Adversarial Robustness of XGLM-564M Across High School and Undergraduate Tutoring Dialogues in English and Indonesian

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the difference in adversarial robustness scores for XGLM-564M when classifying tutoring dialogue acts across high school versus undergraduate level datasets in English and Indonesian. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Classifying German Language Proficiency Levels Using Large Language Models. Research question: What is the difference in adversarial robustness scores for XGLM-564M when classifying tutoring dialogue acts across high school versus undergraduate level datasets in English and Indonesian?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

13 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 4.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The model achieved a precision of 0.471 and recall of 0.640 for intermediate levels like B1.	×	0.02
The model misclassified 21 out of 25 A1 texts and all 25 A2 texts as B1 level.	×	0.02
The mean classification distance for the English Base Prompt was 1.12.	×	0.03
The LLaMA-3-8B-Instruct model was selected as the base model for fine-tuning.	✓	0.17
The dataset includes a balanced distribution of 1,567 learner texts across all six CEFR levels.	×	0.06
The English Base Prompt served as the initial method for instructing the model to classify texts according to CEFR level	×	0.09
The model tended to default to middle-range levels (B1–B2) when using the English Base Prompt.	×	0.04
The German Zero-Shot Prompt showed improved performance over the English Base Prompt.	×	0.05
The dataset was constructed by combining multiple existing corpora and synthetic generated data.	×	0.15
The model achieved a precision of 0.471 and recall of 0.640 for intermediate levels like B1.	×	0.02

References

- http://arxiv.org/abs/2304.07499v1
- http://arxiv.org/abs/2512.06483v1
- http://arxiv.org/abs/2402.17377v2