Comparative analysis on classic polish literature using leading small models and Bielik
Project Report for NLP Course, Winter 2024

inż. Łukasz Jaremek **inż. Tomasz Krupiński** **inż. Mieszko Mirgos** **inż. Patrycja Wysocka**
Jaremek.Lukasz@gmail.com mieszko.mirgos.stud@pw.edu.pl
01151416@pw.edu.pl 01151707@pw.edu.pl

supervisor: dr inż. Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

This paper evaluates the capabilities of smaller language models in analyzing classical Polish literature. We compare four models (Qwen2.5, LLaMA3.1, Bielik, and Mistral0.3) using a dataset of eleven seminal Polish works. Our evaluation framework includes question-answering. The findings demonstrate the effectiveness of smaller models in processing non-English literary texts and provide insights into preserving cultural heritage through accessible NLP technologies.

1 Introduction

The rapid evolution of Large Language Models (LLMs) has fundamentally transformed natural language processing, yet a significant gap remains in understanding their effectiveness for non-English languages, particularly in specialized domains like literature. While recent advances have demonstrated impressive capabilities in English-language tasks, the application of smaller, more accessible models to non-English literary traditions remains relatively unexplored. This research gap is particularly notable in the context of Polish literature, which possesses a rich cultural heritage and unique linguistic characteristics that pose distinct challenges for computational analysis.

1.1 Background and Significance

Recent developments in language models have primarily focused on scaling up model sizes and architectures, with models like GPT-4 and PaLM demonstrating unprecedented capabilities. However, these advances have predominantly centered on English-language applications, leaving significant questions about the effectiveness of more practical, smaller-scale models in processing culturally specific literary texts. The Polish language, with its complex morphology, free word order, and rich literary tradition, presents unique challenges and opportunities for such analysis.

The significance of this project lies in three key aspects:

1. It addresses the critical need for evaluating smaller, more accessible language models (7B parameters) in processing non-English literary texts, providing insights into their practical utility for cultural heritage analysis.

2. It contributes to the understanding of how language models handle the specific linguistic and stylistic features of Polish literature, particularly works from the Romantic period onward.

1.2 Scientific Goals and Research Questions

This project aims to evaluate and compare the capabilities of several 7B parameter language models in analyzing and understanding classical Polish literature. The primary research questions include:

1. To what extent can smaller language models effectively process and understand the linguistic and stylistic nuances of classical Polish literature?

2. How does the performance of Polish-specific models (like Bielik) compare with multilingual models in analyzing Polish literary texts?

3. What are the specific strengths and limitations of different model architectures when processing Polish literary texts?

Our research hypotheses posit that Polish-specific models will demonstrate superior performance in understanding context-specific literary references and stylistic nuances.

1.3 Report Structure

The remainder of this report is organized as follows:

- Section 2 presents a comprehensive literature review, covering recent developments in LLMs and their applications to non-English languages.
- Section 3 details our methodology, including the dataset composition, model specifications, and evaluation frameworks.
- Section 4 provides detailed exploratory data analysis with various plots to present chosen literature.
- Section 5 presents conducted experiments and its results.
- Section 6 discusses the results ans highlights the challenges of evaluation of Polish Large Language Models.

This research represents an effort in systematically evaluating the capabilities of smaller language models in processing classical Polish literature, with potential implications for both computational linguistics and literary studies. The findings will contribute to our understanding of how to effectively leverage modern NLP technologies for analyzing and preserving cultural heritage in non-English languages.

2 Literature Review

This section examines the key concepts and recent developments that form the foundation of our research on analyzing Polish literature using language models. We explore two main areas: the evolution of Large Language Models and recent advances in non-English language model applications.

2.1 Evolution of Large Language Models

Large Language Models (LLMs) have transformed natural language processing through significant architectural innovations and scaling achievements. The fundamental breakthrough came with the Transformer architecture [6], which introduced the self-attention mechanism as a more efficient alternative to traditional recurrent neural networks. This architecture has become the foundation for modern language models, enabling them to capture long-range dependencies and contextual relationships in text more effectively.

The development of models like BERT [2] marked a crucial advancement by introducing bidirectional training, allowing models to understand context from both directions. This innovation was particularly significant for tasks requiring deep contextual understanding, such as literary analysis. However, the real breakthrough in generative capabilities came with the GPT series, especially GPT-3 [1], which demonstrated that scaling up model parameters could significantly improve few-shot learning capabilities.

2.2 Language Models for Non-English Languages

The application of language models to non-English languages presents unique challenges and opportunities. These challenges stem from limited training data, diverse linguistic structures, and varying cultural contexts, while opportunities arise from creating models tailored to specific communities and needs. Recent research has focused on several key areas:

2.2.1 Cross-Lingual Transfer

Studies by [7] have shown that pre-trained LLMs can be effectively adapted to non-English languages through careful alignment strategies. Their work demonstrates the importance of:

- Semantic alignment through cross-lingual instruction tuning, which bridges the gap between languages with differing grammatical and lexical structures
- Combination of translation tasks with general language tasks to enhance model versatility across multilingual datasets
- Adaptation of pre-training strategies for specific language features, such as morphology, syntax, or word order, to better capture linguistic nuances

2.2.2 Language-Specific Models

The development of language-specific models has shown promising results in addressing linguistic nuances and achieving state-of-the-art performance for specific languages. Notable examples include:

- **Bielik** [4]: A Polish-specific model demonstrating strong performance on Polish language tasks, showcasing the potential of targeted pre-training and fine-tuning
- **HerBERT** [3]: A transformer-based model specifically optimized for Polish language understanding, achieving high accuracy in tasks such as sentiment analysis, named entity recognition, and syntactic parsing

These advancements underline the importance of both cross-lingual transfer methods and the creation of language-specific models for broadening the accessibility and applicability of language models to diverse linguistic contexts.

2.2.3 Evaluation Frameworks

The development of language-specific evaluation frameworks has been crucial for measuring model performance. For Polish, the KLEJ benchmark [5] has emerged as a standard evaluation tool, providing:

- Comprehensive assessment across multiple linguistic tasks
- Standardized evaluation metrics for Polish language processing
- Domain-specific evaluation capabilities

Information about the datasets is in the next section.

3 Methodology

This section outlines our approach to evaluating language model performance on Polish literary texts, detailing our dataset construction, model selection, and evaluation methods.

3.1 Dataset Construction

3.1.1 Source Material

Our dataset comprises eleven seminal works of Polish literature, sourced from public domain repositories on Wikisource. The corpus includes:

- **Romantic Poetry and Drama**:
  - ”Pan Tadeusz” (1834) by Adam Mickiewicz - National epic poem
  - ”Dziady” (1822) by Adam Mickiewicz - Dramatic cycle
  - ”Konrad Wallenrod” (1828) by Adam Mickiewicz - Narrative poem
  - ”Sonety” (1825) by Adam Mickiewicz - Poetic collection
  - ”Balladyna” (1839) by Juliusz Słowacki - Tragic drama
  - ”Kordian” (1834) by Juliusz Słowacki - Dramatic poem
- **Novels**:
  - ”Lalka” (1889) by Bolesław Prus - Realist novel
  - ”Quo Vadis” (1896) by Henryk Sienkiewicz - Historical novel
  - ”Trylogia” by Henryk Sienkiewicz:
    - ”Ogniem i mieczem” (1884)
    - ”Potop” (1886)
    - ”Pan Wołodyjowski” (1887)

3.1.2 Data Preprocessing and Task Preparation

Question-Answer Pairs was a main task for all models, questions were generated using GPT-4, with three questions crafted per paragraph. Duplicate questions were filtered out to ensure variety. The process included batch handling to manage errors effectively, and the final output consisted of Polish language question-and-answer pairs.

3.2 Models

To provide a comprehensive evaluation, we selected four multilingual language models that represent distinct approaches to natural language processing, differing in parameter sizes, target use cases, and design philosophies. These models include both general-purpose architectures and systems optimized for specific tasks.

Qwen2.5 (7B parameters) is a general-purpose multilingual model with a focus on broad domain coverage and a large context window, making it suitable for various NLP tasks. LLaMA3.1 (8B parameters) builds on advanced instruction-tuning techniques, ensuring enhanced performance on tasks requiring complex reasoning and multilingual understanding. Mistral v0.3 (7B parameters) is designed for efficient instruction-following, excelling in tasks with conversational or task-specific prompts. Finally, Bielik (7B parameters) is a Polish-specific model, fine-tuned for tasks in the Polish language, providing a targeted solution for our evaluation.

A detailed summary of these models, including their parameter counts and unique characteristics, is presented in Table 1.

Model Name	Parameters	Description
Qwen2.5	7B	Multilingual general-purpose model.
LLaMA3.1	8B	Advanced multilingual model.
Bielik	7B	Polish-specific model.
Mistral0.3	7B	Multilingual general-purpose model.

Table 1: Overview of evaluated language models.

3.3 Evaluation Framework

Our evaluation framework employs multiple metrics tailored to each task to comprehensively assess model performance.

3.3.1 Performance Metrics

Question and Answering (Q&A) tasks often rely on multiple evaluation metrics to assess the quality of generated responses, particularly in terms of linguistic precision and semantic accuracy. Among these, BLEU evaluates how closely the generated text matches the reference text by comparing n-grams, emphasizing precision. METEOR improves upon BLEU by considering synonyms, stemming, and word order, providing a more nuanced understanding of linguistic alignment. ROUGE focuses on recall by comparing overlapping n-grams and sequences, making it especially effective for tasks where capturing the full meaning is critical. Because of the nuanced nature of Q&A tasks, the BERT-score metric is also employed, which uses contextual embeddings from pre-trained language models to evaluate semantic similarity between generated and reference texts. Together, these metrics offer a comprehensive evaluation framework for generated responses.

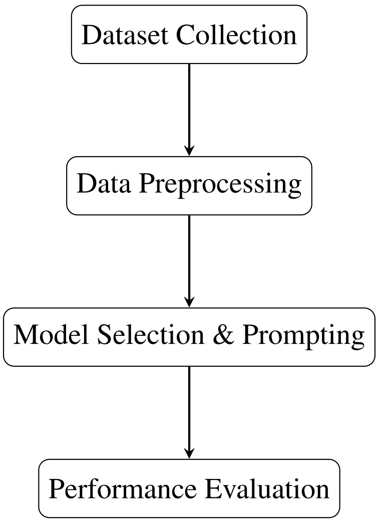
3.4 Evaluation Process

The evaluation process for each model consists of asking it 100 questions about every book. Models are directly evaluated on all tasks to establish a baseline.

Results are aggregated and compared across models and evaluation metrics to provide a comprehensive assessment of each model's capabilities in processing Polish literary texts.

4 Experiment Procedure

The experiment was structured to analyze classic Polish literature using small NLP models. The workflow comprised the following stages:



Each of the block, is described in details in its own corresponding section.

5 Exploratory Data Analysis results

5.1 Statistics aggregated between books

Figure 1 compares mean sentence lengths across works, with ”Pan Tadeusz” being the longest and ”Balladyna” the shortest. Figure 2 compares mean word lengths, showing relatively consistent values across all works. Figure 3 compares sentence counts across works, with ”Potop” having the most sentences and ”Konrad Wallenrod” the fewest. Figure 4 compares word counts across works, with ”Potop” having the highest count and ”Konrad Wallenrod” the lowest.

5.2 Example statistics drawn from a single book

The figure 5 shows the distribution of word lengths in ”Kordian,” highlighting a peak around shorter word lengths. We can see that the data follows normal distribution. Figure 6 shows the most frequent words in ”Kordian,” with ”kordian,” ”car,” and ”lud” being the top three. On figure 7 bigrams for ”Kordian” can be seen, that the most frequent one is ”wielki ksiaże” referring to main character of the book. On figure 8 distribution of sentence

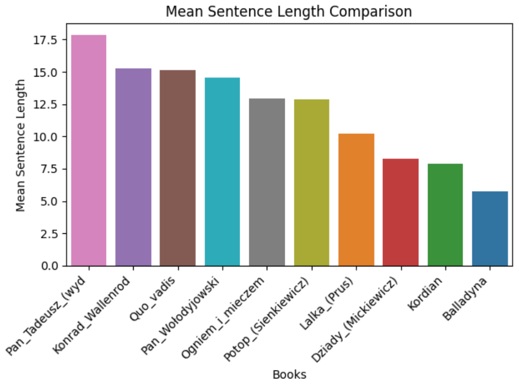


Figure 1: Mean sentence length across works

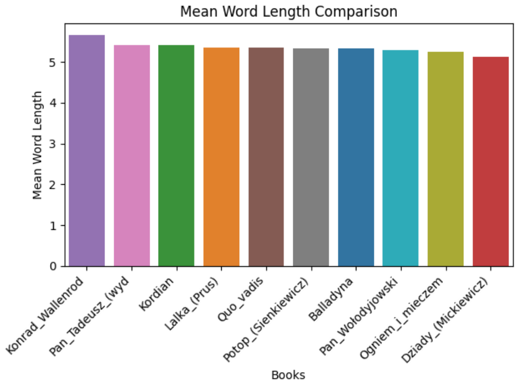


Figure 2: Mean word length

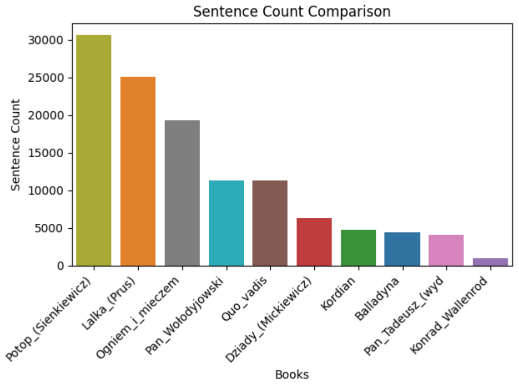


Figure 3: Sentence count

length is presented, showing that rarely very log sentences appear, achieving even 60 words.

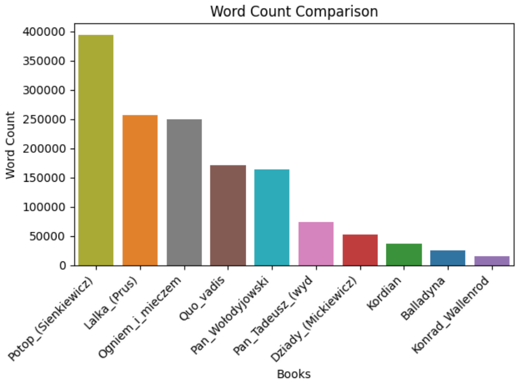


Figure 4: Word count

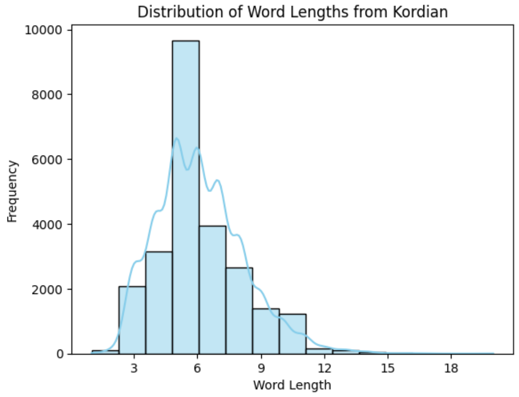


Figure 5: Word count Kordian

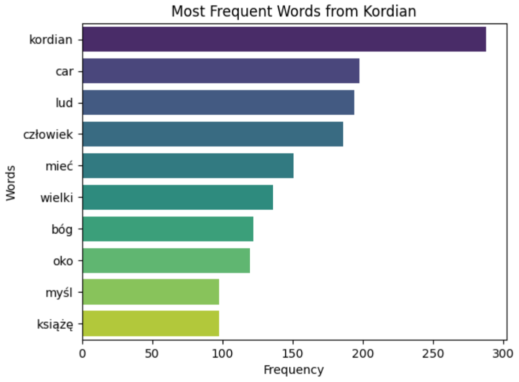


Figure 6: Common words Kordian

6 Experiments Results

To evaluate the Q&A task, we have decided to use the metrics defined in section 3.3.1. Those metrics are:

- BLEU,
- METEOR,
- ROUGE1,

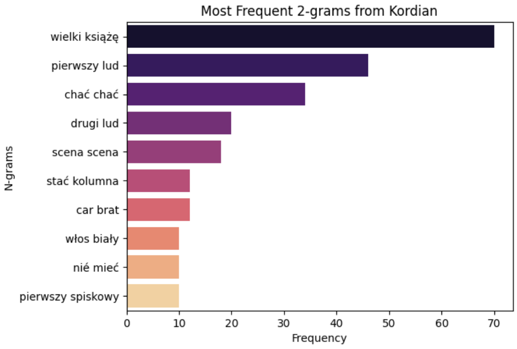


Figure 7: Frequent 2 grams Kordian

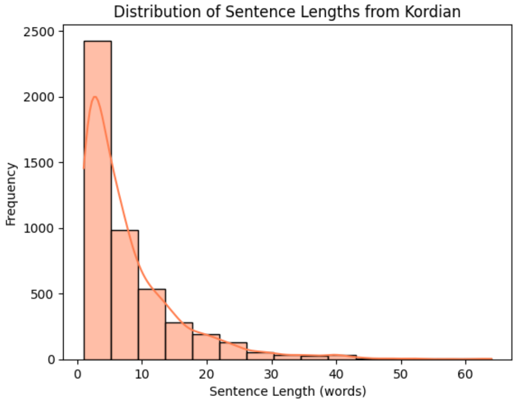


Figure 8: Mean sentence length Kordian

- BERT-score

For each dataset, of which there are 11, we prepared 100 questions with corresponding true answer. Also, each corresponding true answer has been paraphrased, to create more references. In total there are six references per one question. Them, we got the answers from the model by prompting it the same questions. With data prepared that way, we used already implemented metrics in various python libraries to evaluate the model.

The tables in appendix (table 2, table 3, table 4) show the results of our evaluation. It is demonstrated that BERT-score consistently performs as the best metric for evaluating the models, as it shows higher and more stable values compared to BLEU, METEOR, and ROUGE1. While BLEU, METEOR, and ROUGE1 metrics exhibit extremely low or even zero scores across many datasets, particularly for Llama, BERT-score provides a more nuanced and consistent assessment of the models' capabilities in understanding Polish text.

The Figure 11 visualizes the BERTScore comparison of four models—Qwen2.5, Llama3.1, Bielik, and Mistralv0.3—across various classic Polish literature books. Qwen2.5 consistently achieves the highest scores, followed by Bielik and Mistralv0.3, while Llama3.1 demonstrates slightly lower performance overall. The comparison highlights the effectiveness of models tailored for multilingual tasks (like Qwen2.5) or Polish-specific tasks (like Bielik) in understanding and evaluating Polish literary texts.

The Figure 9 below, shows the averaged result over all the datasets per model. As we can see, clearly the best model is Qwen2.5, followed by Mistralv0.3, bielik and at last by Llama3.1.

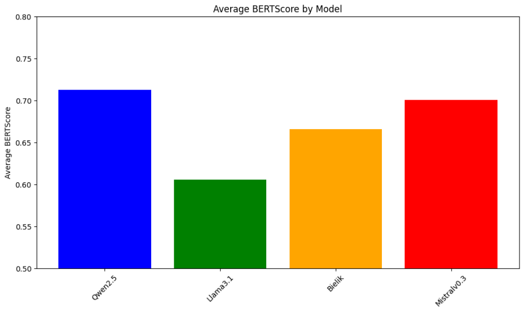


Figure 9: Average model result

The Figure 10 shows the results per dataset. There no real visible difference between them, except only for Ogniem i Mieczem dataset, suggesting that the datasets are balanced in difficulty.

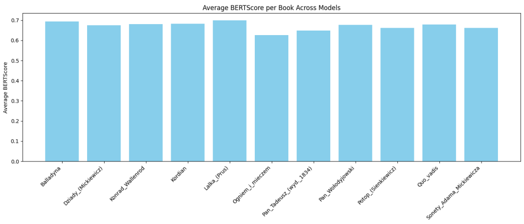


Figure 10: Average dataset result

7 Discussion

The results highlight the performance of four models - Qwen2.5, Llama3.1, Bielik, and Mistralv0.3 - on Polish literary datasets using BLEU, METEOR, ROUGE1, and BERTScore.

Key Observations:

- Qwen2.5 consistently achieves the highest BERT-scores, demonstrating strong multilingual capabilities.

- Mistralv0.3 performs competitively, but cannot outperform Qwen2.5
- Bielik achieves solid results but falls slightly below Mistralv0.3.
- Llama3.1 struggles significantly.
- BERTScore is the most reliable metric, while BLEU, METEOR, and ROUGE1 often return inadequate or zero scores.

Surpringly Qwen2.5 is the most effective model for Polish texts, followed closely by Mistralv0.3, while Bielik turned out to be on the 3rd place. This result was unexpected, as Bielik is a Polish-specific model, and we anticipated it to outperform general-purpose architectures like Qwen2.5 and Mistralv0.3.

Metrics like BLEU, METEOR and ROUGE1 are insufficient for Polish text evaluation. That is because of the multiple different words in polish language and also the lexical variety of the words. This highlights the inadequacy of traditional metrics for Polish text evaluation, emphasizing the need for advanced metrics like BERTScore. It stands out as the most appropriate evaluation metric for these tasks.

Our research has shown, that there is indeed a need for improvements in terms of polish based LLMs and also evaluation metrics tailored specifically to the Polish language, such as advanced contextual methods. These two aspects are essential for more accurate assessments. For future directions, further fine-tuning of Polish-specific models like Bielik with additional high-quality data could significantly enhance their performance and potentially surpass general-purpose models. Finally, expanding datasets to include more diverse linguistic structures and historical vocabulary will address challenges associated with processing complex Polish texts.

8 Answers to reviewers

In this section, we provide a detailed response to the feedback and comments provided by the reviewers. This section aims to demonstrate how their feedback guided us in refining the project and enhancing its overall impact.

8.1 Data visualisation

The reviewers drew attention to the colouring of some of the graphs, which was not informative enough. We have corrected the charts indicated and taken the valuable comments into account when creating new ones so that a similar mistake will not occur again.

8.2 Ethical concerns

The reviewers drew attention too to the ethical issues of the work, specifically the possibility of generating erroneous summaries or false information about school readings. The problem is, unfortunately, impossible for us to circumvent. Even the best language models make mistakes, generate unsubstantiated hallucinations and mislead their users. In addition, the small language models we worked on in this study did not cope perfectly with such a specific text as old books.

8.3 Code

Reviewers praised the high quality and readability of the code, which was a great compliment for us. However, they noted the lack of documentation on some features. In response to this comment, we added the missing descriptions in the code.

8.4 Limitations of work

The subject of this work is unique because of the times in which the books analysed were written. They are not among the most recent and are often written in archaic and outdated language. The reviewers pointed out that the models might have a problem working on such a text and, as our results showed, they were right. Perhaps if we had trained the models with a large group of old books or had advanced graphics cards in their possession, we would have been able to overcome this problem.

References

[1] Tom B Brown et al. “Language Models are Few-Shot Learners”. In: *arXiv preprint arXiv:2005.14165* (2020). URL: `https://arxiv.org/abs/2005.14165`.

[2] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *arXiv preprint arXiv:1810.04805* (2018). URL: `https://arxiv.org/abs/1810.04805`.

[3] Robert Mroczkowski et al. “HerBERT: Efficiently Pretrained Transformer-based Language Model for Polish”. In: *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. Ed. by Bogdan Babych et al. Kiyv, Ukraine: Association for Computational Linguistics, Apr. 2021, pp. 1–10. URL: https://aclanthology.org/2021.bsnlp-1.1.

[4] Krzysztof Ociepa et al. *Bielik 7B v0.1: A Polish Language Model – Development, Insights, and Evaluation*. 2024. arXiv: 2410.18565 [cs.CL]. URL: https://arxiv.org/abs/2410.18565.

[5] Piotr Rybak et al. *KLEJ: Comprehensive Benchmark for Polish Language Understanding*. 2020. arXiv: 2005.00630 [cs.CL]. URL: https://arxiv.org/abs/2005.00630.

[6] Ashish Vaswani et al. “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems* 30 (2017). URL: https://arxiv.org/abs/1706.03762.

[7] Wenhao Zhu et al. *Extrapolating Large Language Models to Non-English by Aligning Languages*. 2023. arXiv: 2308.04948 [cs.CL]. URL: https://arxiv.org/abs/2308.04948.

9 Appendix

Dataset	BLEU	METEOR	ROUGE1	BERT-score
Balladyna	0.252	0.000	0.310	0.751
Dziady_(Mickiewicz)	0.155	0.000	0.221	0.705
Konrad_Wallenrod	0.158	0.000	0.234	0.724
Kordian	0.180	0.000	0.251	0.732
Lalka_(Prus)	0.305	0.000	0.368	0.772
Ogniem_i_mieczem	0.038	0.000	0.097	0.660
Pan_Tadeusz_(wyd._1834)	0.093	0.000	0.138	0.676
Pan_Wołodyjowski	0.149	0.000	0.235	0.726
Potop_(Sienkiewicz)	0.124	0.000	0.191	0.699
Quo_vadis	0.135	0.000	0.215	0.721
Sonety_Adama_Mickiewicza	0.085	0.000	0.135	0.674

Table 2: QWEN results

Dataset	BLEU	METEOR	ROUGE1	BERT-score
Balladyna	0.068	0.000	0.111	0.613
Dziady_(Mickiewicz)	0.074	0.000	0.109	0.625
Konrad_Wallenrod	0.067	0.000	0.107	0.620
Kordian	0.065	0.000	0.105	0.625
Lalka_(Prus)	0.071	0.000	0.113	0.611
Ogniem_i_mieczem	0.015	0.000	0.035	0.567
Pan_Tadeusz_(wyd._1834)	0.049	0.000	0.073	0.598
Pan_Wołodyjowski	0.042	0.000	0.080	0.598
Potop_(Sienkiewicz)	0.044	0.000	0.078	0.596
Quo_vadis	0.044	0.000	0.080	0.599
Sonety_Adama_Mickiewicza	0.068	0.000	0.093	0.610

Table 3: Llama results

Dataset	BLEU	METEOR	ROUGE1	BERT-score
Balladyna	0.107	0.000	0.155	0.678
Dziady_(Mickiewicz)	0.085	0.000	0.134	0.667
Konrad_Wallenrod	0.084	0.000	0.130	0.668
Kordian	0.082	0.000	0.124	0.666
Lalka_(Prus)	0.111	0.000	0.145	0.676
Ogniem_i_mieczem	0.033	0.000	0.064	0.636
Pan_Tadeusz_(wyd._1834)	0.059	0.000	0.086	0.644
Pan_Wołodyjowski	0.071	0.000	0.112	0.672
Potop_(Sienkiewicz)	0.075	0.000	0.122	0.666
Quo_vadis	0.110	0.000	0.170	0.693
Sonety_Adama_Mickiewicza	0.073	0.000	0.120	0.662

Table 4: Bielik results

Dataset	BLEU	METEOR	ROUGE1	BERT-score
Balladyna	0.246	0.000	0.310	0.732
Dziady_(Mickiewicz)	0.142	0.000	0.196	0.703
Konrad_Wallenrod	0.145	0.000	0.218	0.710
Kordian	0.136	0.000	0.207	0.706
Lalka_(Prus)	0.248	0.000	0.315	0.739
Ogniem_i_mieczem	0.026	0.000	0.070	0.641
Pan_Tadeusz_(wyd._1834)	0.097	0.000	0.155	0.679
Pan_Wołodyjowski	0.139	0.000	0.220	0.711
Potop_(Sienkiewicz)	0.117	0.000	0.183	0.689
Quo_vadis	0.125	0.000	0.200	0.704
Sonety_Adama_Mickiewicza	0.110	0.000	0.188	0.698

Table 5: Mistral results

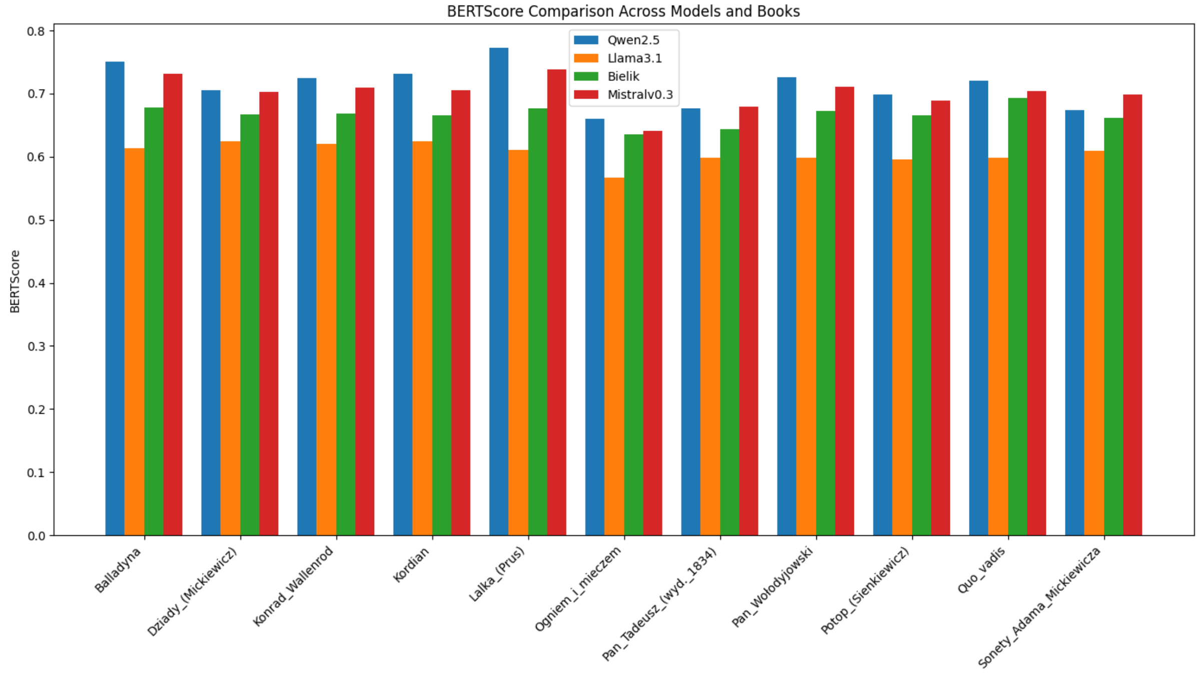


Figure 11: Comparison across 4 models

Reproducibility checklist

Overall results:

- MODEL DESCRIPTION – we used Bielik 7B (finetune of Mistral 7B), Mistralv0.3 7B, Qwen 2.5 and Llama 3.1. All models are LLMs and their architectures are publicly availabile

- LINK TO CODE – https://github.com/grant-TraDA/NLP-2024W/tree/main/projects/PoCs/Mirgos-Krupinski-Jaremek-Wysocka

- INFRASTRUCTURE – Huggingface hosted models and local personal computers.

- RUNTIME PARAMETERS – Hardware dependent, around 3 hours for a low grade hosted endpoint to execute tests.

- PARAMETERS – 7B

- VALIDATION PERFORMANCE – Mistralv0.3 and Qwen 2.5 dominated the other models, Bielik outeprforemd Llama3.1.
- METRICS – Four different metrics: BLEU, METEOR, ROUGE and BERT score, all calculated in a single .ipynb file in folder metrics_scripts.

Multiple Experiments:

- NO TRAINING EVAL RUNS – N/A
- HYPER BOUND – N/A
- HYPER BEST CONFIG – N/A
- HYPER SEARCH – N/A
- HYPER METHOD – N/A
- EXPECTED PERF – N/A

Datasets – utilized in the experiments and/or the created ones:

- DATA STATS – 11 datasets, each with 100 QA pairs
- DATA SPLIT – N/A
- DATA PROCESSING – N/A
- DATA DOWNLOAD – https://github.com/grant-TraDA/NLP-2024W/tree/main/projects/PoCs/Mirgos-Krupinski-Jaremek-Wysocka/data_processing_scripts
- NEW DATA DESCRIPTION – Manual check of the data
- DATA LANGUAGES – Polish