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 HerBERT) using a dataset of eleven seminal Polish works. Our evaluation framework includes blank filling, question-answering, summary generation, and translation assessment, with and without Retrieval-Augmented Generation (RAG). 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.

3. It explores the potential of Retrieval-Augmented Generation (RAG) in enhancing model performance for domain-specific literary analysis, offering insights into cost-effective approaches for processing culturally significant texts.

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. Can RAG techniques significantly improve the performance of these models in literary analysis tasks?

4. 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.

- RAG implementation will significantly improve model performance, particularly in handling historical and cultural references.

- Different model architectures will show varying strengths in different aspects of literary analysis (e.g., stylistic analysis vs. content understanding).

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, RAG systems, and their applications to non-English languages.

- Section 3 details our methodology, including the dataset composition, model specifications, and evaluation frameworks.

- Section 4 provides a summary of our approach and expected outcomes, including preliminary evaluation metrics.

- Section 5 lists the references and related works that inform our research methodology.

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.

- Background and significance of the project (state of the art, justification for tackling a specific scientific problem, justification for the pioneering nature of the project, the impact of the project results on the development of the research field and scientific discipline);

- Scientific goal of the project (description of the problem to be solved, research questions and hypotheses) and the project contributions;

- Description of the rest of the report (what contains the following sections).

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 three main areas: the evolution of Large Language Models, the emergence of RAG systems, 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 [8], 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 Retrieval-Augmented Generation (RAG)

RAG represents a significant advancement in improving language model accuracy and reliability. As detailed in recent surveys [4], RAG systems

combine the generative capabilities of language models with explicit knowledge retrieval, addressing key limitations of traditional LLMs:

- **Knowledge Integration**: RAG enables models to access external knowledge sources, reducing hallucinations and improving factual accuracy.
- **Context Enhancement**: The retrieval component allows models to incorporate relevant background information, particularly valuable for domain-specific tasks like literary analysis.
- **Scalability**: RAG provides a more efficient alternative to continuous model retraining, particularly important for specialized domains like Polish literature.

Recent work by [3] has demonstrated RAG's effectiveness in knowledge-intensive tasks, showing particular promise for applications in cultural and literary analysis.

2.3 Language Models for Non-English Languages

The application of language models to non-English languages presents unique challenges and opportunities. Recent research has focused on several key areas:

2.3.1 Cross-Lingual Transfer

Studies by [9] 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
- Combination of translation tasks with general language tasks
- Adaptation of pre-training strategies for specific language features

2.3.2 Language-Specific Models

The development of language-specific models has shown promising results. Notable examples include:

- **Bielik** [6]: A Polish-specific model demonstrating strong performance on Polish language tasks
- **HerBERT** [5]: A transformer-based model specifically optimized for Polish language understanding

2.3.3 Evaluation Frameworks

The development of language-specific evaluation frameworks has been crucial for measuring model performance. For Polish, the KLEJ benchmark [7] 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

This research aims to address these gaps by providing a comprehensive evaluation of how different model architectures perform in analyzing classical Polish literature, with a specific focus on the capabilities of smaller, more accessible models enhanced with RAG techniques.

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

The texts underwent systematic preprocessing to create four distinct evaluation tasks.

For the **Blank Filling Task**, words were automatically masked at a 30% ratio, with the selection process being random. Each blank was replaced with a length-preserving underscore, while the original words were stored as solutions to maintain traceability.

In the **Question-Answer Pairs** task, 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.

The **Summary Generation** task involved processing text in 2000-character chunks. A random sampling of up to 100 chunks was selected for summarization using GPT-4, with a focus on Polish language summaries. The process was conducted in batches, with chunking ensuring manageable input sizes for the model.

For the **Translation Assessment** task, entire paragraphs were processed and randomly sampled, with a limit of 100 paragraphs. GPT-4 was employed to perform Polish-to-English translations, and the original structure of the paragraphs was preserved throughout the task.

3.2 Models

To provide a comprehensive evaluation, we selected four language models that represent distinct approaches to natural language processing and differ in their target use cases, parameter sizes, and design philosophies. These models include multilingual general-purpose architectures, advanced multilingual systems, and models specifically fine-tuned for Polish. A detailed summary of these models, including their parameter counts and unique characteristics, is presented in Table 1.

3.3 Evaluation Framework

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

3.3.1 Performance Metrics

For the **Blank Filling** task, the Exact Match (EM) score is used to measure how accurately the model predicts masked words. In the **Question and Answering (Q&A)** task, BLEU, METEOR, and ROUGE scores evaluate the generated responses in terms of linguistic precision, recall, and fluency. The **Summary Generation** task is assessed using ROUGE variants, including ROUGE-1, ROUGE-2, and ROUGE-L, which focus on capturing different aspects of overlap between the generated and reference summaries. For the **Translation** task, BLEU and METEOR scores are employed to measure the quality and accuracy of Polish-to-English translations.

3.3.2 KLEJ Benchmark Integration

To further evaluate Polish language understanding, the models are tested on the KLEJ benchmark. This includes tasks such as sentence similarity assessment to evaluate semantic closeness, textual entailment to assess logical relationships between sentence pairs, named entity recognition to identify proper nouns and categories, and sentiment analysis to determine the emotional tone of text.

3.4 Evaluation Process

The evaluation process for each model consists of two distinct phases. In the **Base Performance** phase, the models are directly evaluated on all tasks to establish a baseline. In the **RAG-Enhanced Performance** phase, the models are augmented with Retrieval-Augmented Generation (RAG) techniques, enabling them to access external knowledge sources to enhance task-specific performance.

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

4 Exploratory Data Analysis results

4.1 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

Model Name	Parameters	Description
Qwen2.5	7B	Multilingual general-purpose model.
LLaMA3.1	8B	Advanced multilingual model.
Bielik	7B	Polish-specific model.
HerBERT	0.34B	Polish BERT variant.

Table 1: Overview of evaluated language models.

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.

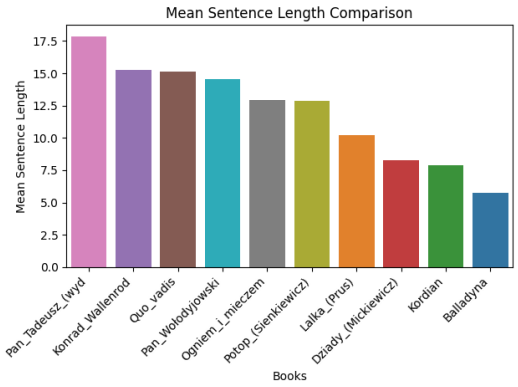

Figure 1: Mean sentence length across works

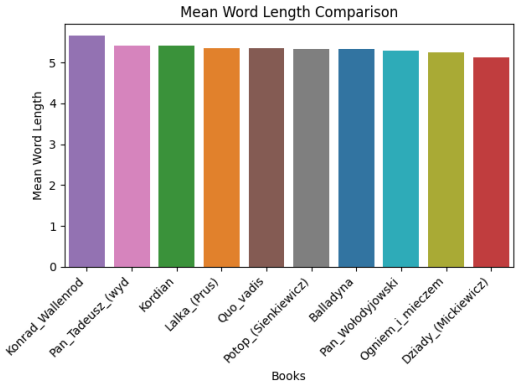

Figure 2: Mean word length

4.2 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

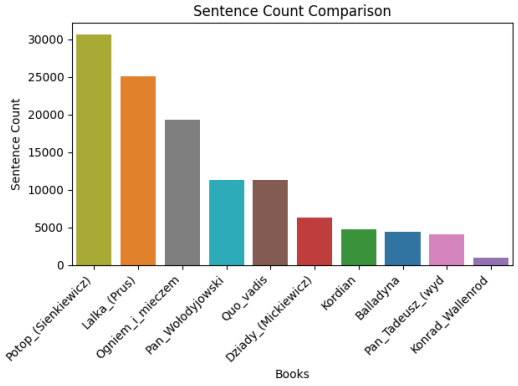

Figure 3: Sentence count

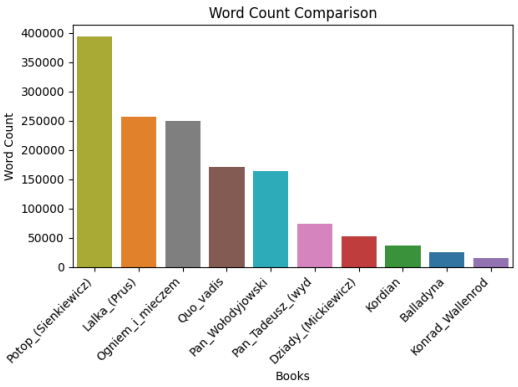

Figure 4: Word count

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

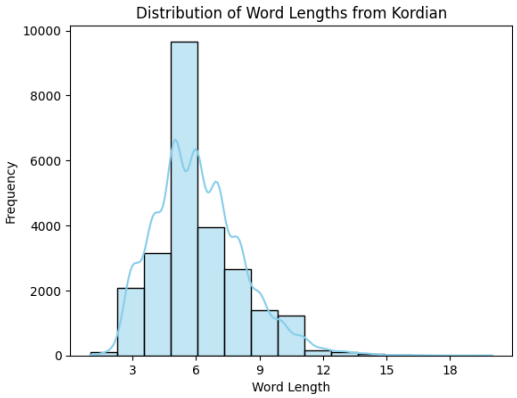

Figure 5: Word count Kordian

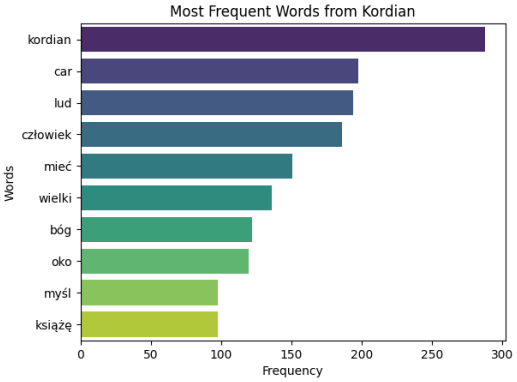

Figure 6: Common words Kordian

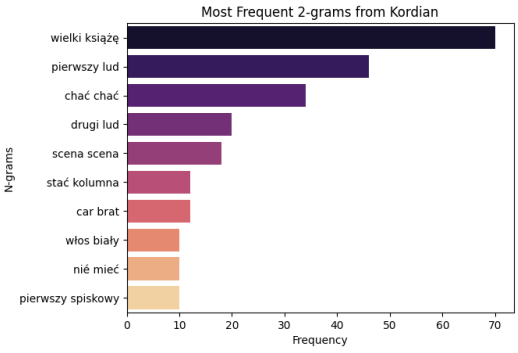

Figure 7: Frequent 2 grams Kordian

4.3 Preliminary evaluation metrics

For the POC we only tested one model, and only done the Q&A task. To evaluate our current solutions, we have decided to use the metrics defined in section 3.3.1. Those metrics are:

- BLEU
- METEOR
- ROUGE (1,2,L)

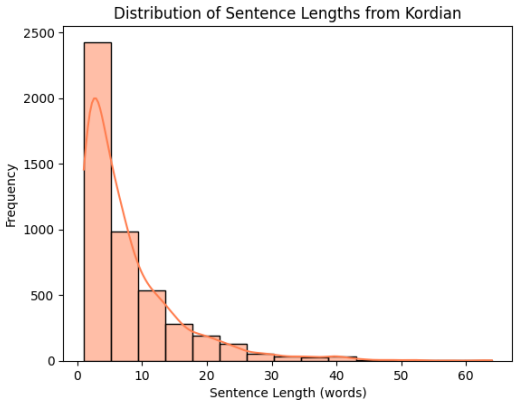

Figure 8: Mean sentence length Kordian

For each dataset, of which there are 10, we prepared 100 questions with corresponding true answer. 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 table below shows the results of our evaluation. Bolded results show the biggest values:

Dataset	BLEU	METEOR	ROUGE1
Balladyna	**0.054**	**0.250**	**0.335**
Dziady	0.013	0.174	0.248
Konrad Wallenrod	0.019	0.209	0.294
Kordian	0.014	0.167	0.238
Lalka	**0.056**	**0.221**	**0.318**
Ogniem i mieczem	0.000	0.033	0.049
Pan Tadeusz	0.006	0.111	0.157
Pan Wołodyjowski	0.009	0.155	0.219
Potop	0.006	0.111	0.1851
Quo vadis	0.014	0.159	0.224

Dataset	ROUGE2	ROUGEL
Balladyna	**0.168**	**0.305**
Dziady	0.101	0.228
Konrad Wallenrod	0.137	0.284
Kordian	0.099	0.224
Lalka	**0.181**	**0.298**
Ogniem i mieczem	0.006	0.045
Pan Tadeusz	0.051	0.145
Pan Wołodyjowski	0.091	0.209
Potop	0.075	0.171
Quo vadis	0.089	0.211

As we can see, the results are not satisfactory. We would hope for the metrics to be much closer to value of 1 rather than 0. However, we have learned from that experiment, that we should make

much more reference sentences - as currently we have got only one per question. Therefore the metrics more often than not evaluate answers as incorrect, even if the meaning of the sentences is actually correct.

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] Yunfan Gao et al. "Retrieval-Augmented Generation for Large Language Models: A Survey". In: *arXiv preprint arXiv:2312.10997* (2023). URL: https://arxiv.org/abs/2312.10997.

[4] Shailja Gupta et al. "A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions". In: *arXiv preprint arXiv:2410.12837* (2024). URL: https://arxiv.org/abs/2410.12837.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.