LLM of Babel:
Evaluation of LLMs on code for Non-English use cases

Author: Maksym Ziemlewski (M.Ziemlewski@student.tudelft.nl)
Supervisor: Jonathan Katzy
Responsible Professors: Maliheh Izadi, Arie van Deursen

TU Delft

1. Introduction

- Integration of Large Language Models (LLMs) into software development processes has significantly increased [1].
- Persistent challenge still stands out: the disparity in performance across multiple natural languages [2].
- The exploration of LLM evaluation and performance metrics outside of English-speaking contexts remains under-explored [3].

2. Objective

This research aims to highlight the limitations and errors encountered in generating comments for code in Polish using Meta's Code Llama.

3. Research Question

How does Code Llama perform in comment generation in Java when applied in Polish language settings?

1) How often does Code Llama generate erroneous comments for code written in Polish?
2) What are the most frequent types of limitations encountered in comment generation for code written in Polish?
3) What is the influence of temperature parameter for code comment completion in Java for the Polish language?

4. Methodology

Dataset preprocessing

↓

Quantitative Analysis with BLEU and ROUGE-L

↓

Qualitative analysis with Open Coding [4]

↓

Taxonomy of Errors

5 Results

- 25.2% of generated comments were manually evaluates as correct (RQ1).
- Taxonomy obtained through open coding [4] is displayed on Figure 4.
- The results of quantitative analysis are visualized in Figure 2.
- Figure 3 presents the frequency of each error in the manually evaluated dataset.
- Quality of comments decrease with higher temperature values (Figure 1). Results were obtained through qualitative evaluation.

$$softmax(z_i, T) = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}}$$

Equation 1: Softmax Activation Function with Temperature T

Temperature's influence is measured on correct predictions. We re-generate the comments with different temperature values and evaluate their correctness.

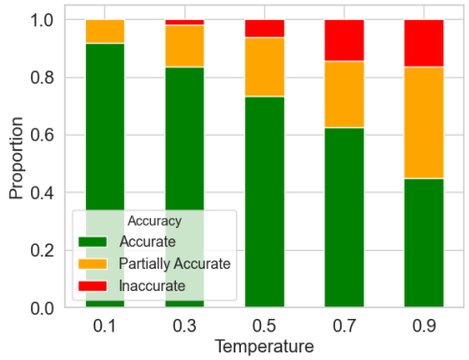

Figure 1: Accuracy Summary by Temperature (RQ3)

Majority of BLEU scores are close to 0, despite the expert evaluation indicating different observations. ROUGE-L is more evenly distributed.

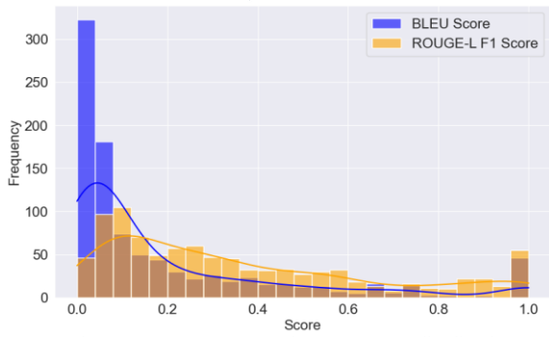

Figure 2: ROUGE-L F1 and BLEU scores distribution

The most common errors encountered were inclusion of code snippet, copying context and late termination. Code Llama underperformed with grammatical mistakes in comparison to other languages.

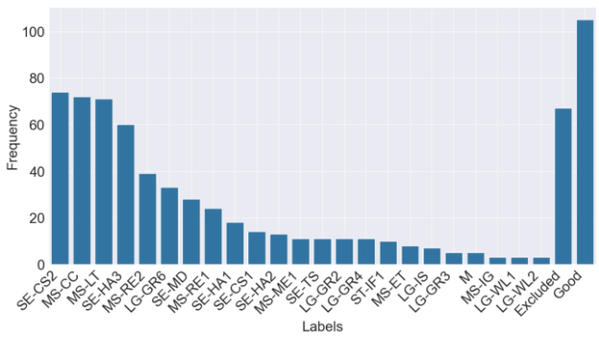

Figure 3: Distribution of Error Categories (RQ2)

6. Limitations

- **Language Scope:** This study focuses exclusively on Polish, with no English benchmark for comparison.
- **Bias in Labelling:** potential bias from individual experts, which could affect the accuracy of error classification and analysis.

7. Conclusion

- We identified common error categories in comment generation in Polish.
- Polish language-specific challenges, such as frequent grammatical errors or incoherent comments, underline the model's limitations in handling complex language syntax.
- BLEU underperforms as a metric for evaluating code comment quality in non-English languages.
- Model's inability to produce meaningful comments severely degrades perceived usability and trust in LLMs and general AI technology [5].
- We encourage the inclusion of diverse languages in the training corpuses of large language models.

8. References

[1] L. Belzner, T. Gabor, and M. Wirsing, 'Large language model assisted software engineering: prospects, challenges, and a case study', in International Conference on Bridging the Gap between AI and Reality, 2023, pp. 355–374.
[2] Y. Chang et al., 'A survey on evaluation of large language models', ACM Transactions on Intelligent Systems and Technology, 2023.
[3] F. Koto, N. Aisyah, H. Li, and T. Baldwin, 'Large Language Models Only Pass Primary School Exams in Indonesia: A Comprehensive Test on IndoMMLU', in The 2023 Conference on Empirical Methods in Natural Language Processing, 2023.
[4] S. H. Khandkar, 'Open coding', University of Calgary, vol. 23, no. 2009, p. 2009, 2009.
[5] H. Choung, P. David, and A. Ross, 'Trust in AI and its role in the acceptance of AI technologies', International Journal of Human--Computer Interaction, vol. 39, no. 9, pp. 1727–1739, 2023.

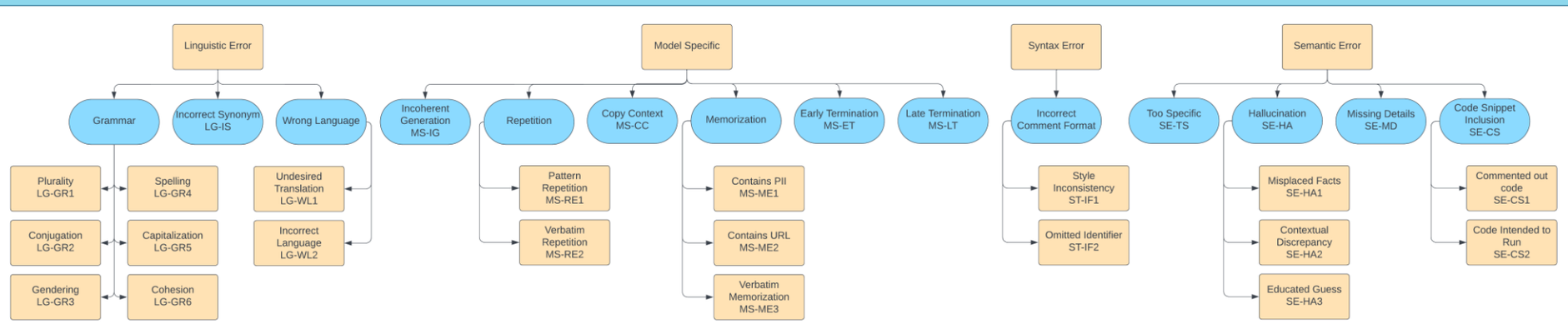

Figure 4: Error Taxonomy Tree