When Models Lie, We Learn: Multilingual Span-Level Hallucination Detection with PsiloQA

Elisei Rykov[1], Kseniia Petrushina[1,5], Maksim Savkin[2,5], Valerii Olisov[5], Artem Vazhentsev[2,1], Kseniia Titova[3,1], Alexander Panchenko[1,2], Vasily Konovalov[2,1,5], and Julia Belikova[4,1]
[1]Skoltech, [2]AIRI, [3]MWS AI, [4]Sber AI Lab,
[5]Moscow Institute of Physics and Technology
{Elisei.Rykov, A.Panchenko, Julia.Belikova}@skol.tech

Abstract

Hallucination detection remains a fundamental challenge for the safe and reliable deployment of large language models (LLMs), especially in applications requiring factual accuracy. Existing hallucination benchmarks often operate at the sequence level and are limited to English, lacking the fine-grained, multilingual supervision needed for a comprehensive evaluation. In this work, we introduce PsiloQA, a large-scale, multilingual dataset annotated with span-level hallucinations across 14 languages. PsiloQA is constructed through an automated three-stage pipeline: generating question–answer pairs from Wikipedia using GPT-4o, eliciting potentially hallucinated answers from diverse LLMs in a no-context setting, and automatically annotating hallucinated spans using GPT-4o by comparing against golden answers and retrieved context. We evaluate a wide range of hallucination detection methods – including uncertainty quantification, LLM-based tagging, and fine-tuned encoder models – and show that encoder-based models achieve the strongest performance across languages. Furthermore, PsiloQA demonstrates effective cross-lingual generalization and supports robust knowledge transfer to other benchmarks, all while being significantly more cost-efficient than human-annotated datasets. Our dataset and results advance the development of scalable, fine-grained hallucination detection in multilingual settings.[1]

1 Introduction

Large Language Models (LLMs) became a crucial component in a wide range of text generation applications, including summarization, translation, and question-answering systems in various domains. However, even state-of-the-art models, such as GPT-4 (OpenAI, 2023), or open-weight models, such as LLaMa (Dubey et al., 2024) and DeepSeek (DeepSeek-AI et al., 2025) are inevitably prone to production of hallucinations or unsupported facts in their generated output (Xiao and Wang, 2021; Dziri et al., 2022; Xu et al., 2024). This phenomenon poses a crucial obstacle for the real-world deployment of LLMs, particularly in safety-critical domains such as medicine (Ben Abacha and Demner-Fushman, 2019; He et al., 2025). A single hallucinated word can substantially alter the overall meaning of the generation, potentially causing harm to end-users. Consequently, hallucination detection has become a critical challenge in the development and application of LLMs (Huang et al., 2025).

Hallucination detection is typically categorized into three standard tasks: *sequence-level*, *span-level*, and *entity-level*. Sequence-level detection focuses on identifying entire generations that contain some factual inconsistencies. In contrast, span-level and entity-level detection addresses the more challenging task of precisely highlighting factually misaligned spans or individual entities within the generated text.

Several approaches have been proposed for the detection of hallucination of LLMs. Uncertainty quantification (UQ) emerging as one of the most prominent research directions (Gal and Ghahramani, 2016; Shelmanov et al., 2021; Baan et al., 2023; Geng et al., 2024; Vashurin et al., 2025; Belikova et al., 2024). Recently, numerous UQ methods have been developed specifically for LLMs (Kuhn et al., 2023; Lin et al., 2023). However, most of these methods focus on sequence-level verification (Farquhar et al., 2024), while only a few methods operate at the token or span level (Zhang et al., 2023; Fadeeva et al., 2024). While Rykov et al. (2025) proposed combining uncertainty estimation methods with a fine-tuned LLM, Qwen2.5-7B-Instruct, using a weighted averaging approach, where the weights were optimized separately for each language.

[1]https://github.com/s-nlp/psiloqa

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 11663–11682
November 4-9, 2025

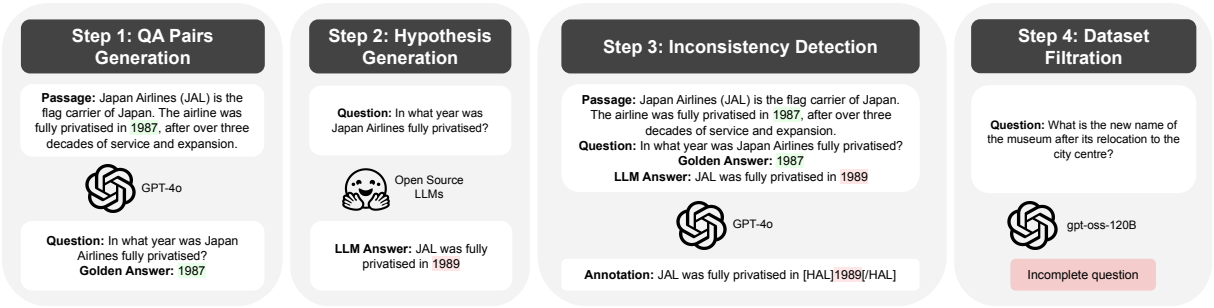

Figure 1: PsiloQA generation pipeline, where a dataset is built in fourth steps. **Step 1:** Generation of multilingual question-answer pairs using the GPT-4o and randomly retrieved passages from Wikipedia articles. **Step 2:** Generation of an answer to the question without the supporting passages from Wikipedia. By using only internal knowledge without external sources of information, LLMs cannot easily answer hard factual questions. **Step 3:** Span-level inconsistency detection between the golden answer generated by GPT-4o and the LLM hypothesis. **Step 4:** Filtration of incomplete or subjective questions and cases when LLM refuses to answer.

Although UQ is a rapidly growing area of research, even the most advanced methods still face significant limitations. For instance, sampling-based methods (Duan et al., 2024) require substantial computational overhead and operate only on the sequence level. Information-based methods (Fomicheva et al., 2020; Fadeeva et al., 2024) demonstrate strong performance in span-level tasks but still fail to detect some hallucinations and remain far from ideal performance.

Another set of approaches focuses on fact-checking techniques based on external knowledge sources (Niu et al., 2024) or auxiliary LLMs (Mishra et al., 2024). While these methods achieve high performance, they require substantial computational overhead. These approaches first extract atomic claims from the generated response and compare them to a retrieved context using an auxiliary LLM (Min et al., 2023). This process produces a verification score, indicating the degree to which the extracted claims supported with the retrieved evidence. Moreover, the performance of such systems is heavily dependent on the quality of both the retrieved context and the auxiliary LLM, which typically requires fine-tuning for better performance (Mishra et al., 2024).

To evaluate the quality of both systems, including those based on uncertainty quantification and external knowledge, we require a dataset with annotated hallucinations. High-quality, fine-grained annotations, particularly at the span level, are labor-intensive, requiring expert human annotators (Vazquez et al., 2025) or costly automated pipelines (Min et al., 2023). Span-level annotation, though more practical for pinpointing unsupported text, introduces additional complexity compared to sequence-level verification. Multilingual contexts increase these challenges due to linguistic diversity, limited non-English data availability, and the need for language-specific pipeline adjustments (Vashurin et al., 2025).

To address these limitations, we introduce a novel methodology for automatically generating multilingual data with fine-grained hallucination annotations, which includes (i) synthetically creating question-answering pairs from Wikipedia article summaries, (ii) generating hypotheses using LLMs in a zero-context setting to produce both hallucinated and accurate answers, (iii) automating span-level inconsistency annotation by comparing responses to context and ground truth via an advanced LLM, and (iv) automated filtering step through both rule-based and prompt-based methods. Further details of this pipeline are outlined in Section 3.

The contributions of this work could be summarized as follows:

- We propose an automated and scalable pipeline for generating and annotating synthetic data.
- We introduce a large multilingual dataset with high-quality and fine-grained span-level hallucination annotations for numerous open-weighted and proprietary LLMs.
- We conduct comprehensive empirical evaluations of various state-of-the-art hallucination detection methods of different types across 14 languages.

Dataset	Domain	Annotation	Generation	Lang	# LLMs	# Train	# Val	# Test	Licence
Mu-SHROOM (Vazquez et al., 2025)	General	Manual	Natural	Mult	38	3,351*	499	1,902	CC-BY-4.0
HalluEntity (Yeh et al., 2025)	Biography	Manual	Natural	En	1	–	–	157	MIT
$\text{RAGTruth}_{\text{QA}}$ (Niu et al., 2024)	General	Manual	Natural	En	6	5,034	–	900	MIT
FAVA-Bench (Mishra et al., 2024)	General	Auto	Synthetic	En	3	–	–	902	CC-BY-4.0
PsiloQA (ours)	General	Auto	Natural	Mult	24	63,792	3,355	2,897	CC-BY-4.0

Table 1: Comparative overview of span-level hallucination detection datasets. The Mu-SHROOM dataset has an unlabeled training set (*) comprising 4 languages (en, es, fr, zh). The **Generation** column distinguishes whether LLM answers were generated with intentional error insertion (synthetic) or used as-is (natural).

2 Related Work

2.1 Hallucination Detection Datasets

Most hallucination detection benchmarks operate at the sentence or paragraph level, such as TruthfulQA (Lin et al., 2022), ANAH (Ji et al., 2024; Gu et al., 2024) and HaluEval (Li et al., 2023). These benchmarks categorize each generated response as either hallucinated or correct. However, instance-level detection is unable to identify specific hallucinated content, which is essential for correcting misinformation. This limitation is especially concerning in long-form text where a single response may include both supported and unsupported information, rendering binary quality assessments insufficient (Min et al., 2023).

To tackle these issues, recent studies have improved benchmarks for more detailed hallucination detection. For instance, Min et al. (2023) introduced FActScore, a dataset focusing on fine-grained hallucination detection in Wikipedia bibliographies. It aids in assessing hallucination detection techniques that utilize external knowledge and detailed fact-level annotations.

Similarly, HalluEntity (Yeh et al., 2025) includes biographies created by ChatGPT, with each entry comprising a name, a ChatGPT-generated biography, and a list of atomic facts marked as True or False, aligning with the relevant entity in the language model's output.

RAGTruth (Niu et al., 2024) is a large-scale benchmark with nearly 18,000 human-annotated examples designed for retrieval-augmented generation (RAG) tasks. It provides fine-grained word-level annotations marking hallucinated spans that contradict reference documents, covering question answering, data-to-text, and summarization tasks.

FAVA (FavaBench) (Mishra et al., 2024) is a dataset containing fine-grained hallucination annotations with external knowledge support, used to evaluate hallucination detection and editing. It includes diverse generation tasks and has been shown effective in fine-grained hallucination detection research.

The multilingual Mu-SHROOM dataset has been suggested in the shared task on Multilingual Hallucinations and Related Observable Overgeneration Mistakes (Vazquez et al., 2025). In particular, the Mu-SHROOM task aims to detect hallucination spans in the outputs of instruction-tuned LLMs in multilingual context models for Arabic, Basque, Catalan, Chinese, Czech, English, Farsi, Finnish, French, German, Hindi, Italian, Spanish, and Swedish.[2] The datasets with their respective annotation levels and splits are outlined in Table 1.

2.2 Hallucination Detection Methods

Most hallucination detection methods operate at the sentence-level. For instance, recent works propose various sampling-based UQ methods that measure the consistency of multiple sampled generations (Kuhn et al., 2023; Lin et al., 2023; Duan et al., 2024; Zhang et al., 2024). On the contrary, reflexive methods aims to assess an LLM's confidence in its generation by directly prompting it for self-evaluation (Kadavath et al., 2022; Tian et al., 2023). Supervised methods are applicable at both sequence and token levels, but they often require model-specific training due to differences in number of hidden features and attention heads (Azaria and Mitchell, 2023; Vazhentsev et al., 2024; Chuang et al., 2024; CH-Wang et al., 2024; Vazhentsev et al., 2025).

Token probability and entropy (Fomicheva et al., 2020) are trivial baselines for span-level detection, which utilize the distribution of token probabilities. Fadeeva et al. (2024) propose to analyze the consistency of the top most probable token candidates by leveraging the natural language inference (NLI) model using the Claim Conditioned Probability (CCP) method. Zhang et al. (2023) propose to model the conditional dependencies between the

[2] https://helsinki-nlp.github.io/shroom

generated tokens by reweighing token uncertainty scores, leveraging uncertainty of the previous tokens and attention weights after max-pooling.

Several methods leverage external knowledge to evaluate the factuality of the generations. Among the most well-known is FActScore (Min et al., 2023), which extracts atomic facts from the model's response and compares them to a retrieved context using an additional LLM. This fact-checking process generates a score that indicates whether the claims are supported by the retrieved context.

Niu et al. (2024) introduce RAGTruth, a pipeline for the detection of word-level hallucinations in retrieval-augmented generation (RAG) systems. This work introduces a dataset and benchmark to evaluate the factuality of LLM responses for various tasks, such as summarization, question-answering, and others. Moreover, this framework could be easily adapted for the fact-checking task. The hallucinations annotations in the presented data were created by human annotators.

Furthermore, the task of hallucination detection can naturally be extended to hallucination editing. For example, the FAVA (Mishra et al., 2024) model is specifically trained for word-level hallucination detection and editing tasks according to the introduced hallucination taxonomy. To collect training data, the authors asked LLMs to insert errors from the introduced taxonomy into the responses.

Despite their advantages, both RAGTruth and FAVA are limited in their applicability, as they are designed only for English-language tasks and require human annotations.

3 PsiloQA: A Synthetic Span-Level Hallucination Dataset

3.1 Dataset Generation Process

Figure 1 illustrates the dataset generation pipeline. Our objectives in developing the PsiloQA generation process include: (i) utilizing real LLM hallucinations rather than artificially inserted errors; (ii) ensuring the process is cost-effective, quick, and scalable; (iii) encompassing multiple languages and domains.

The initial stage of the PsiloQA pipeline involves generating context-based question–answer pairs. Given the scarcity of such multilingual data in the general domain, we constructed a multilingual context-based QA dataset from scratch. This is achieved by utilizing passages from Wikipedia to source diverse, multilingual data. The passages, along with a specific prompt, are submitted to GPT-4o to generate QA pairs. To achieve varying question complexity, 3 different question-answer pairs are produced with varying levels of complexity, as demonstrated in Figure 7.

To generate answers that contain hallucinations, we asked various LLMs to respond to previously generated questions without referring to Wikipedia for support. When relying solely on internal knowledge, LLMs often produce hallucinations in their responses to factual questions. The same models employed in Mu-SHROOM (Vazquez et al., 2025) were used to generate these inaccurate answers.

To catch hallucinations, we prompt GPT-4o to review the passage and question, comparing the golden answer with the LLM hypothesis. Any discrepancies are marked with `[HAL]` tags. Otherwise, the LLM's response is copied unchanged when no inconsistencies are found. The prompt for detecting inconsistencies is illustrated in Figure 8. For span-level annotation, we followed the RAGTruth pipeline (Niu et al., 2024) and asked GPT-4o to annotate spans at the word level, encouraging precise labeling and discouraging overgeneralization (e.g., marking the entire answer as a hallucination).

Additionally, we conducted several automatic filtering steps, including both rule-based and prompt-based. The rule-based filter removes samples where `[HAL]` tags did not properly match, cases where the annotator model generated empty spans, and cases where the annotated answer is not consistent with the initial LLM's answer after removing all `[HAL]` tags. The prompt-based filter removes subjective questions, incomplete questions, and answers where LLM refuses to answer. Subjective questions are non-factual and thus do not require context for answering. Consequently, it is challenging to recognize any inconsistencies when comparing the answers to subjective questions with the original contexts. Incomplete questions are artifacts of the QA pairs generation process, as they have a reference to the original context, or use pronouns with no clear antecedent. The absence of an explicit subject in a question makes it unanswerable without context. Finally, cases when LLM refuses to answer also introduce difficulties in identifying inconsistent segments, as they do not constitute responses to factual questions. Prompt-based filtering was performed using `gpt-oss-120B`[3] model.

[3] https://hf.co/openai/gpt-oss-120b

Instructions are shown in Appendix H. In total, approximately 6,500 samples were filtered.

Consequently, PsiloQA fulfills all our previously outlined criteria. It utilizes LLM for question answering, which results in samples containing genuine hallucinations. Furthermore, using GPT-4o for annotation makes the generation process scalable, more cost-effective, and faster compared to human annotation. Wikipedia is used as a dependable and varied source of multilingual seeds.

After all filtration steps, the training set of PsiloQA consists of 63,792 samples. For each language and LLM checkpoint combination, we select 100 random samples for benchmarking. The PsiloQA testing split contains 2,897 samples.

3.2 Dataset Statistics

Figure 4 illustrates the distribution of samples by language in the PsiloQA dataset. English is the most prevalent, with nearly 23,000 samples. Hindi, Finnish, Catalan, Chinese, Swedish, and Czech each range between 5,000 and 7,000 samples. The dataset contains roughly 3,700 Farsi samples, and approximately 2,000 to 2,500 samples per language for Spanish, Euskara, French, Italian, and Arabic. German appears as the least represented language, with about 1,500 samples. Figure 2 displays the statistics by LLMs. Figure 5 illustrates the number of hallucination spans present in each sample. There are 14,000 samples with no hallucinations and 50,000 samples containing just one span of hallucination. Other samples have 2 or more spans, with a maximum of 10 in rare cases. Figure 6 depicts the word distribution of span lengths. PsiloQA spans are relatively short, with 50,000 spans including fewer than 5 words.

Also, we analyzed the distribution of predicted domains in the PsiloQA dataset. Each passage was assigned a single domain using zero-shot classification with the `bart-large-mnli`[4] model across 34 candidate domains. Geography and Sports are the most prevalent, with roughly 25–30% of samples each. Overall, the dataset exhibits a diverse distribution, with some highly represented domains and a long tail of less frequent categories.

3.3 Dataset Production Cost

The estimated cost of GPT-4o labeling is $535, based on token generation pricing where $4 is charged for every 1M input tokens and $16 for every 1M tokens. To compare this with RAGTruth, the only span-level hallucination dataset with a labeled training set, we estimated its cost. To annotate RAGTruth, annotators proficient in English and holding a bachelor's degree in English, Communications, or relevant fields were employed to ensure accuracy and reliability. They were recruited from a professional vendor and compensated at $25 per hour per individual. Each response was labeled by two annotators, achieving a 91.8% consistency rate at the response level and 78.8% at the span level. The total cost of labeling is not specified in the paper, but our rough estimation suggests the span-level labeling of $\text{RAGTruth}_{\text{QA}}$ was approximately $3,000.

3.4 Manual Dataset Analysis

To validate the quality of our automatic annotation pipeline, we conducted a manual verification study on 100 randomly selected samples from the English PsiloQA test split. Three annotators with MS degrees in relevant fields were tasked with identifying hallucination spans within these samples, using general instruction presented in Figure 8.

Following our dual-level evaluation approach (detailed in Section 4), we assessed annotation quality using two metrics: average precision (AP) and intersection over union (IoU). For inter-annotator agreement, we computed the mean of all pairwise comparisons between annotators, yielding an AP of 80.1% and IoU of 76.8%, demonstrating substantial consensus among human annotators. To compare human labels against GPT-4o's automatic predictions, we first aggregated the three manual annotations: for IoU, we computed the character-level union, while for AP, we calculated the mean score for each character position. The aggregated reference showed strong alignment with GPT-4o's predictions, achieving an AP of 84.3% and IoU of 71.0%.

The result indicates that GPT-4o is a reliable annotator of span-level hallucination given ground-truth context. The chosen sample size yields a worst-case margin of error of approximately 9.8% at the 95% confidence level (Klie et al., 2024), providing reasonable confidence of the pipeline's overall adequacy. To further validate annotation quality, we present cross-lingual transfer results in Section 5.2 using the Mu-SHROOM (Vazquez et al., 2025) dataset.

[4] https://hf.co/facebook/bart-large-mnli

4 Experimental Setup

In our experiments, we pursue the following objectives: (i) to evaluate the performance of uncertainty quantification (UQ) baselines, large language models (LLMs), and state-of-the-art methods on the PsiloQA dataset; (ii) to demonstrate the transferability of knowledge from models trained on PsiloQA and RAGTruth to a range of downstream benchmarks; (iii) to assess the impact of multilingual training in PsiloQA by comparing two configurations of mmBERT (Marone et al., 2025): one trained on the full multilingual PsiloQA dataset and the other trained separately on each individual language subset.

4.1 Datasets

In addition to PsiloQA (described in Section 3), we employ four different QA-based benchmarks to evaluate several methods for identifying span-level hallucinations, with the dataset specifics outlined in Table 1.

Mu-SHROOM: a multilingual benchmark for 14 languages. The dataset contains 3,351 unlabeled samples in four languages: English, Spanish, Chinese, French. The test set contains 1,902 samples (Basque, Catalan, Czech and Farsi containing around 100 items, and other languages containing around 150 items).

FAVA-Bench: An English-language, human-annotated benchmark designed to identify and correct various types of hallucinations according to the FAVA taxonomy. Due to the inability of most models to detect errors using this taxonomy, the original benchmark's focus was limited to span-level hallucination detection.

HalluEntity: A benchmark in English for detecting entity-level hallucinations, comprising 157 human-annotated samples. Annotations were gathered by having ChatGPT generate biographies of various well-known individuals.

RAGTruth$_{\text{QA}}$: An English-language, human-annotated benchmark for detecting hallucinations, consisting of 900 samples with questions sourced from the MS MARCO dataset. For generating responses, six different models were utilized: `GPT-3.5-turbo-0613`, `GPT-4-0613`, `Llama-2-7B-chat`, `Llama-2-13B-chat`, `Llama-2-70B-chat` and `Mistral-7B-Instruct`.

4.2 Metrics

Due to the complex nature of hallucination detection, we employ a dual-level evaluation approach combining span-level and character-level assessment. As with Mu-SHROOM, we selected the intersection over union (IoU) metric to evaluate span-level hallucination detection. Additionally, we use average precision (AP) for ranking-based evaluation on character-level.

First, the span-level annotation is converted into a set of binary labels for each character. Next, the IoU is calculated as follows:

$$\text{IoU} = \left|\hat{C}_{\text{bin}} \cap C_{\text{bin}}\right| \,/\, \left|\hat{C}_{\text{bin}} \cup C_{\text{bin}}\right|, \quad (1)$$

where C_{bin} is the set of binarized character-level annotations, and $\hat{C}_{\text{bin}}$ is the set of characters that the model predicts as hallucinated.

Average precision provides a threshold-independent evaluation by computing the area under the precision-recall curve. In practice, it is calculated as:

$$\text{AP} = \sum_n p(n)\Delta r(n) \quad (2)$$

where $p(n)$ is the precision at cut-off n in the ranked list and $\Delta r(n)$ is the change in recall between items $n-1$ and n. This metric is particularly valuable for hallucination detection as it handles imbalanced datasets effectively and provides robust evaluation across different prediction confidence distributions, making it suitable for both in-domain evaluation and cross-dataset transfer scenarios.

4.3 Baselines

4.3.1 Uncertainty Quantification

Most uncertainty quantification methods either operate at the sequence-level or require model-specific training. Therefore, for a given generated text $\tilde{y}$ of a length N, for each token $t_i \in \tilde{y}$, $i = 1 \dots N$, we compute three uncertainty quantification methods, which are designed for token-level tasks. The set of baselines includes Maximum Token Probability (MaxProb; Fomicheva et al. (2020)), Claim Conditioned Probability (CCP; Fadeeva et al. (2024)), and Focus (Zhang et al., 2023). To compute the IoU metrics, we employ language-specific threshold calibration on the validation set.

Method	Mode	Metrics	ar	ca	cs	de	en	es	eu	fa	fi	fr	hi	it	sv	zh
Uncertainty Quantification																
MSP	–	AP	43.38	39.41	40.12	30.86	59.38	52.04	50.76	41.08	65.95	49.84	56.75	47.71	45.39	49.18
		IoU	35.70	28.36	33.68	30.03	45.69	33.72	33.04	22.13	53.13	37.67	43.45	31.61	26.96	28.42
CCP	–	AP	48.90	41.17	40.98	31.18	62.52	52.55	51.63	44.87	66.75	50.43	59.62	49.75	45.30	52.67
		IoU	35.70	28.37	33.68	33.25	45.69	33.72	33.04	22.13	53.13	37.67	43.45	32.20	26.96	27.39
Focus	–	AP	49.87	39.88	43.86	32.71	63.61	61.72	52.84	47.12	68.90	53.65	60.09	48.07	56.20	53.05
		IoU	36.93	28.37	33.68	32.05	45.69	42.24	34.65	29.94	53.13	39.26	43.45	32.20	36.15	27.83
Encoder Models																
lettuce-detect-base	–	AP	46.23	57.71	32.53	32.15	54.21	51.27	30.78	32.45	57.43	37.51	33.17	35.36	48.97	31.01
		IoU	37.81	44.37	30.08	30.31	43.28	40.08	33.35	32.45	56.44	35.60	16.95	34.97	49.11	35.94
ModernBERT-base	SFT	AP	60.37	75.48	53.46	44.77	81.63	81.71	58.72	53.84	66.87	68.48	71.94	72.00	79.94	66.84
		IoU	55.27	65.70	44.73	46.27	68.23	61.69	**50.43**	68.63	64.68	53.90	54.15	62.75	**67.09**	56.95
mmBERT-base	SFT	AP	**70.71**	**77.22**	**67.62**	61.40	**84.88**	**84.84**	65.30	**75.24**	**75.85**	**73.52**	**78.33**	**73.81**	**84.04**	73.79
		IoU	**58.10**	**67.01**	**48.81**	**54.97**	**70.67**	**66.18**	50.27	**76.61**	**68.16**	**56.38**	**61.19**	**66.57**	66.24	**61.58**
Language Models																
FActScore (GPT-4o)	–	AP	53.62	45.24	58.65	43.32	62.38	51.75	**66.82**	70.04	74.49	71.12	50.35	69.81	69.68	58.68
		IoU	20.75	28.99	10.44	26.68	25.84	28.54	19.68	26.62	28.16	10.21	21.03	43.92	19.25	25.18
Qwen2.5-32B-it	3-shot	AP	54.52	67.66	61.18	**70.50**	63.17	54.69	57.10	59.68	72.42	67.41	76.14	57.65	64.43	**77.20**
		IoU	35.54	51.71	46.83	23.57	39.98	40.51	36.52	19.18	34.69	31.92	44.56	37.95	50.89	42.77

Table 2: Performance comparison of span-level hallucination detection methods on the PsiloQA test set across 14 languages. Encoder models were supervised fine-tuned (SFT) on the complete PsiloQA train set, while Qwen2.5-32B-it used 3-shot prompting.

4.3.2 Encoder Models

We evaluate several encoder-based transformer models fine-tuned for token-level hallucination detection. These models process context-question-answer triples to identify unsupported claims at the token level. LettuceDetect models (Kovács and Recski, 2025), built on ModernBERT (Warner et al., 2025), were trained on the RAGTruth dataset and offer extended context processing capabilities (8,192 tokens) through a local-global attention mechanism (Rivière et al., 2024). However, they were pre-trained primarily on English data from various sources like web documents, code, and scientific literature. This implies that the models are not directly suitable for use with other languages, and their effectiveness may be further reduced for low-resource languages, even though the tokenizers might include subtokens pertaining to non-English languages.

To address multilingual requirements, we fine-tuned `mmBERT-base`[5], Modern Multilingual Encoder (307M parameters) that extends ModernBERT with native support for multiple languages. We also fine-tuned `ModernBERT-base`[6] on the complete PsiloQA dataset for comparison. All models were fine-tuned using identical hyperparameters as those in LettuceDetect, detailed in Table 5 (Appendix C).

4.3.3 Language Models

We also use two LLM-based approaches. FActScore (Min et al., 2023) decomposes model responses into atomic facts and verifies each against the provided context using an LLM (GPT-4o in our implementation). To adapt this sentence-level method for token-level annotation, we compute token-level hallucination scores based on their frequency in unsupported claims, applying a threshold of 0.5 for binary classification. Tokens appearing in all claims or present in the original input are excluded from hallucination marking. Few-shot prompting with `Qwen2.5-32B-Instruct`[7] using 3-shot learning with examples randomly selected from the validation set. The prompting template is detailed in Figure 11.

5 Results

5.1 Performance on PsiloQA

Table 2 presents the performance of span-level hallucination detection methods on PsiloQA across 14 languages.

Uncertainty Quantification methods show moderate performance, with Focus consistently outperforming MSP and CCP across both metrics, achieving the highest AP scores (e.g., 68.90 for Finnish, 63.61 for English) among UQ approaches. However, IoU scores remain relatively low across all languages, indicating limited precision in span-level detection.

Encoder models demonstrate superior performance, with a clear hierarchy emerging. The pre-trained LettuceDetect model shows mixed results, performing reasonably on some languages. Fine-tuned models significantly outperform the pre-trained baseline: ModernBERT achieves strong results, while mmBERT obtains the best overall per-

[5] `https://hf.co/jhu-clsp/mmBERT-base`
[6] `https://hf.co/answerdotai/ModernBERT-base`
[7] `https://hf.co/Qwen/Qwen2.5-32B-Instruct`

Strategy	Metrics	ar	ca	cs	de	en	es	eu	fa	fi	fr	hi	it	sv	zh
PsiloQA															
Per language	IoU	45.7	59.88	45.64	39.68	**72.05**	55.05	42.33	69.37	61.98	49.2	52.11	59.27	62.74	51.62
	AP	57.1	70.79	53.71	58.29	82.82	67.71	49.81	65.54	65.1	62.57	71.3	69.88	83.69	69.73
Multilingual	IoU	**58.1**	**67.01**	**48.81**	**54.97**	70.67	**66.18**	**50.27**	**76.61**	**68.16**	**56.38**	**61.19**	**66.57**	**66.24**	**61.58**
	AP	**70.71**	**77.22**	**67.62**	**61.4**	**84.88**	**84.84**	**65.3**	**75.24**	**75.85**	**73.52**	**78.33**	**73.81**	**84.04**	**73.79**
Mu-SHROOM															
Per language	IoU	47.17	52.4	31.43	38.42	**58.51**	31.27	38.11	49.62	**61.69**	53.53	62.91	61.27	33.58	31.05
	AP	67.18	64.33	51.05	60.34	70.18	44.7	51.85	68.91	**81.37**	76.93	75.91	79.07	70.52	**60.8**
Multilingual	IoU	**65.87**	**65.12**	**42.14**	**64.13**	58	**45.28**	**48.16**	**69.01**	51.69	**59.49**	**71.69**	**72.6**	**49.45**	**38.19**
	AP	**82.4**	**78.87**	**63.5**	**81.51**	**72.27**	**53.12**	**66.45**	**78.61**	78.51	**80.88**	**79.73**	**84.02**	**74.7**	56.01

Table 3: Cross-lingual transfer results comparing two training strategies of mmBERT-base: language-specific models trained independently on each language subset of PsiloQA (per language) versus a single multilingual model trained on the complete PsiloQA dataset (multilingual). Both approaches are evaluated on test sets from PsiloQA and Mu-SHROOM datasets.

Test	Metrics	Train		
		$\text{RAGTruth}_{\text{QA}}$	$\text{PsiloQA}_{\text{en}}$	Both
FAVA-Bench	IoU	14.46	14.29	**14.88**
	AP	18.55	**23.10**	17.36
HalluEntity	IoU	28.12	**30.80**	25.53
	AP	40.94	56.33	**63.37**
$\text{Mu-SHROOM}_{\text{en}}$	IoU	40.27	**58.51**	55.90
	AP	46.45	**70.18**	67.31

Table 4: Generalization performance of mmBERT-base across different hallucination detection benchmarks. Models were fine-tuned on $\text{RAGTruth}_{\text{QA}}$, $\text{PsiloQA}_{\text{en}}$, or both datasets, and evaluated on FAVA-Bench, HalluEntity, and $\text{Mu-SHROOM}_{\text{en}}$ test sets.

formance, achieving the highest scores in 12 of 14 languages for both metrics. This superiority highlights the importance of multilingual pre-training for cross-lingual hallucination detection.

Language model approaches exhibit divergent patterns. FActScore achieves competitive AP scores in certain languages (e.g., Finnish, French) but consistently shows poor IoU performance, suggesting it identifies hallucinated regions but struggles with precise span boundaries. `Qwen2.5-32B-it` with 3-shot prompting demonstrates language-specific strengths, achieving the best AP for German and Chinese, though its IoU scores remain moderate.

Takeaway

Fine-tuned multilingual encoder models consistently outperform both uncertainty-based and LLM-based approaches. But the gap between AP and IoU metrics across all methods indicates that precise span-level boundary detection remains a significant challenge.

5.2 Cross-lingual Transfer

To evaluate the cross-lingual transferability of PsiloQA, we compare mmBERT-base models trained on the complete multilingual PsiloQA dataset against models trained on individual language subsets. We evaluate performance on both PsiloQA and Mu-SHROOM test sets to measure within-distribution and cross-dataset generalization.

Table 3 demonstrates that training on the full multilingual PsiloQA dataset enables robust cross-lingual transfer. The multilingual model consistently outperforms language-specific models across most target languages, with improvements observed even for languages with distinct scripts (Arabic, Hindi) and those from different language families.

Takeaway

Multilingual training on PsiloQA enables superior cross-lingual transfer compared to language-specific training, with benefits extending across different scripts and language families.

5.3 Knowledge Transfer

To evaluate generalization across datasets, we assess model transferability (Karpov and Konovalov, 2023) by comparing models fine-tuned on synthetic PsiloQA versus human-annotated RAGTruth. We utilize `mmBERT-base` as our strongest baseline. We fine-tune it in three configurations: (i) exclusively on $\text{RAGTruth}_{\text{QA}}$, (ii) solely on $\text{PsiloQA}_{\text{en}}$, and (iii) on both datasets combined. All configurations use identical hyperparameters and are evaluated on three independent benchmarks: FAVA-Bench, HalluEntity, and $\text{Mu-SHROOM}_{\text{en}}$.

Table 4 reveals that PsiloQA$_{en}$-trained models consistently outperform RAGTruth$_{QA}$ across benchmarks. Most notably, PsiloQA achieves substantial gains on Mu-SHROOM$_{en}$ – representing a 45% improvement in IoU. Similar advantages appear on HalluEntity, while FAVA-Bench shows limited performance across all configurations, suggesting valuable task differences. Combining both datasets demonstrates selective benefits, with PsiloQA-based configurations (alone or combined) dominating. Joint training achieves the highest AP on HalluEntity. On Mu-SHROOM$_{en}$, the combined model maintains strong performance, ranking second only to PsiloQA alone.

The advantage of PsiloQA$_{en}$ might stem from its larger training set size, yet PsiloQA's synthetic generation remains more than 17 times cheaper than manually curated RAGTruth$_{QA}$ (Section 3).

Takeaway

The superior ability of PsiloQA to transfer knowledge could be attributed to its larger size compared to RAGTruth, yet its generation cost is significantly lower than the labeling cost of RAGTruth.

Conclusion

In this work, we introduced PsiloQA, a large-scale, multilingual, span-level hallucination detection dataset constructed using a scalable and cost-effective pipeline. Our approach leverages real hallucinations produced by LLMs in a zero-context setting and employs GPT-4o for automated span-level annotations. PsiloQA offers extensive language coverage and supports diverse LLM architectures, providing a valuable resource for evaluating and training hallucination detection models.

Through comprehensive evaluations across multiple baselines – including uncertainty quantification, encoder-based detectors, and LLM-based methods – we demonstrated the effectiveness of PsiloQA for benchmarking hallucination detection. Our results reveal that fine-tuned encoder-based methods, particularly multilingual models like mmBERT and ModernBERT, outperform other baselines, though significant challenges remain. Additionally, we showed that PsiloQA supports strong cross-lingual generalization and outperforms human-annotated datasets like RAGTruth in knowledge transfer experiments, despite being over 17 times cheaper to produce.

Our findings highlight the feasibility and advantages of using synthetically generated datasets with automated high-quality annotations to improve the robustness and factuality of LLMs. Future work will explore extending the PsiloQA pipeline to other generation tasks, such as summarization and data-to-text generation, further broadening its utility in hallucination research.

Limitations

While PsiloQA presents significant advancements in span-level hallucination detection across languages, several limitations remain:

Annotation Source Bias: PsiloQA relies exclusively on GPT-4o for both generating question–answer pairs and annotating hallucination spans. This introduces potential bias in annotation and generation patterns, as the judgment of a single model may not reflect broader consensus or generalize well across diverse use cases. This bias could be substantially mitigated by using an ensemble of annotators composed of several state-of-the-art models with span averaging. We consider this a promising direction for future work.

Task Narrowness: The current version of PsiloQA is limited to the question-answering (QA) task. While QA is a strong proxy for factual reasoning, other generative tasks such as summarization, dialogue, and data-to-text generation also suffer from hallucinations and warrant similar treatment.

Hallucination Type Coverage: Unlike datasets that inject controlled hallucination types (e.g., FAVA), PsiloQA does not explicitly cover a diverse taxonomy of hallucinations. The hallucinations in PsiloQA arise naturally from LLM errors in a zero-context setting, which may result in skewed distributions and underrepresentation of certain error types.

Language Resource Imbalance: Despite covering 14 languages, the sample distribution across languages is uneven, and lower-resource languages may suffer from fewer high-quality examples. Additionally, many baselines used for comparison are predominantly trained or optimized for English, potentially underestimating performance in other languages.

Dependency on Wikipedia: Using Wikipedia as the sole source of context limits the topical, stylistic, and cultural diversity of the dataset. While Wikipedia provides clean, factual content across

many languages, its coverage is uneven: some languages, cultures, and topics are better represented than others, potentially introducing cultural or regional biases into the dataset. Consequently, models trained on this data may inherit these biases. Moreover, real-world applications often involve noisier or domain-specific data

Ethical Considerations

This work involves the creation and analysis of a multilingual question-answering (QA) dataset, PsiloQA, designed for evaluating span-level hallucination detection in large language models (LLMs). We address several ethical aspects to ensure responsible data generation, annotation, and usage:
Data Source and Privacy: All questions and corresponding answers were generated using publicly available information from Wikipedia. The dataset contains no personally identifiable information (PII), and no sensitive or private data was collected, stored, or processed at any stage.
Intended Use and Limitations: PsiloQA and the associated models are intended solely for research purposes, particularly in the development of more trustworthy and interpretable QA systems. While our models aid in detecting hallucinations, they are not error-free and should not be considered reliable for deployment in high-stakes or real-time decision-making systems (e.g., healthcare, legal domains) without rigorous domain-specific evaluation and validation.
Fairness and Inclusivity: The dataset includes samples across 14 languages to support multilingual research. However, language coverage is uneven, and performance disparities may exist due to varying model training resources and language-specific complexities. Researchers should account for these disparities when interpreting results or deploying tools based on PsiloQA.
Avoidance of Misuse: We explicitly discourage the use of our dataset or trained models for surveillance, censorship, or automated moderation without human oversight. The tools are not intended for identifying or suppressing content and must not be used to enforce ideologically biased or discriminatory practices.
Transparency and Reproducibility: We provide complete documentation of our data generation and annotation pipeline, including prompt designs and filtering mechanisms, to ensure transparency and reproducibility. Our approach emphasizes the use of real hallucinations generated by LLMs in a zero-context setting, promoting authentic error analysis.
Model Dependency and Bias: Since both the generation and annotation of PsiloQA rely on GPT-4o, there is an inherent risk of model bias influencing the dataset. Although GPT-4o was among the state-of-the-art models available during dataset development, its judgments may reflect underlying model biases or fail to align with human consensus in edge cases. Furthermore, GPT-4o's proficiency varies across languages, which may affect the consistency and quality of cross-lingual annotations. Future iterations of PsiloQA may incorporate diverse model perspectives and human-in-the-loop validation to mitigate this concern.

By making PsiloQA publicly available, we aim to support the development of robust, multilingual hallucination detection systems while promoting ethical, fair, and responsible AI research.

Acknowledgements

The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the R.F. in accordance with the agreement 000000C313925P4F0002 and the agreement with Skoltech №139-10-2025-033.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *CoRR*, abs/2307.15703.

Julia Belikova, Evegeniy Beliakin, and Vasily Konovalov. 2024. JellyBell at TextGraphs-17 shared task: Fusing large language models with external knowledge for enhanced question answering. In *Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing*, pages 154–160, Bangkok, Thailand. Association for Computational Linguistics.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

Sky CH-Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. 2024. Do androids know they're only dreaming of electric sheep? In *Findings of the Association for Computational Linguistics: ACL 2024*,

pages 4401–4420, Bangkok, Thailand. Association for Computational Linguistics.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436, Miami, Florida, USA. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 9367–9385. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nat.*, 630(8017):625–630.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Yuzhe Gu, Ziwei Ji, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. Anah-v2: Scaling analytical hallucination annotation of large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Inf. Fusion*, 118:102963.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55.

Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. ANAH: Analytical annotation of hallucinations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8135–8158, Bangkok, Thailand. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *CoRR*, abs/2207.05221.

Dmitry Karpov and Vasily Konovalov. 2023. Knowledge transfer between tasks and languages in the

multi-task encoder-agnostic transformer-based models. In *Computational Linguistics and Intellectual Technologies*, volume 2023.

Jan-Christoph Klie, Juan Haladjian, Marc Kirchner, and Rahul Nair. 2024. On efficient and statistical quality estimation for data annotation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15680–15696, Bangkok, Thailand. Association for Computational Linguistics.

Ádám Kovács and Gábor Recski. 2025. Lettucedetect: A hallucination detection framework for RAG applications. *CoRR*, abs/2502.17125.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Junyi Li, Xiaoxue Cheng, 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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6449–6464. Association for Computational Linguistics.

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 (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang 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, Singapore. Association for Computational Linguistics.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *First Conference on Language Modeling*.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 80 others. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Elisei Rykov, Valerii Olisov, Maksim Savkin, Artem Vazhentsev, Kseniia Titova, Alexander Panchenko, Vasily Konovalov, and Julia Belikova. 2025. SmurfCat at SemEval-2025 task 3: Bridging external knowledge and model uncertainty for enhanced hallucination detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1034–1045, Vienna, Austria. Association for Computational Linguistics.

Artem Shelmanov, Evgenii Tsymbalov, Dmitri Puzyrev, Kirill Fedyanin, Alexander Panchenko, and Maxim Panov. 2021. How certain is your transformer? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1833–1840, Online. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy

Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Artem Vazhentsev, Ekaterina Fadeeva, Rui Xing, Alexander Panchenko, Preslav Nakov, Timothy Baldwin, Maxim Panov, and Artem Shelmanov. 2024. Unconditional truthfulness: Learning conditional dependency for uncertainty quantification of large language models. *CoRR*, abs/2408.10692.

Artem Vazhentsev, Lyudmila Rvanova, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2025. Token-level density-based uncertainty quantification methods for eliciting truthfulness of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2246–2262, Albuquerque, New Mexico. Association for Computational Linguistics.

Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona De Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2472–2497, Vienna, Austria. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 2526–2547. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.

Ziwei Xu, Sanjay Jain, and Mohan S. Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *CoRR*, abs/2401.11817.

Min-Hsuan Yeh, Max Kamachee, Seongheon Park, and Yixuan Li. 2025. Halluentity: Benchmarking and understanding entity-level hallucination detection. *Transactions on Machine Learning Research*.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. LUQ: Long-text uncertainty quantification for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262, Miami, Florida, USA. Association for Computational Linguistics.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 915–932. Association for Computational Linguistics.

A License and Infrastructure

Experiments utilized 2 NVIDIA A100 GPUs, totaling around 50 GPU-hours. Models were used according to their licenses: Qwen 2.5 under Apache 2.0. We release our dataset under CC-BY-4.0.

B Packages

To generate the PsiloQA dataset, we used the VLLM (Kwon et al., 2023) package for efficient inference of LLMs on available GPUs, following the default hyperparameters recommended in the Hugging Face README files. Encoder model training was performed using the Transformers library (Wolf et al., 2020), with training hyperparameters detailed in Appendix C.

C Encoder Hyperparameters

Hyperparameter	Value
learning_rate	1e-5
num_train_epochs	6
weight_decay	0.01
batch_size	8

Table 5: Training followed LettuceDetect (Kovács and Recski, 2025) hyperparameters for six epochs, with the best validation checkpoint selected.

D PsiloQA Dataset Statistics

Language	Language Model	# of parameters	Avg # of spans	Avg span length	# of samples
ar	SeaLLMs/SeaLLM-7B-v2.5	7-9B	1.35	9.09	2072
ca	occiglot/occiglot-7b-es-en-instruct	7-9B	1.07	11.18	6240
cs	mistralai/Mistral-7B-Instruct-v0.3	7-9B	1.60	16.18	4984
de	malteos/bloom-6b4-clp-german-oasst-v0.1	3-7B	0.97	9.96	1378
en	HuggingFaceH4/zephyr-7b-beta	7-9B	2.41	22.46	665
en	HuggingFaceTB/SmolLM2-1.7B-Instruct	1-3B	2.20	21.36	608
en	HuggingFaceTB/SmolLM2-135M-Instruct	<1B	2.09	36.80	581
en	HuggingFaceTB/SmolLM2-360M-Instruct	<1B	1.85	25.97	578
en	ServiceNow-AI/Apriel-5B-Instruct	3-7B	1.37	10.66	3525
en	TinyLlama/TinyLlama-1.1B-Chat-v1.0	1-3B	1.87	21.46	2151
en	tiiuae/falcon-7b-instruct	7-9B	1.70	14.96	1595
en	togethercomputer/Pythia-Chat-Base-7B-v0.16	7-9B	1.29	9.31	2042
es	Iker/Llama-3-Instruct-Neurona-8b-v2	7-9B	1.43	14.84	2364
eu	google/gemma-7b-it	7-9B	1.03	10.81	3853
fa	Qwen/Qwen2-7B-Instruct	7-9B	1.22	7.46	4550
fi	BSC-LT/salamandra-7b	7-9B	0.83	2.37	4512
fi	Finnish-NLP/llama-7b-finnish-instruct-v0.2	7-9B	1.09	9.38	2561
fr	croissantllm/CroissantLLMChat-v0.1	1-3B	1.56	26.84	2026
hi	google/gemma-7b-it	7-9B	1.03	10.81	3853
hi	nickmalhotra/ProjectIndus	1-3B	1.25	25.65	1801
hi	sarvamai/sarvam-1	1-3B	0.98	11.67	3331
it	sapienzanlp/modello-italia-9b	7-9B	1.39	15.13	2181
sv	utter-project/EuroLLM-9B-Instruct	7-9B	1.12	11.27	7729
zh	Qwen/Qwen2-7B-Instruct	7-9B	1.22	7.46	4550
zh	Qwen/Qwen2.5-3B-Instruct	3-7B	1.11	1.46	1170
zh	ikala/bloom-zh-3b-chat	3-7B	1.28	3.04	3309

Table 6: The list includes the utilized LLMs along with their corresponding languages and statistics like the average number of spans and the average span length.

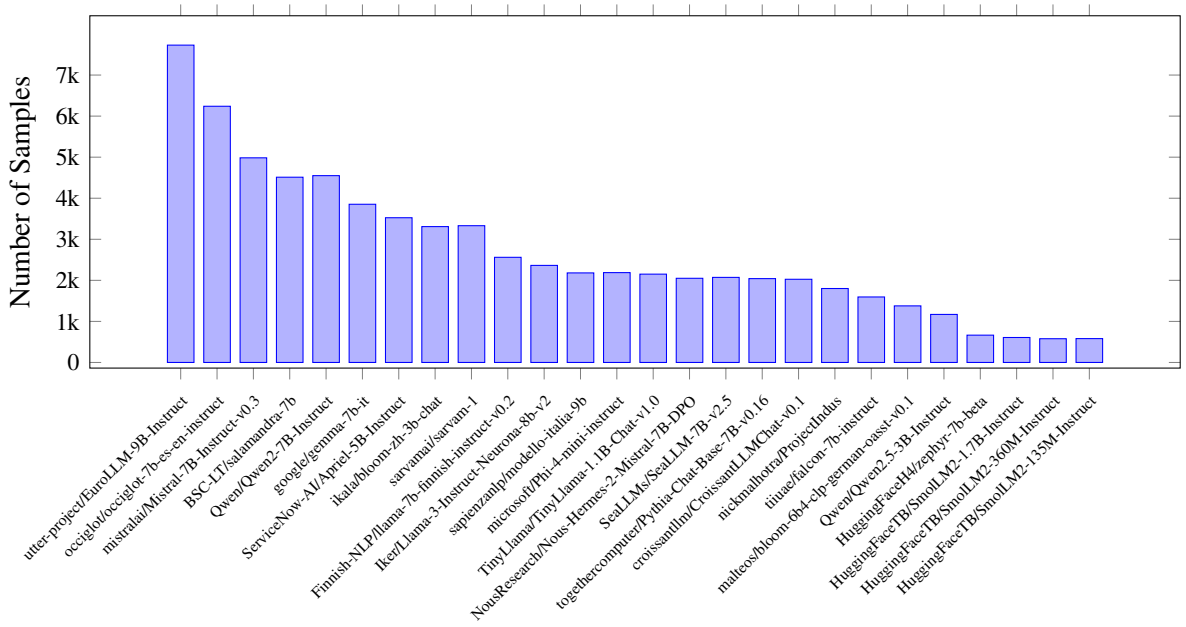

Figure 2: Number of samples generated by every model in PsiloQA.

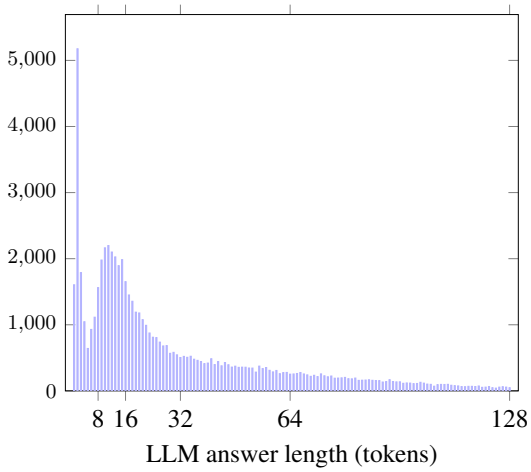

Figure 3: Distribution of LLM answer lengths in PsiloQA, measured in tokens using spaCy.

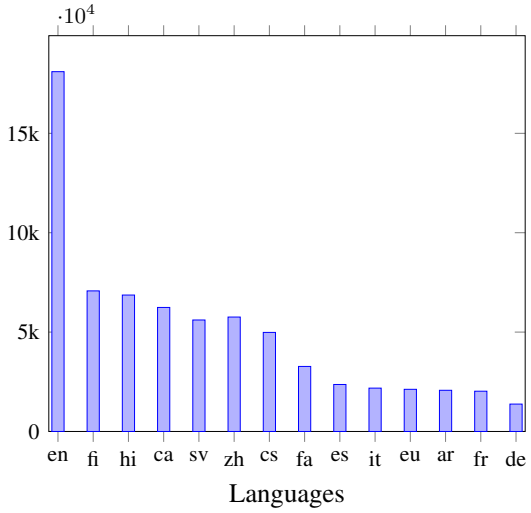

Figure 4: Distribution of languages in PsiloQA dataset.

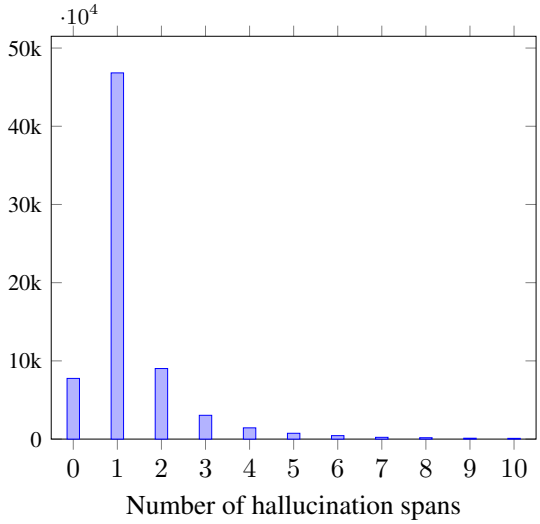

Figure 5: Distribution of hallucination spans in PsiloQA.

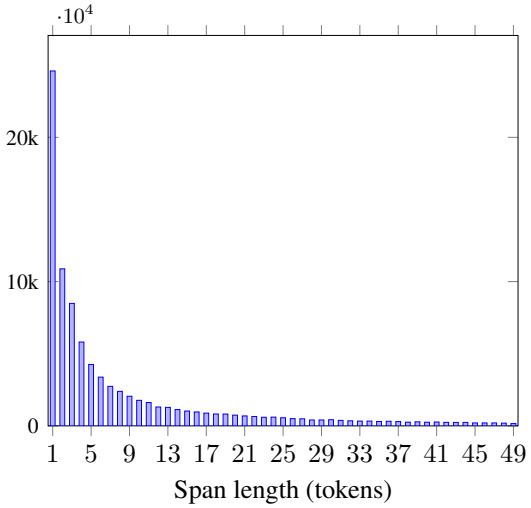

Figure 6: Distribution of hallucination span lengths in PsiloQA, measured in tokens using spaCy.

E PsiloQA Dataset Sample

Passage	Question	Complexity	Golden Answer	LLM Answer
Lillian Richter (1915–2000) was an American lithographer. Richter did work for the Works Progress Administration (WPA).	When was Lillian Richter born?	Easy	1915	Lillian Richter was born in 1932.
Things I Carry Around is the eleventh studio album by Australian country music artist Troy Cassar-Daley.	What is the title of the eleventh studio album by Troy Cassar-Daley?	Easy	Things I Carry Around	The eleventh studio album by Troy Cassar-Daley is "The Greatest Hits"
Albert Herman (1887–1958) was an American actor, screenwriter and film director. Herman was a prolific director, working mainly on low-budget movies for companies such as Producers Releasing Corporation	For which company did Albert Herman mainly work as a director?	Hard	Producers Releasing Corporation	Albert Herman mainly worked as a director for the company Albert Herman & Cie.
Edmond Armand René Thorailler, né le 17 juin 1911 à Nogent-le-Roi (Eure-et-Loir) et mort le 24 août 2004 à Notre-Dame-de-Monts (Vendée), est un notaire et homme politique français.	Quelle est la date de naissance d'Edmond Thorailler?	Easy	17 juin 1911	1923.
Bloomfield Road est le nom d'un stade de football localisé à Blackpool, en Angleterre. C'est l'enceinte du club principal de la ville, le Blackpool Football Club, depuis 1901. Ce stade de 16 220 places a été inauguré le 28 octobre 1899.	Quel est le nom du stade situé à Blackpool, en Angleterre?	Easy	Bloomfield Road	Blackpool Stadium
Der Rotstielige Zwerg-Täubling oder Rotstieliger Birken-Täubling (Russula font-queri) ist ein Pilz aus der Familie der Täublingsverwandten. Es ist ein seltener Täubling und typischer Birkenbegleiter mit kupferrotem Hut und leicht rosa überlaufenem Stiel. Sein Sporenpulver ist gelblich.	Wie heißt der Rotstielige Zwerg-Täubling auf Latein?	Easy	Russula font-queri	Amanita rubescens-Team

Table 7: Example of PsiloQA samples. Each sample contains the passage retrieved from Wikipedia, the question and the Golden Answer of some complexity level generated by GPT-4o, the LLM answer generated by some LLM, and the hallucination ranges (highlighted in red) annotated by GPT-4o by comparing the LLM answer with the Golden Answer.

F QA Pairs Generation Prompt

```
Generate 3 question-answer pairs with different levels of complexity (easy,
    medium and hard).
You must only create questions that require knowledge of the passage.
Format your answer as a Python list with 3 jsons.
Each json should contain "question", "answer" and "complexity" fields.

Do not generate questions that require imagination.
The questions should be factual, so do not generate questions that ask for
    subjective opinion or reasoning.
It should be enough to know the facts from the provided passage.
Each question must contain exactly one question.
Do not make any reference to the passage in the question-answer pair.

Use the language in which the passage is written.

{answer_length_constraint}

**Passage:** {p}
```

Figure 7: Prompt for question-answer pairs generation used for PsiloQA creation. We ask GPT-4o to generate three different question-answer pairs of different complexity using the retrieved passage from Wikipedia. We also control the length of the answer. In 33% of the cases we ask GPT-4o to generate long and detailed answers.

G Inconsistency Detection Prompt

```
Your task is to find any inconsistencies with the correct information in LLM's
    answer.
Carefully read the user's question, the golden answer, the relevant passage, and
    LLM's answer.
The relevant passage contains information for answering the question.
LLM did not see the relevant passage while generating the response.

General instructions:
- If LLM refused to answer the question, then answer by simply copying LLM's
    answer. There is no inconsistency if LLM did not provided any answer to the
    question.
- If LLM's answer does not contain information relevant to the question, and the
    information does not contradict the relevant passage, then answer by simply
    copying LLM's answer.
- If LLM's answer is relevant to the question, use the opening tag [HAL] and
    closing tag [/HAL] to highlight areas of inconsistency with the golden
    answer and relevant passage information. Inconsistency means something that
    is related to the topic of the question, but contradicts the relevant
    passage or introduces new information.
- If LLM's answer is consistent with the golden answer, do not highlight
    anything, answer by simply copying LLM's answer.

How to highlight spans:
- Each span could contain from 1 to several words. In rare, specific cases, it
    could be longer.
- Make spans as precise as possible, do not highlight the entire answer.
- Highlight spans at the word level, not the character level.

DO NOT ADD ANY CHANGES TO LLM'S ANSWER EXCEPT [HAL] and [/HAL]!

Begin your answer with "**Highlighted LLM Response:**".
```

Figure 8: Prompt for inconsistencies detection in LLM answers. We pass to prompt questions, answers of different LLMs and the golden answers generated by GPT-4o. In the prompt, we ask GPT-4o to find any spans of inconsistencies in the golden answer. Here we consider the obtained inconsistencies with the gold and the context as hallucinations.

H Filtering Prompts

```
Classify the following question into one of three categories:
1. INCOMPLETE_QUESTION
Incomplete questions that use pronouns with no clear antecedent ("Where is THIS
    located?" / "What is HE famous for?").
Questions that refer to a list/paragraph/excerpt that isn't included in the
    question itself.

2. SUBJECTIVE
Questions whose answers require value judgments or a subjective opinion.

3. NORMAL
Any question with a clear subject ("Where is Jacksonville located?" / "Who is
    Alexander Montenegro?").
The subject may be ambiguous or polysemous, but it must not be impersonal.
Open-world questions that name their subject (even if broad or multi-answer) are
    NORMAL.

Instructions:
Output only one of the three categories: INCOMPLETE_QUESTION, SUBJECTIVE, or
    NORMAL.

Examples:
INCOMPLETE_QUESTION
Q: What were the set scores in the final match?
A: INCOMPLETE_QUESTION
Q: Which religious order was the abbey associated with?
A: INCOMPLETE_QUESTION
Q: What is the chemical formula of perrhenic acid as stated in the passage?
A: INCOMPLETE_QUESTION

SUBJECTIVE
Q: What are some of the primary aims of the Athletics Club Lechia Gdansk?
A: SUBJECTIVE
Q: What is the significance of the Fahey-Murray ministry in the context of New
    South Wales government history?
A: SUBJECTIVE

NORMAL
Q: What was the population of Hazleton according to the 2020 census?
A: NORMAL
Q: What role did William Arrington hold in the Illinois State Senate between
    1955 and 1973?
A: NORMAL
Q: Which Olympic Games did Lee Jun-ho represent South Korea in?
A: NORMAL
Q: How many league appearances did James Ryan make in the EFL?
A: NORMAL
```

Figure 9: Prompt for the detection of subjective and incomplete questions. Subjective questions require a personal opinion rather than a factual answer. Incomplete questions lack a clear subject and are therefore unanswerable.

```
You are a strict detector of 'I don't know' style answers.
Given an answer string, return only the word TRUE if the answer expresses
    refusal to answer due to lack of information, or otherwise clearly states
    that it cannot answer.
If the model provides any meaningful information regarding the topic or makes
    assumptions, return only the word FALSE, even if there was uncertainty
    anywhere in the answer.
Do not explain.
```

Figure 10: Prompt for detecting cases when LLM refuses to answer.

I LLM Baseline Evaluation Prompt

```
Please act as an objective error detector. You will be given the user's
    question, a relevant passage, and the LLM's response.
Your task is to analyze the response to the question and identify the parts of
    the response that are most likely to contain errors or inconsistencies.
Taking into account the question and the answer provided by the language model,
    you need to wrap erroneous words/phrases in the answer with special tokens:
    [HAL] erroneous word/phrase  [/HAL].
Before the question, you will be given a passage - a factual reference that will
    help you identify factual errors. Use it as a guide when highlighting
    hallucinations - mark words/phrases as hallucinations if they do not
    correspond to the information in the passage.

Example {x3}:
Knowledge source: {sampled passage (optional)}
Question: {sampled question}
Answer: {sampled answer}
Answer with highlighted spans: {sampled highlighted answer}

Knowledge source: {passage}
Question: {question}
Answer: {answer}
Answer with highlighted spans:
```

Figure 11: Prompt for evaluating the baseline model `Qwen2.5-32B-instruct` in English. We evaluate baseline in 3-shot mode, with examples sampled from the PsiloQA validation set. The prompt is translated for each language, and the few-shot examples are picked from the corresponding language.