Evaluation of Multi-Model Translation Strategies for Lithuanian Medical Multimodal Dataset Construction

Augustė Raišytė[1], Dalius Matuzevičius[2], Tadas Žvirblis[3], Linas Petkevičius[1]
[1]Vilnius University, [2]Vilnius Gediminas Technical University, [3]National Cancer Institute
Corresponding author: auguste.raisyte@mif.vu.lt

Artificial Intelligence Methods Lab

Follow us: aiml.lt

Abstract

This study investigates the influence of different translation approaches on vision–language models for medical visual question answering in Lithuanian. PathVQA, VQA-RAD and SLAKE are translated using three large language models, producing three individual and one merged dataset variant. Individual test-set translations are compared with human-reviewed and corrected Lithuanian references. MedGemma 4B and Gemma 4B are fine-tuned separately on each training variant using Quantized Low-Rank Adaptation (QLoRA). The eight configurations are evaluated on the same Lithuanian human-reviewed test set to assess downstream Visual Question Answering (VQA) performance.

Problem

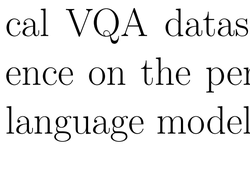
Most vision–language models and medical VQA datasets are developed for English, limiting their applicability to lower-resource languages such as Lithuanian. Large language models offer a practical solution for constructing Lithuanian medical datasets through automatic translation. However, different translation models and strategies may produce translations of varying quality, potentially affecting the performance of fine-tuned vision–language models.

Objective: To evaluate different translation approaches for constructing Lithuanian medical VQA datasets and investigate their influence on the performance of fine-tuned vision–language models.

Datasets

Three English medical VQA datasets are used: PathVQA, VQA-RAD, and SLAKE. Questions and answers are translated into Lithuanian. To ensure consistency across dataset variants, only question–answer pairs translated by all three models are retained. Additional data preprocessing includes removing some logically inconsistent question–answer pairs and correcting errors in the original English texts.

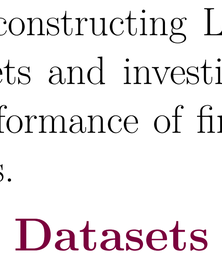
PathVQA
Pathology



Q: Where is this from?

A: Gastrointestinal system

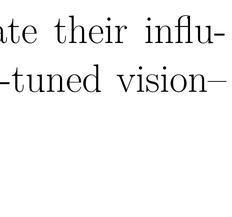
VQA-RAD
Radiology



Q: Are the kidneys visible in this image?

A: Yes

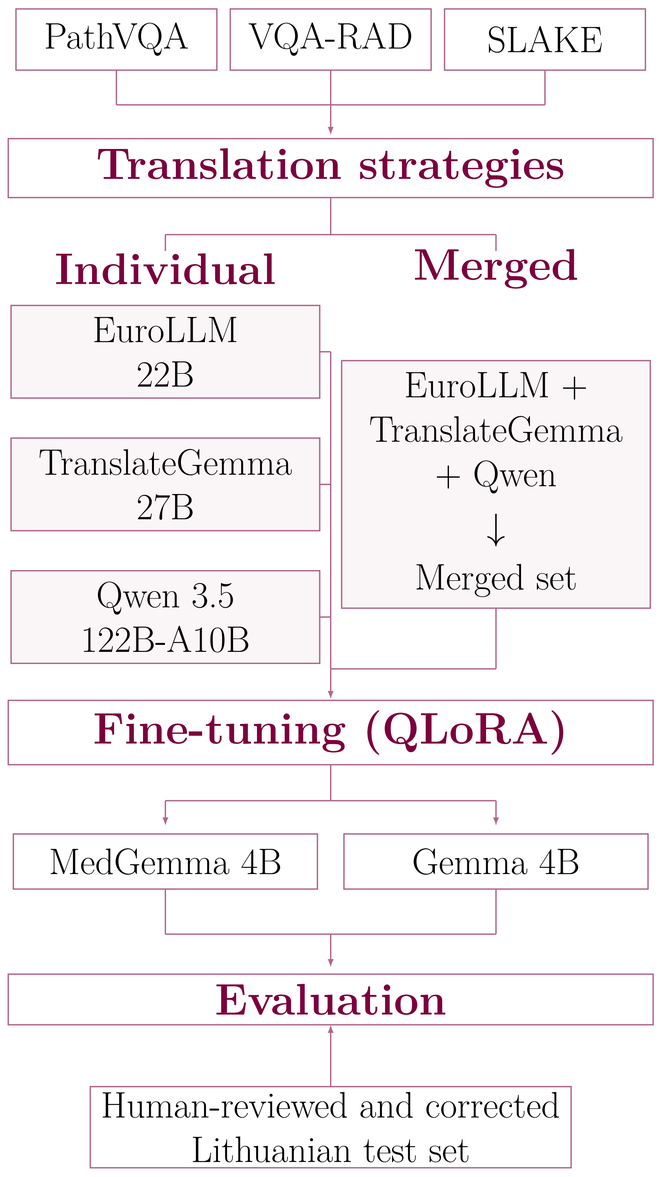
SLAKE
Radiology



Q: Is the lung healthy?

A: Yes

Experimental design



All four training variants contain the same aligned question–answer pairs, differing only in their Lithuanian translations. Each variant is used to fine-tune MedGemma 4B and Gemma 4B independently, resulting in eight model configurations. The human-reviewed and corrected Lithuanian test set is reserved exclusively for evaluation.

Evaluation

Translation quality. Questions and answers translated by each of the three LLMs are compared with human-reviewed and corrected Lithuanian reference translations using BLEU and chrF++ separately for questions and answers.

Downstream VQA performance. All eight fine-tuned configurations are evaluated on the same human-reviewed and corrected Lithuanian test set across the combined datasets. Closed-ended yes/no answers are assessed by accuracy, open-ended answers are assessed by exact match (EM), ROUGE-L F1 and BERTScore F1.

Translation quality

Comparison with human-reviewed and corrected reference translations on the test set

	BLEU		chrF++	
Translation model	**Q**	**A**	**Q**	**A**
Qwen 3.5 122B-A10B	30.85	38.52	58.23	**60.84**
EuroLLM 22B	**34.82**	**38.92**	**60.63**	60.27
TranslateGemma 27B	21.43	30.77	50.15	54.21

Q: questions, A: answers. Higher scores indicate greater agreement with the reference.

Fine-tuned VLMs

VQA performance of MedGemma 4B and Gemma 4B fine-tuned on different Lithuanian translation variants.

Training translation	Yes/No Acc. (%)	EM (%)	ROUGE-L F1 (%)	BERT Score F1
MedGemma 4B				
Qwen 3.5 122B-A10B	**83.52**	**16.87**	**21.92**	**0.7602**
EuroLLM 22B	81.89	14.18	19.53	0.7480
TranslateGemma 27B	81.56	9.54	14.00	0.7347
Merged	78.19	6.81	9.71	0.7319
Gemma 4B				
Qwen 3.5 122B-A10B	75.29	8.27	10.71	0.7177
EuroLLM 22B	76.05	11.96	16.10	0.7412
TranslateGemma 27B	55.22	8.05	10.90	0.7170
Merged	63.30	11.98	15.82	0.7458

Accuracy applies to yes/no answers; EM, ROUGE-L and BERTScore apply to open-ended answers. Accuracy, EM and ROUGE-L F1 are percentages, BERTScore F1 is on a 0–1 scale. Best scores are bolded.

Downstream VQA performance is evaluated on a pooled test set of 8018 question–answer pairs, comprising 81.3% PathVQA, 13.1% SLAKE, and 5.6% VQA-RAD.

Conclusions

- EuroLLM leads in question translation BLEU and chrF++, Qwen achieves the highest answer chrF++.
- MedGemma 4B fine-tuned on Qwen translations achieves the best scores on all four VQA metrics (83.52% yes/no accuracy, 0.7602 BERTScore F1).
- TranslateGemma achieves the lowest translation similarity scores and generally the weakest downstream VQA performance among individual translation variants.
- Merging translations does not consistently improve performance.

Acknowledgments. This project has received funding from the Research Council of Lithuania (LMTLT), agreement No S-ITP-26-11.