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] Julia Belikova[4,1]

Hallucinations Undermine Trust

Large Language Models (LLMs) frequently produce **unsupported or fabricated facts** – so-called **hallucinations**. Reliable detection of such misinformation requires **robust and scalable benchmarks**, yet existing ones fall short:

- ✗ **Expensive & slow** – rely on manual human annotation
- ✗ **Coarse labels** – only judge entire answers as correct or wrong
- ✗ **English-centric** – lack multilingual coverage

Why PsiloQA Fills the Gap?

- 🌍 **Multilinguality:** Covers **14 languages**, each with its own set of LLMs
- 🤖 **Model Variety:** Authentic hallucinations from **26 different LLMs**
- 📈 **Scale: 66,000+ samples** (63,792/2,208 train/test)
- 📝 **Annotation:** Span-level, validated by ~**80% human agreement**
- 💵 **Cost:** Created for ~**535$**, proving an affordable, scalable alternative

Dataset	Domain	Annot.	Gen.	Lang	# LLMs	Train	Test
Mu-SHROOM	General	Human	Real	Multi	38	3,351*	1,902
HalluEntity	Biography	Human	Real	EN	1	–	157
RAGTruth-QA	General	Human	Real	EN	6	5,034	900
FAVA-Bench	General	LLM	Synth	EN	3	–	902
PsiloQA (ours)	General	LLM	Real	Multi	24	63,792	2,897

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 (synth) or used as-is (real)

Performance on PsiloQA

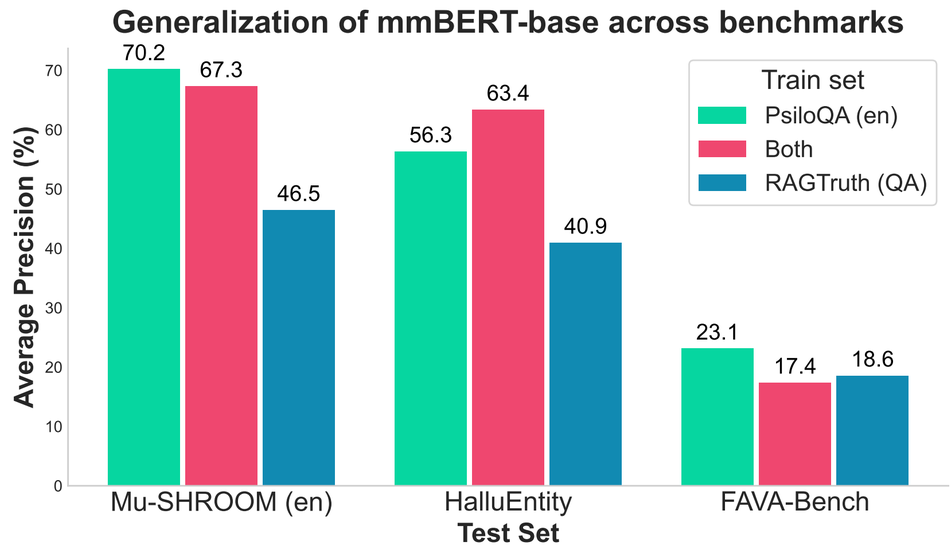
Evaluation metrics: **AP (Average Precision)** at character-level, **IoU (Intersection over Union)** at span-level.

Encoder models dominate, with mmBERT-base achieving **74.8 AP / 62.3 IoU**, significantly outperforming:

- **LLM-based methods:** Best is Qwen2.5-32B (64.6 AP / 38.3 IoU)
- **Uncertainty Quantification:** Best is Focus (52.3 AP / 36.8 IoU)

Transfer to External Benchmarks

mmBERT-base **trained on PsiloQA** outperforms RAGTruth-trained models on Mu-SHROOM (en), HalluEntity, and FAVA – achieving **+45% AP and +18% IoU** on Mu-SHROOM while being **17× cheaper**.



Cross-Lingual Transfer

Multilingual training (single mmBERT-base model on 14 languages) outperforms language-specific training (separate models per language) – achieving **+8–10 AP and higher IoU** across all languages.

PsiloQA Generation Pipeline

Step 1: QA Pairs Generation

Retrieve a random passage from Wikipedia and generate QA pairs using the passage

Passage: Japan Airlines (JAL) is the flag carrier of Japan. After over three decades of service and expansion, the airline was fully privatised in 1987.

GPT-4o with Generation Prompt

Question: In what year was Japan Airlines fully privatised?
Golden Answer: 1987

Step 2: Hypothesis Generation

Pass question to LLM without supporting passage to get answer with hallucinations

Question: In what year was Japan Airlines fully privatised?

Open Source LLMs

LLM Answer: JAL was fully privatised in 1989

Step 3: Inconsistency Detection

Pass the passage, the question, the golden answer and LLM's hypothesis to GPT-4o to find any inconsistencies

Passage: Japan Airlines (JAL) is the flag carrier of Japan. The airline was fully privatised in 1987.
Question: In what year was Japan Airlines fully privatised?
Golden Answer: 1987
LLM Answer: JAL was fully privatised in 1989

GPT-4o with Detection Prompt

Annotation: JAL was fully privatised in [HAL]1989[/HAL]

Step 4: Dataset Filtration

Filter incomplete and subjective questions and cases when LLM refuses to answer

Question: What is the new name of the museum after its relocation?

Question: Who directed the 1961 epic film “Francis of Assisi”?

LLM Answer: I couldn't find any information about Li Chengjiang.

Question: Discuss the significance of Helmut Reichelt's contributions.

GPT-oss-120B with Filtration Prompt

- Incomplete question
- Normal
- Refuse
- Subjective

Skoltech[1] AIRI[2] MTS AI[3] Sber AI Lab[4] MIPT[5]

Fall into ML 2025