al. (Yang et al., 2025), Gemma-3-12B-IT (Gemma Team, 2025), cogito-v1-preview-qwen-32B and cogito-v1-preview-llama-70B (Deep Cogito, 2025). The results are presented in Table 2.

4. Discussion

Across all models, high-resource languages (English and German) exhibit consistently lower hallucination rates than the lower-resource languages Danish and Icelandic, with Icelandic showing the highest rates. For the high-resource languages, the

Model	Language	Supported-F1	Unsupported-F1	Accuracy
Ettin-17m	Danish	0.8239	0.6560	0.7670
EuroBERT-210m	Danish	0.9062	0.8206	0.8768
mmBERT-small (140m)	Danish	**0.9143**	**0.8689**	**0.8963**
Ettin-17m	German	0.8761	0.7291	0.8299
EuroBERT-210m	German	0.7737	0.4759	0.6839
mmBERT-small (140m)	German	**0.9147**	**0.8627**	**0.8948**

Table 1: Classifiers finetuned with LettuceDetect on the MultiWikiQA train dataset with 4000 samples. The F1-scores and accuracies were evaluated from the test dataset with 1000 samples. The mmBERT-small model performed best in both Danish and German, and therefore we use the mmBERT-small model as the model to finetune for hallucination detection for European languages

Metric	Language	Qwen3-0.6B	Qwen3-14B	Gemma-3 -12B-IT	Cogito-Qwen -32B	Cogito-Llama -70B
Hallucination rate	DA	0.17	0.08	0.08	**0.07**	**0.07**
	DE	0.09	**0.03**	0.05	0.05	0.05
	EN	0.03	**0.01**	0.02	**0.01**	0.02
	IS	0.36	0.17	0.20	0.18	**0.15**
Answer-level rate	DA	0.52	0.12	0.13	0.09	**0.08**
	DE	0.17	**0.04**	0.06	0.06	0.06
	EN	0.07	0.02	0.03	**0.01**	0.03
	IS	0.60	0.26	0.27	**0.18**	0.19

Table 2: Hallucination scores by the finetuned mmBERT-small classifier for four languages: English (EN), Danish (DA), German (DE), and Icelandic (IS). *Hallucination rate* is the token-level rate (hallucinated tokens / total tokens); *Answer-level rate* is the fraction of answers containing at least one hallucinated token. Bold indicates the best (lowest) score per language per metric.

token-level hallucination rate remains low across all models except the smallest, whereas Danish and especially Icelandic reach notably higher rates. This pattern is more pronounced in the answer-level metric: for Icelandic, up to 60% of answers contain at least one hallucinated token with Qwen3-0.6B.

The two larger models, cogito-v1-preview-qwen-32B and cogito-v1-preview-llama-70B, achieve the lowest or tied-lowest hallucination rates on three of the four languages, while Qwen3-14B performs best on German. On Icelandic, cogito-v1-preview-llama-70B achieves the lowest token-level rate of 0.15, while cogito-v1-preview-qwen-32B achieves the lowest answer-level rate of 0.18. The smallest model, Qwen3-0.6B, shows substantially higher hallucination rates across all languages, with Icelandic being particularly affected.

Notably, the relationship between model size and hallucination rate is not strictly monotonic. For example, Qwen3-14B outperforms the larger cogito-v1-preview-qwen-32B on German, and cogito-v1-preview-llama-70B does not always outperform cogito-v1-preview-qwen-32B. This suggests that architecture, training data composition, and multilingual coverage may matter as much as raw parameter count for hallucination behaviour across languages, however a larger sample size is needed in order to draw conclusions.

The LettuceDetect approach proved practical for our multilingual setting. Although dataset generation and classifier training are one-time costs, inference-time hallucination scoring is fast, making the approach scalable for large-scale evaluation across many languages. However, the classifier may overestimate the hallucination rate due to false positives, particularly for lower-resource languages where training signal is noisier. Further experiments such as varying the hallucination intensity distribution or cross-validating against larger human annotation sets are needed to quantify this bias.

Another potential confound is tokenization: low-resource languages tend to produce more tokens per sentence than high-resource languages (Rust et al., 2021), because subword tokenizers trained predominantly on high-resource data split unfamiliar words into smaller pieces. This means that, for the same semantic content, a low-resource language may present more tokens to the classifier, increasing the opportunity for hallucination labels and inflating token-level hallucination rates. Disentangling the effect of tokenization granularity from

genuine hallucination behaviour is an important direction for future work.

5. Conclusion

In this work, we presented a multilingual hallucination benchmark leveraging the LettuceDetect framework and the MultiWikiQA dataset. We released a synthetic hallucination dataset for 306 languages and token-level hallucination classifiers for 30 European languages, and evaluated five language models (Qwen3-0.6B, Qwen3-14B, Gemma-3-12B-IT, cogito-v1-preview-qwen-32B, and cogito-v1-preview-llama-70B) on English, Danish, German, and Icelandic. Our finetuned mmBERT-small classifiers showed strong calibration on gold answers and revealed that hallucination rates are consistently higher for the lower-resource language Icelandic. Among the evaluated models, cogito-v1-preview-qwen-32B and cogito-v1-preview-llama-70B achieved the lowest hallucination rates on most languages, while Qwen3-14B performed best on German. Model size alone did not determine hallucination behaviour, suggesting that architecture and multilingual training data composition play an important role.

6. Resources

All resources are publicly available. Note that the **dataset** covers 306 languages (the full scope of MultiWikiQA), the **classifiers** are released for 30 European languages (the subset for which we finetuned models), and the **evaluations** in this paper cover four languages (English, Danish, German, and Icelandic).

- **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.

7. Acknowledgements

This research was funded by the EU Horizon project TrustLLM (grant agreement number 101135671).

8. Bibliographical References

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. HalluLens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156, Vienna, Austria. Association for Computational Linguistics.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno M. Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. Eurobert: Scaling multilingual encoders for european languages.

Deep Cogito. 2025. Cogito v1 preview. Model release.

Gemma Team. 2025. Gemma 3 technical report.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitel, Sumit Sahrawat, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920.

Lianjin Huang, Weize Yu, Weiguang Ma, Wei Zhong, Zhen Feng, Haoran Wang, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Akos Kovacs and Gabor Recski. 2025. Lettucedetect: A hallucination detection framework for rag applications.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3214–3252.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Matt Marone, Oren Weller, William Fleshman, Eric Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. Blog post.

Donghao Ru, Liang Qiu, Xiaoyang Hu, Tong Zhang, Peng Shi, Shiyu Chang, and Zheng Zhang. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 37:21999–22027.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819.

Binjie Wang, Steffi Chern, Ethan Chern, and Pengfei Liu. 2024. Halu-j: Critique-based hallucination judge. *arXiv preprint arXiv:2407.12943*.

Jason Wei, Nanyun Karina, Hyung Won Chung, Yao Jie Jiao, Stuart Papay, Aidan Glaese, and William Fedus. 2024. Measuring short-form factuality in large language models.

Oren Weller, Kaleab Ricci, Matt Marone, Alexander Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2025. Seq vs seq: An open suite of paired encoders and decoders.

Aohan Yang, An Li, Bo Yang, Bingchao Zhang, Bin Hui, Bo Zheng, and Zheng Qiu. 2025. Qwen3 technical report.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert.

9. Language Resource References

Cheng Niu and Yuanhao Wu and Juno Zhu and Siliang Xu and Kashun Shum and Randy Zhong and Juntong Song and Tong Zhang. 2024. *RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models*.

Dan Saattrup Smart. 2025. *MultiWikiQA: A Reading Comprehension Benchmark in 300+ Languages*.