Out-of-distribution robustness in multilingual instruction-tuned models

Jakub Kozanyi and **Piotr Grabowski**
jkoz@itu.dk
piog@itu.dk

Abstract

1 Introduction

Since the introduction of **ChatGPT** (**?**), interest in **large language models (LLMs)** has surged, driving an AI revolution that reshapes industries, communication, and knowledge access (**?**). However, most state-of-the-art LLMs are predominantly trained on English data, leading to **linguistic biases** that disadvantage speakers of other languages (**??**). This imbalance results in significantly better performance in English compared to other languages (**?**), reinforcing structural inequalities in AI accessibility (**??**). Consequently, non-native English speakers face reduced accuracy and reliability when interacting with these models, exacerbating the **digital language divide** (**?**).

To address this gap, **multilingual LLMs** such as **Aya101** (**?**) and **EuroLLM** (**?**) have been developed with the goal of providing more inclusive language coverage. While these models claim robust performance across languages, their actual capabilities—particularly in **underrepresented languages like Polish and Slovak**—remain underexplored (**??**).

In this paper, we evaluate the **out-of-distribution (OOD) robustness** (**?**) of multilingual LLMs in Polish and Slovak, focusing on **prompt perturbation in a question-answering setting** (**?**). Our results reveal [summary of key findings], shedding light on the strengths and limitations of current multilingual LLMs and the challenges of ensuring **equitable AI performance across languages** (**?**).

2 Datasets

To evaluate the robustness of multilingual QA models, we construct two datasets in Slovak and Polish, ensuring both in-distribution (ID) and out-of-distribution (OOD) evaluation settings. In-distribution (ID) evaluation refers to testing a model on data that is drawn from the same distribution as its training data, ensuring that the model encounters familiar patterns. Out-of-distribution (OOD) evaluation, on the other hand, assesses the model's performance on data that differs from its training distribution, which helps measure its generalization ability. Evaluating multilingual QA models under both settings is crucial to understand their robustness, ensuring they perform well not only on known data but also in real-world scenarios with unexpected linguistic variations. The datasets are carefully annotated and validated for fluency, language fidelity, and grammatical correctness.

2.1 Dataset Construction

We use two types of QA datasets:

- **Out-of-Distribution (OOD) Dataset: SK-QuAD** (**?**) and Po-QuAD (**?**)
  Extracted from Slovak and Polish Wikipedia articles, ensuring no overlap with the training data of common QA models. This dataset is human-translated to maintain high-quality, natural linguistic expressions. We use it to test the model's generalization ability to unseen, authentic, and well-formed QA data.

- **In-Distribution (ID) Dataset: MULTIQ** (**?**)
  Machine-translated from an existing multilingual QA dataset, MULTIQ. Due to its machine translation origins, some questions and answers may exhibit translation artifacts. We use it to evaluate how well models handle data they are likely to have seen or been trained on.

2.2 Validation of datasets

To ensure the quality of the datasets, we picked and manually evaluated 100 MultiQ Slovak questions, 100 MultiQ Polish questions, 100 SK-QuAD

question-answer pairs, and 100 Po-QuAD question-answer pairs. We used the following evaluatiuon metrics:

- **Fluency:** Ensuring the text reads naturally in Slovak and Polish, without awkward phrasing or unnatural expressions.
- **Grammatical Correctness:** Verifying proper sentence structure, spelling, and punctuation.
- **Correct answer(SQuAD only):** Checking whether the answer matched to the question is correct.
- **Correct translation(MultiQ only):** Checking whether translations retain the original meaning without introducing semantic shifts.

3 Experimental Setup

To evaluate the robustness of multilingual QA models, we use the two constructed datasets from the **Datasets** section and apply controlled perturbations to measure performance degradation using different metrics for evaluation.

3.0.1 Models

We test the following models:

- **Aya-101** (**?**): A 101-language instruction-tuned model.
- **EuroLLM** (**?**): Optimized for European languages.
- **+more models maybe**

3.0.2 Datasets & Perturbations

We use the two constructed datasets with systematic prompt perturbations (Table 2):

- **OOD Data**: SK-QuAD (Slovak) and Po-QuAD (Polish), human-translated from Wikipedia.
- **ID Data**: MULTIQ, a machine-translated dataset where some quality degradation may occur compared to human-translated text.

3.1 Evaluation Metrics

We employ four complementary metrics:

3.1.1 Answer Accuracy

F1 Score: Token-level precision/recall for partial matches.

3.1.2 Semantic Robustness

BERTScore : Semantic similarity using BERT embeddings.
BLEURT : Learned metric for generation quality.

3.1.3 Language Fidelity

Proportion of responses in the target language (vs. English), measured using `fastText`, an efficient open-source library for learning word representations and text classification, developed by Facebook AI Research. It enriches word vectors with subword information, making it especially useful for morphologically rich languages and rare words (**?**).

3.1.4 Manual Evaluation

We score answers on a scale of 1–5 scale for **Fluency**, which defines naturalness of language. We will also manually evaluate **Correctness**, which defines factual accuracy.

3.2 Procedure

1. **Baseline Evaluation**: Run models on unperturbed ID/OOD data.
2. **Perturbation Testing**: Apply perturbations and remeasure metrics.
3. **Analysis**:
   - Compute performance drop: $\frac{OriginalEM - PerturbedEM}{OriginalEM} \times 100$.
   - Track language switch rates.
   - Compare semantic drift via BERTScore/BLEURT.

3.3 Research Questions

To synthesize insights, we analyze:

- Does *Aya101* perform better than *EuroLLM* across ID and OOD datasets?
- Is the performance drop larger for OOD data, indicating weaker generalization?
- Which perturbation type is most damaging, and why?

This evaluation framework provides a comprehensive analysis of QA model robustness in multilingual settings, helping to uncover key failure points and areas for improvement.

4 Experiment Results

W.I.P.

Dataset	Evaluation Criteria	Score
Slovak MultiQ	Fluency	88%
	Grammatical Correctness	97%
	Correct Translation	91%
Polish MultiQ	Fluency	95%
	Grammatical Correctness	98%
	Correct Translation	96%
SK-QuAD	Fluency	95%
	Grammatical Correctness	93%
	Correct Answer	99%
Po-QuAD	Fluency	99%
	Grammatical Correctness	73%
	Correct Answer	99%

Table 1: Evaluation of Slovak and Polish datasets based on fluency, grammatical correctness, and correctness of answers or translations.

Perturbation Type Example	Description
Synonym Substitutions "Who is the **leader** of Poland?" → "Who is the **chief** of Poland?"	Replace words with synonyms to test lexical robustness
Typographical Errors "What is the **capittal** of Slovakia?"	Introduce spelling mistakes to simulate input noise
Rewording Questions "When was Napoleon born?" → "What is Napoleon's birth year?"	Restructure while preserving meaning
Code-Switching "Jaki jest **capital** Polski?"	Mix English words in Polish/Slovak prompts

Table 2: Types of prompt perturbations used to evaluate model robustness.