KRZYSZTOF OCIEPA

ŁUKASZ FLIS

REMIGIUSZ KINAS

KRZYSZTOF WRÓBEL

ADRIAN GWOŹDZIEJ

BIELIK 11B V3: MULTILINGUAL LARGE LANGUAGE MODEL FOR EUROPEAN LANGUAGES

Abstract

We present Bielik 11B v3, a state-of-the-art language model highly optimized for the Polish language, while also maintaining strong capabilities in other European languages. This model extends the Mistral 7B v0.2 architecture, scaled to 11B parameters via depth up-scaling. Its development involved a comprehensive four-stage training pipeline: continuous pre-training, supervised fine-tuning (SFT), Direct Preference Optimization (DPO), and reinforcement learning.

Comprehensive evaluations demonstrate that Bielik 11B v3 achieves exceptional performance. It significantly surpasses other specialized Polish language models and outperforms many larger models (with 2-6 times more parameters) on a wide range of tasks, from basic linguistic understanding to complex reasoning.

The model's parameter efficiency, combined with extensive quantization options, allows for effective deployment across diverse hardware configurations. Bielik 11B v3 not only advances AI capabilities for the Polish language but also establishes a new benchmark for developing resource-efficient, high-performance models for less-represented languages.

Note

This is a preliminary version; results and exposition may change.

1. Introduction

The development of large language models for underrepresented languages has gained significant momentum, with several initiatives targeting European languages. Our work builds upon our previous Bielik 11B v2 model [26] and complements our smaller Bielik v3 models [25]. This effort aligns with broader European initiatives, including EuroLLM [22], which targets multilingual capabilities across EU languages, Apertus [12], focused on democratizing open LLMs for global language environments, and Polish-specific projects such as the PLLuM family [17].

2. Model and Tokenizer

This section provides an overview of the model architecture and tokenizer setup, explaining the main design principles and implementation configurations.

2.1. Model Architecture

Parameter	Value
Layers	50
Model Dimension	4096
Attention Heads	32
Key/Value Heads	8
Head Size	128
Intermediate Size	14336
Activation Function	SwiGLU
Vocabulary Size	32128
Positional Embeddings	RoPE ($\theta = 1000000$)
Native Context Length	32768
Extended Context Length (YaRN)	131072

Table 1

Model architecture.

The Bielik 11B v3 architecture is founded on the Transformer model [45], with its principal configuration parameters summarized in Table 1. The design integrates a suite of contemporary architectural and algorithmic improvements intended to enhance computational efficiency, stability, and overall model performance.

Self-attention with causal masking [45] enables the network to dynamically allocate attention across different elements of the input sequence. The application of a causal mask constrains

attention to preceding tokens, thereby preserving the autoregressive property critical for effective language modeling.

Grouped-query attention (GQA) [1] provides a more resource-efficient variant of multi-head attention by employing fewer key„value heads than query heads. This configuration substantially reduces both computational and memory overheads while maintaining comparable representational fidelity, thus facilitating more efficient processing of extended input sequences.

SwiGLU activation [9, 38] combines the Swish nonlinearity with Gated Linear Units (GLU), offering improved gradient flow and optimization stability. Empirical evidence indicates that SwiGLU yields superior performance and trainability relative to conventional activation functions such as ReLU.

Rotary Positional Embeddings (RoPE) [43] are employed to encode relative positional information within token representations. In contrast to absolute positional embeddings, RoPE provides enhanced generalization to longer sequences and improved modeling of positional dependencies, particularly in tasks sensitive to word order and spatial context.

Root Mean Square Layer Normalization (RMSNorm) [14] normalizes activations based on their root mean square, rather than their mean and variance as in standard Layer Normalization. This approach yields marginally faster computation and greater numerical stability during training, contributing to overall model robustness.

Pre-normalization applies layer normalization prior to the self-attention and feed-forward sublayers. This architectural choice enhances gradient flow throughout the network, resulting in improved convergence behavior and more consistent model performance.

Building upon the substantial success of the previous Bielik 11B v2 [26] iteration, we adopted an identical, validated architectural approach. This strategic decision to leverage an existing, high-performing model structure, rather than commencing development from scratch, was fundamentally motivated by the imperative for efficient resource allocation. By concentrating our efforts on the linguistic and structural refinement of a proven base model, we achieved significant optimization in both time and computational resources.

The resulting Bielik 11B v3 model is an adaptation of the Mistral 7B v0.2 model [13], meticulously scaled via the Depth Up-Scaling methodology [15], followed by a subsequent phase of continued pretraining. Specifically, we initiated the process with the original $n = 32$-layer architecture. The up-scaling procedure involved duplicating the model's existing layers, followed by the strategic excision of the final eight and initial eight layers at the central junction ($m = 7$), culminating in a $s = 50$-layer model, as schematically detailed in Figure 1. This target of 50 layers was chosen to achieve a final model size of approximately 11 billion parameters, which retains the crucial operational characteristic of being runnable on mainstream, consumer-grade GPUs equipped with up to 24 GB of VRAM.

2.2. Tokenizer

The tokenization efficiency is primarily assessed by the resulting token count for a given input sequence, where a reduction in this number typically correlates with faster and more resource-efficient text generation by the language model. The original tokenizer from the
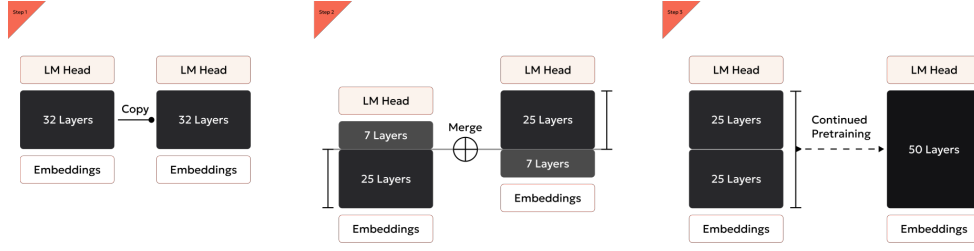
Figure 1. Depth up-scaling with $n = 32$, $m = 7$, and $s = 50$.

Tokenizer	Vocab size	Avg tokens	Polish			English		
			Tokens	CpT	TpW	Tokens	CpT	TpW
EuroLLM 9B	128000	420	437	4.11	1.88	404	4.79	1.27
Mistral Small 3.2 24B	131072	462	547	3.28	2.36	377	5.14	1.19
APT3	31980	480	344	5.22	1.48	615	3.15	1.93
Qwen3	151669	499	625	2.87	2.69	373	5.19	1.17
APT4	32000	503	375	4.78	1.62	631	3.07	1.98
Llama2	32000	554	681	2.63	2.94	427	4.53	1.34
Mistral 7B v0.2	32000	578	747	2.40	3.22	408	4.75	1.28

Table 2: Comparison of token count, characters per token (CpT), and tokens per word (TpW) for the preamble of the Constitution of the Republic of Poland in Polish and English versions, processed by various tokenizers.

Mistral 7B v0.2 model, however, was not specifically trained on or optimized for the Polish language.

To establish a relevant baseline for comparative analysis, we selected the Preamble of the Constitution of the Republic of Poland as our benchmark text. This text was chosen for its comprehensive representation of formal Polish writing, and critically, because it possesses an official English translation, allowing for direct comparison. Table 2 provides a detailed comparison of key metrics, including token count, characters per token (CpT), and tokens per word (TpW), to illustrate the performance of different tokenizers when applied to both the Polish and English versions of the preamble.

Ultimately, we elected to retain the original tokenizer derived from the Mistral 7B v0.2 model, which encompasses a base vocabulary of 32,000 tokens. The sole modification introduced was the necessary addition of special tokens required for the standardized chat template, increasing the total vocabulary size to 32,128 tokens, which aligns precisely with the tokenizer used in the preceding Bielik 11B v2.6 version [24]. The rationale for maintaining this specific tokenizer was twofold: first, the Bielik 11B v3 model is designed to support over 30 languages; and second, the current tokenizer provides a satisfactory trade-off between tokenization efficiency for Polish and its performance across other supported languages, including English, while maintaining a relatively small vocabulary size.

3. Pre-training

The central objective of the pre-training phase was to substantially enhance the model's linguistic proficiency, specifically targeting high levels of accuracy and fluency across Polish and other major European languages. To realize this goal, a heterogeneous corpus comprising high-quality textual data was employed. These materials underwent a rigorous preprocessing pipeline and comprehensive quality evaluation to ensure optimal standards for the training data utilized in the subsequent fine-tuning processes.

3.1. Pre-training Data

The pre-training phase for Bielik v3 involved a substantial expansion in both the volume and variety of training corpora compared to its predecessor, Bielik 11B v2 [26]. Specifically, the total token count was increased fivefold, encompassing a multilingual corpus across 32 languages (detailed in Table 3). This newly curated dataset integrates a broad spectrum of high-quality sources, including:

- **Legal and Official Records:** Court rulings, legislative acts, regulations, and official resolutions.
- **Academic and Press:** Scientific literature from the Science Library and commercial press publications.
- **Public Discourse:** Parliamentary proceedings and thematic web forums.
- **Cultural and Digital Repositories:** Content from Wikipedia, Europeana, and public web domains.
- **Synthetic Data:** Purpose-built datasets generated to augment training.

A primary emphasis was placed on Polish and European contexts. This includes specialized efforts to capture regional nuances within Poland, such as the integration of Silesian and Kashubian Wikipedia editions. The collection also draws from established repositories like the Parliamentary Discourse Corpus (*Korpus Dyskursu Parlamentarnego*), the Science Library (*Biblioteka Nauki*), and various anonymized journals of laws and court gazettes.

3.1.1. Data Processing and Quality Control

To ensure the integrity of the model, we implemented a rigorous preprocessing pipeline. While the corpus includes copyright-protected materials and documents that may contain personal data, a robust anonymization mechanism is standard for all pipelines. This system identifies and removes sensitive identifiers, including PESEL numbers, phone numbers, email addresses, and URLs.

Furthermore, we applied deduplication and quality-filtering heuristics to minimize redundancy. These measures are applied proportionally, adjusting for the specific licensing terms and the nature of the data source (e.g., API-based access versus public dumps). Detailed methodologies for these filtering processes are documented in the technical reports for our previous Bielik 11B v2 model [26] and our smaller Bielik v3 models [25].

Language	Number of documents	Share of dataset
Polish	428.6M	54.25%
English	162.0M	20.50%
Dutch	12.8M	1.62%
Portuguese	12.2M	1.55%
Swedish	11.9M	1.50%
German	11.5M	1.45%
Spanish	10.7M	1.36%
Italian	10.3M	1.31%
Turkish	9.7M	1.22%
French	9.7M	1.22%
Czech	8.6M	1.09%
Romanian	8.6M	1.09%
Bosnian	8.4M	1.06%
Finnish	8.1M	1.02%
Hungarian	7.5M	0.95%
Greek	6.5M	0.82%
Ukrainian	6.2M	0.79%
Slovenian	6.2M	0.78%
Danish	6.0M	0.76%
Russian	5.9M	0.74%
Croatian	5.9M	0.74%
Bulgarian	5.6M	0.71%
Slovak	5.6M	0.71%
Lithuanian	5.2M	0.66%
Norwegian	5.2M	0.65%
Estonian	3.9M	0.49%
Latvian	3.6M	0.46%
Albanian	1.1M	0.14%
Serbian	1.1M	0.14%
Belarusian	920.1K	0.12%
Icelandic	759.4K	0.10%
Serbo-Croatian	84.0K	0.01%

Table 3: Language distribution of documents used in pretraining the Bielik 11B v3 model

3.1.2. Ethical Compliance and Data Rights

Our curation process adheres to strict ethical and legal standards. We systematically monitor `robots.txt` files, meta-tag-based Text and Data Mining (TDM) reservations, and Terms of Service (ToS) declarations. Any source expressing a reservation of rights is excluded from the corpus.

This exclusion policy is applied retroactively: should a data provider declare a reservation after initial collection, the relevant data is purged from all future iterations of the model. This

protocol extends to our own publicly released datasets; if a manifest explicitly prohibits AI training, those datasets are excluded and the restriction is publicly recorded.

3.1.3. Linguistic Distribution and Sources

The dataset is predominantly Polish and English, supplemented by other European languages. The collection effort, which commenced in November 2022, remains an ongoing initiative. The final training mix involves a sophisticated blending of various web domains, further refined through deduplication. Key external resources utilized include:

- **CulturaX** [23] and **HPLT v2.0** [5] for multilingual depth.
- **FineWeb** [30] and **FineWeb-Edu** [20] for high-quality web content.
- **SlimPajama-627B** [41] and **Common Crawl**[1] for large-scale general knowledge.

3.2. Pre-training Stages

The Bielik 11B v3 model was pre-trained through a three-stage process:

1. **General Stage:** In the first pre-training stage, the model was trained on over 1 trillion tokens using a context length of $8,192$ tokens. During this stage, the models achieved full pre-training on language proficiency and general world knowledge, with training data covering multiple languages.
2. **Full Context Stage:** To increase the context length to the full size of $32,768$ tokens, we optimized the pre-training corpus for this stage by increasing the proportion of long documents. Specifically, we selected only documents with a length exceeding $7,168$ tokens. The models were further pre-trained with approximately 50 billion tokens at a context length of $32,768$ tokens. We also increased the learning rate at this stage back to the initial value of 2.5×10^{-5}.
3. **Long Context Stage:** In the final pre-training stage, we introduced YARN [31] and collected a high-quality long context corpora to extend the model's context length. The model was trained on over 1 billion tokens using a maximum context length of $65,536$ tokens.

At the conclusion of each stage, we merged selected recent checkpoints, which helped stabilize the model's performance. The selection process involved choosing checkpoints that achieved the highest scores on evaluation benchmarks. We then created various combinations of these selected checkpoints for merging, subsequently evaluated the merged models using the same benchmarks, and finally chose the best-performing one for continuation of the training process.

3.3. Pre-training Hyperparameters

The optimization process for the model was driven by the AdamW optimizer [19]. We configured the optimizer with decay coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.95$, alongside a weight decay coefficient of 0.1. To regulate the learning rate, we implemented a cosine decay schedule

---

[1]https://commoncrawl.org/

that peaked at 2.5×10^{-5} and tapered to a minimum of 9×10^{-6}. This schedule was preceded by a linear warm-up phase spanning the initial 50 iterations. The entire training run was executed over a total of $270,000$ iterations, processing an aggregate of 1.1 trillion tokens.

The training was divided into distinct phases to manage computational efficiency and context scaling:

- **Baseline Phase:** Initial training was performed with a sequence length of $8,192$ tokens and a global batch size of 256. During this stage, tensor parallelism was not utilized.
- **Context Expansion Phase:** To extend the model's capabilities to $32,768$ and $65,536$ tokens, we enabled tensor parallelism and integrated gradient checkpointing to manage memory constraints.

Stability and efficiency were further ensured by setting a gradient clipping threshold of 1.0 and utilizing mixed-precision training with the bfloat16 format.

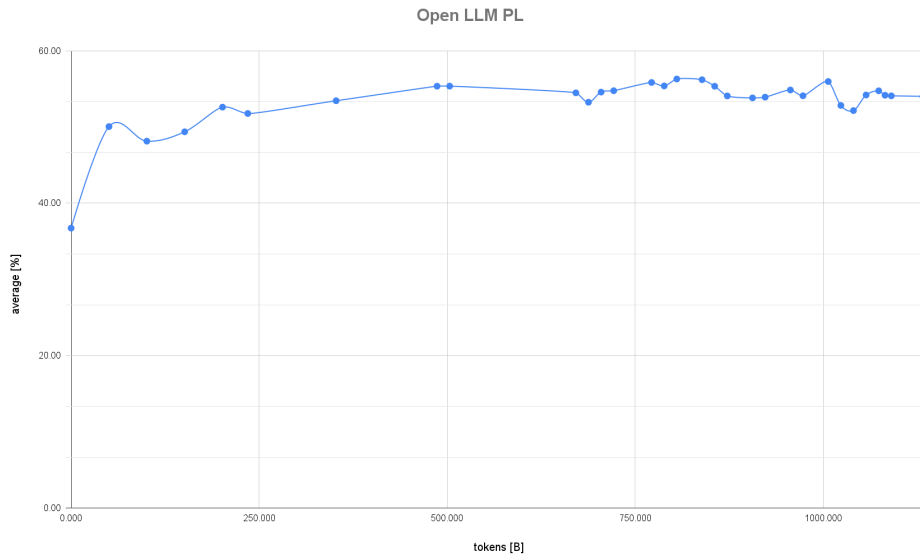## 3.4. Pre-training Process Monitoring

Figure 2. Average score of the Bielik 11B v3 model checkpoints during pretraining in the Open PL LLM benchmark

To guarantee the desired quality throughout the pre-training execution, we diligently tracked the model's performance evolution on the Open LLM PL [47] and Open LLM EN [3] benchmark suites. Maintaining performance on both of these benchmarks is crucial for preventing model degradation. Model checkpoints were captured at predefined intervals during training and subsequently subjected to performance evaluation, as shown in Figure 2. Our observations indicate that a decline in benchmark scores over the training period typically

signals problems within the dataset and seldom proves reversible. Furthermore, a significant deterioration in either benchmark usually serves as an early indicator of wider model quality issues.

The performance assessment, utilizing the specified benchmarks, was executed concurrently with the core training on an isolated cluster of compute nodes to avoid resource contention. This strategy facilitated the prompt identification of training anomalies, eliminating the need to wait for full epoch completion and thus yielding considerable efficiencies in computational resources.

4. Post-training

Following the initial pre-training phase, we executed a rigorous post-training pipeline designed to bolster the model's capabilities in logical reasoning, mathematics, code generation, and complex instruction following. This stage followed a three-tier training protocol, mirroring the methodology developed for the smaller Bielik v3 model variants [25].

4.1. Post-training Data

To address the scarcity of high-quality instruction-tuning data for the Polish language, we developed a proprietary dataset that undergoes continuous expansion and refinement by human annotators. This corpus is manually curated through the formulation of diverse instructions and multi-turn dialogues, ensuring the content is culturally relevant and linguistically precise for Polish and other European languages.

Furthermore, we utilized documents from the pre-training phase to synthesize instructions and dialogues via controlled prompting techniques. This approach ensures a broad representation of task types, inquiry styles, and domain-specific nuances. Throughout this synthetic generation process, we maintained a strict focus on factual consistency, linguistic naturalness, and structural clarity to match the standards of high-quality, human-authored data.

4.2. Supervised Fine-Tuning

The Supervised Fine-Tuning (SFT) stage is designed to achieve several critical objectives beyond basic task execution. It enables the model to identify and adapt to various registers and levels of formality, maintain long-range coherence in multi-turn conversations, and respond accurately to system-level prompts. Additionally, the model is trained to support structured outputs and function calling (tool use).

To optimize the SFT process, we implemented the following technical strategies:

- **Constant Learning Rate and Weight Decay:** Employed to maintain stable optimization throughout the epoch cycles.
- **Instruction Masking:** We utilized a selective masking strategy where the loss function is calculated only on the model's responses. By masking user instructions and control tokens [40], we ensured the gradient updates remained focused on content generation rather than predicting the prompt structure.

- **Sample Packing and FlexAttention:** To maximize computational throughput, we employed sample packing to concatenate multiple sequences up to the maximum context window, significantly outperforming traditional padding methods [46]. We integrated `FlexAttention` [10] to generate attention masks dynamically, preventing cross-contamination between independent dialogues within a single packed sequence.
- **Selective Gradient Checkpointing:** This was applied to specific layers to achieve an optimal balance between memory conservation and processing speed.
- **Checkpoint Merging:** We performed weight averaging (merging) across various SFT training stages. This resulted in a final model that demonstrated superior evaluation performance compared to any individual intermediate checkpoint.

4.2.1. Supervised Fine-Tuning Hyperparameters

Optimization was performed using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$) with a weight decay of 0.05. We maintained a constant learning rate of 5×10^{-6} following a 100-iteration linear warmup. The model was trained for 3 epochs on 20 million instructions and dialogues using a global batch size of 64 and a maximum sequence length of $32,768$ tokens. Training utilized `bfloat16` mixed precision with a gradient clipping threshold of 1.0.

4.3. Preference Learning

To align the Bielik 11B v3 model with human expectations, we employed **Direct Preference Optimization (DPO)** and its variant, **DPO-Positive (DPO-P)**, which proved highly effective in previous iterations of the Bielik series.

4.3.1. Preference Dataset

Compared to previous versions of the Bielik model, we introduced a series of substantial upgrades to our Polish preference instruction corpus and the associated response generation framework. We expanded the dataset to over $114,000$ examples, prioritizing a more sophisticated distribution of tasks.

Key enhancements in the dataset composition include:

- **Domain Expansion:** We integrated a higher volume of instructions focused on logical reasoning and mathematics.
- **Advanced Capabilities:** The alignment categories were broadened to explicitly include function calling and tool-use scenarios.
- **Interaction Realism:** To better simulate authentic user behavior, we incorporated an extensive collection of multi-turn conversational dialogues.

For the generation of contrastive response pairs (preferred vs. rejected), we diversified our model ensemble. In addition to our legacy pipeline, we employed several state-of-the-art models, most notably DeepSeek-V3-0324 and Bielik 11B v2.6.

Despite the increase in scale, our core philosophy regarding data integrity remains unchanged. We continue to employ a hybrid curation strategy that balances manual authorship and perturbation-based augmentation with automated safeguards. This strategy includes rigorous deduplication protocols, quality assessment via reward metamodels, and thorough

manual audits. Collectively, these refinements ensure that the dataset provides a high-fidelity signal for downstream alignment and preference learning.

4.3.2. DPO-P Hyperparameters

The DPO-P phase utilized the AdamW optimizer with a constant learning rate of 5×10^{-7} and 50 warmup iterations. Training was conducted over 3 epochs on the $114,000$ samples with a context window of $32,768$ tokens and a global batch size of 64. Consistent with the SFT phase, we employed `bfloat16` mixed precision and a gradient clipping norm of 1.0.

4.4. Reinforcement Learning

In the final stage of our alignment pipeline, we utilized Reinforcement Learning (RL) to further sharpen the model's analytical capabilities. We implemented Group Relative Policy Optimization (GRPO) [37] alongside its variant, Dr. GRPO [18]. The latter was specifically selected to improve token efficiency by mitigating the common issue where models artificially inflate response length to maximize rewards.

The RL training was performed using the Volcano Engine Reinforcement Learning (VERL) framework [39], which provided a scalable and modular environment for our workflows. The training corpus consisted of 143k curated problems across logic, STEM, mathematics, and tool-use domains. Each sample was selected based on the availability of Reinforcement Learning from Verifiable Rewards (RLVR), ensuring that every problem had a definitive, checkable solution.

To facilitate accurate reward modeling, we developed specialized verifiers tailored to the task type:

- **Mathematics:** Rewards were granted based on the extraction of the correct final answer, which was required to be enclosed within a LATEX `\boxed{...}` command.
- **STEM:** For multiple-choice questions, the verifier assessed whether the model selected the single correct option from the available choices.
- **Tool Use:** The reward system validated whether the model invoked the appropriate function calls with the correct and expected arguments.

The integration of Reinforcement Learning significantly boosted the model's proficiency in mathematical tasks and broadly enhanced its general reasoning and problem-solving faculties. These results indicate that this training phase effectively generalizes across diverse domains, reinforcing the model's ability to follow complex logical paths beyond simple instruction following.

4.4.1. Reinforcement Learning Hyperparameters

Our optimization configuration utilized a learning rate of 1×10^{-6} and a global batch size of 128 (distributed as local batches of 16 per GPU). To maintain stability during policy updates and prevent catastrophic forgetting, we applied KL divergence regularization with a coefficient of 0.001, employing a low-variance KL loss formulation.

5. Evaluation

To comprehensively assess the capabilities of Bielik v3 models, we conducted extensive evaluations across multiple benchmarks covering diverse aspects of language understanding, generation, and reasoning. Our evaluation strategy encompasses both Polish-specific and multilingual benchmarks to demonstrate the models' proficiency in handling various linguistic tasks.

The models were evaluated on the following benchmarks:

- Open PL LLM Leaderboard **(Polish)**
- Polish EQ-Bench **(Polish)**
- CPTUB Leaderboard **(Polish)**
- Polish Medical Leaderboard **(Polish)**
- Polish Linguistic and Cultural Competency Benchmark (PLCC) **(Polish)**
- Open LLM Leaderboard
- include-base-44
- belebele
- flores

5.1. Open PL LLM Leaderboard

The Open PL LLM Leaderboard [47, 27] provides a comprehensive assessment of language models across a diverse range of Polish NLP tasks. Built upon the foundation of Open LLM Leaderboard v1 [4], this benchmark evaluates core language understanding capabilities including sentiment classification, named entity recognition, topic categorization, reading comprehension, and question answering. The evaluation framework employs the lm-evaluation-harness toolkit [11] and primarily focuses on discrete task performance rather than conversational interaction abilities.

Tasks:

- **polemo2:** Sentiment analysis of online consumer reviews across four domains (medicine, hotels, products, university) with four-class labeling (positive, negative, neutral, ambiguous) [16]; metric: accuracy.
- **klej-ner:** Named entity recognition in sentences containing single-type entities, classifying into six categories (no entity, place, person, organization, time, geographical name) [35]; metric: accuracy.
- **8tags:** Topic classification of social media headlines into eight categories (film, history, food, medicine, motorization, work, sport, technology) [8]; metric: accuracy.
- **belebele:** Machine reading comprehension for question answering [2]; metric: accuracy.
- **dyk:** Question answering based on human-annotated pairs from Wikipedia's "Did You Know" section [21]; metric: binary F1.
- **ppc:** Text similarity assessment using manually labeled sentence pairs (exact paraphrases, close paraphrases, non-paraphrases) [6]; metric: accuracy.
- **psc:** Summarization of news articles [28]; metric: binary F1.

- **cbd:** Text classification for cyberbullying and hate-speech detection [33]; metric: macro F1.
- **polqa:** Open-domain question answering from the "Jeden z dziesięciu" TV show, with and without context (abstractive QA/RAG) [36]; metric: accuracy, levenshtein.
- **poquad:** Context-based extractive question answering (QA/RAG) [44]; metric: levenshtein.
- **eqbench:** emotional intelligence benchmark [29]; metric: custom.

The majority of benchmark tasks employ a multiple-choice format where models select from predefined answer options. Two distinct evaluation methodologies are applied:

- **Loglikelihood:** Models select the option with the highest token probability from the available choices (e.g., A, B, C, D). This approach is particularly well-suited for evaluating base models without instruction tuning.
- **Generate:** Models produce free-form text responses, testing their generation capabilities.

Each task undergoes evaluation in both 0-shot (no examples provided) and 5-shot (five examples given) configurations, with final scores representing the average performance across all tasks, normalized against established baselines. Tables 4 and 5 present the results for base and instruction-tuned models, respectively.

As shown in Table 4, the Bielik-11B-v3-Base-20250730 model achieves a score of 55.16, demonstrating competitive performance for a base model of its size. The instruction-tuned variant, Bielik-11B-v3.0-Instruct, significantly outperforms its base counterpart with a score of 65.93 (Table 5), ranking among the top models and surpassing several much larger models including Meta-Llama-3.1-70B-Instruct and Mixtral-8x22B-Instruct-v0.1.

5.2. Polish EQ-Bench

The Polish Emotional Intelligence Benchmark represents a culturally adapted Polish adaptation of the EQ-Bench framework [29]. This benchmark assesses language models' ability to recognize, interpret, and reason about emotional states and interpersonal dynamics. The evaluation encompasses multiple facets of emotional intelligence, including emotion recognition in context, understanding of emotional implications, and sensitivity to nuanced affective states in conversational scenarios. Results are presented in Table 6.

Bielik-11B-v3.0-Instruct achieves a score of 71.20 on the Polish EQ-Bench (Table 6), demonstrating strong emotional intelligence capabilities. While this represents a slight decrease compared to the previous version Bielik-11B-v2.6-Instruct (73.8), the v3.0 model maintains competitive performance, placing it among models with substantially larger parameter counts such as Llama-3.3-70B-Instruct (70.73) and Qwen2.5-32B-Instruct (71.15).

5.3. Complex Polish Text Understanding Benchmark (CPTUB)

CPTUB [42] presents a sophisticated evaluation framework targeting advanced comprehension capabilities in Polish language processing. In contrast to conventional benchmarks that primarily test literal interpretation, CPTUB probes deeper cognitive abilities including inference from context, pragmatic understanding, and reasoning under ambiguity. The benchmark structure encompasses two primary evaluation dimensions:

Model	Parameters (B)	Average
Qwen2.5-72B	72.7	67.38
Qwen2.5-32B	32.8	66.73
Qwen-72B	72.7	66.02
Qwen2.5-14B	14.8	62.71
Meta-Llama-3-70B	70.6	62.07
Qwen1.5-72B	72.7	61.11
Meta-Llama-3.1-70B	70.6	60.87
Mixtral-8x22B-v0.1	141.0	60.75
Mistral-Small-24B-Base-2501	24.0	59.90
Qwen1.5-32B	32.8	58.71
Bielik-11B-v2	11.2	58.14
Llama-4-Scout-17B-16E	109.0	57.77
Bielik-11B-v3-Base-20250730	**11.2**	**55.16**
Qwen2.5-7B	7.0	53.35
EuroLLM-9B	9.2	50.03
Qwen-7B	7.0	49.39
SOLAR-10.7B-v1.0	10.7	47.54
Mistral-Nemo-Base-2407	12.2	47.28
internlm2-20b	20.0	47.15
Bielik-4.5B-v3	4.8	45.47
Meta-Llama-3.1-8B	8.0	43.77
Meta-Llama-3-8B	8.0	43.30
Qwen1.5-72B	72.3	39.51
Mistral-7B-v0.3	7.0	38.88
Mistral-7B-v0.2	7.0	38.81
Qwen1.5-7B	7.0	37.92
Bielik-7B-v0.1	7.2	34.34
Qra-13b	13.0	33.90
Llama-3.2-3B	3.0	31.89
Qra-7b	7.0	16.60

Table 4: Open PL LLM Leaderboard results for base models (5-shot evaluation)

- **Implicatures**: This component measures models' competence in decoding non-literal meanings and contextual implications. It examines understanding of figurative language, ironic expressions, and idiomatic constructions through three distinct evaluation categories:

  - **Sentiment**: Assessing the ability to discern emotional valence that diverges from surface-level lexical content
  - **Language understanding**: Testing comprehension of communicative intent and pragmatic meaning
  - **Phraseology**: Evaluating knowledge of conventionalized multi-word expressions where compositional semantics fail

- **Tricky Questions**: This section challenges models with adversarially constructed queries featuring logical paradoxes, semantic ill-formedness, contradictions, absurdist premises, and humorous misdirection. It specifically probes reasoning robustness and the model's tendency to produce plausible-sounding but incorrect responses when confronted with problematic inputs.

Table 7 presents the comprehensive results across all CPTUB evaluation categories.

Model	Parameters (B)	Average
Mistral-Large-Instruct-2411	123.0	69.84
Meta-Llama-3.1-405B-Instruct-FP8	405.0	69.44
Mistral-Large-Instruct-2407	123.0	69.11
Qwen2.5-72B-Instruct	72.7	67.92
QwQ-32B-Preview	32.8	67.01
Llama-3.3-70B-Instruct	70.6	66.40
Bielik-11B-v3.0-Instruct	**11.2**	**65.93**
Qwen2-72B-Instruct	72.7	65.87
<u>Bielik-11B-v2.3-Instruct</u>	<u>11.2</u>	<u>65.71</u>
<u>Bielik-11B-v2.2-Instruct</u>	<u>11.2</u>	<u>65.57</u>
Meta-Llama-3.1-70B-Instruct	70.6	65.49
<u>Bielik-11B-v2.1-Instruct</u>	<u>11.2</u>	<u>65.45</u>
Mixtral-8x22B-Instruct-v0.1	141.0	65.23
<u>Bielik-11B-v2.0-Instruct</u>	<u>11.2</u>	<u>64.98</u>
Meta-Llama-3-70B-Instruct	70.6	64.45
<u>Bielik-11B-v2.6-Instruct</u>	<u>11.2</u>	<u>64.26</u>
Qwen3-32B	32.8	64.24
Llama-4-Scout-17B-16E-Instruct	109.0	64.21
<u>Bielik-11B-v2.5-Instruct</u>	<u>11.2</u>	<u>63.95</u>
Mistral-Small-24B-Instruct-2501	24.0	62.97
phi-4	14.7	62.57
Qwen3-14B	14.8	62.24
gemma-3-12b-it	12.0	62.20
Mistral-Small-Instruct-2409	22.2	61.41
Qwen2.5-32B-Instruct	32.8	61.21
Qwen2.5-14B-Instruct	14.8	59.91
aya-23-35B	35.0	56.37
<u>Bielik-4.5B-v3.0-Instruct</u>	<u>4.8</u>	<u>56.13</u>
gemma-3-27b-it	27.0	55.92
Qwen3-8B	8.2	55.78
Qwen3-4B	4.0	55.49
Mistral-Nemo-Instruct-2407	12.2	55.27
EuroLLM-22B-Instruct-Preview	22.0	55.17
Qwen2.5-7B-Instruct	7.6	54.93
EuroLLM-9B-Instruct	9.0	50.07
GaMS-9B-Instruct	9.0	48.78
Mistral-7B-Instruct-v0.3	7.2	47.74
Apertus-8B-Instruct-2509	8.0	47.27
Mistral-7B-Instruct-v0.2	7.2	45.95
<u>Bielik-7B-Instruct-v0.1</u>	<u>7.2</u>	<u>44.70</u>
gemma-2-9b-it	9.0	42.12
Qwen2.5-3B-Instruct	3.0	41.23
Mistral-7B-Instruct-v0.1	7.0	33.11
Qwen2.5-1.5B-Instruct	1.5	31.89

Table 5: Open PL LLM Leaderboard results for instruction-tuned models (5-shot evaluation)

Model	Parameters (B)	Score
Mistral-Large-Instruct-2407[†]	123.0	78.07
Mistral-Large-Instruct-2411[†]	123.0	77.29
Meta-Llama-3.1-405B-Instruct-FP8	405.0	77.23
gpt-4o-2024-08-06	Unknown	75.15
gpt-4-turbo-2024-04-09	Unknown	74.59
Bielik-11B-v2.6-Instruct	11.2	73.8
Mistral-Small-Instruct-2409	22.2	72.85
Llama-PLLuM-70B-chat	70.6	72.56
Meta-Llama-3.1-70B-Instruct	70.6	72.53
DeepSeek-V3-0324	685.0	73.46
Bielik-11B-v2.5-Instruct	11.2	72.00
Qwen2-72B-Instruct	72.7	71.23
Meta-Llama-3-70B-Instruct	70.6	71.21
Bielik-11B-v3.0-Instruct	**11.2**	**71.20**
gpt-4o-mini-2024-07-18	Unknown	71.15
Qwen2.5-32B-Instruct	32.8	71.15
Bielik-11B-v2.3-Instruct	11.2	70.86
Llama-3.3-70B-Instruct	70.6	70.73
Llama-PLLuM-70B-instruct	70.6	69.99
WizardLM-2-8x22B	141.0	69.56
Qwen2.5-14B-Instruct	14.8	69.17
Bielik-11B-v2.2-Instruct	11.2	69.05
Bielik-11B-v2.0-Instruct	11.2	68.24
glm-4-9b-chat	9.0	61.79
Mistral-Nemo-Instruct-2407	12.2	61.76
Bielik-11B-v2.1-Instruct	11.2	60.07
pllum-12b-nc-chat-250715	12.2	55.20
EuroLLM-9B-Instruct	9.2	54.10
Bielik-4.5B-v3.0-Instruct	4.8	53.58
PLLuM-12B-chat	12.2	52.26
PLLuM-8x7B-nc-chat[†]	46.7	47.29
Llama-PLLuM-8B-chat	8.0	46.20
PLLuM-8x7B-chat	46.7	45.22
PLLuM-12B-nc-chat[†]	12.2	35.41

[†]Models with a non-commercial license.

Table 6: Polish EQ-Bench results for various models.

Model	Params (B)	Overall Average	Implicatures Average	Senti-ment	Language Understanding	Phrase-ology	Tricky Questions
gemini-2.0-flash-001	4.29	4.39	4.52	4.32	4.34	3.99	
DeepSeek-R1	685.0	4.14	4.14	4.49	4.35	3.60	4.12
gemini-2.0-flash-lite-001	4.09	4.17	4.23	4.05	4.24	3.85	
DeepSeek-V3-0324	685.0	4.03	4.03	4.36	4.20	3.54	4.02
Mistral-Large-Instruct-2411[†]	123.0	4.00	4.10	4.33	3.98	3.99	3.72
Qwen2.5-72B-Instruct	72.7	3.95	3.99	4.08	3.97	3.93	3.81
Mistral-Large-Instruct-2407[†]	123.0	3.93	4.03	4.23	4.00	3.86	3.65
Llama-4-Maverick-17B-128E-Instruct	402.0	3.93	3.99	4.39	4.11	3.48	3.76
gemma-3-27b-it	27.4	3.81	3.90	3.88	3.79	4.03	3.53
Meta-Llama-3-70B-Instruct	70.6	3.78	3.81	4.13	3.82	3.47	3.71
Qwen2.5-32B-Instruct	32.8	3.75	3.80	3.81	3.57	4.04	3.59
Llama-4-Scout-17B-16E-Instruct	109.0	3.75	3.94	4.10	3.81	3.90	3.19
Bielik-11B-v3.0-Instruct	**11.2**	**3.73**	**3.92**	**3.88**	**3.91**	**3.96**	**3.19**
Mistral-Small-24B-Instruct-2501	23.6	3.71	3.80	3.91	3.60	3.88	3.45
pllum-12b-nc-chat-250715[†]	12.2	3.67	3.92	4.36	3.96	3.46	2.90
Bielik-11B-v2.6-Instruct	11.2	3.64	3.82	4.10	3.94	3.41	3.10
Mixtral-8x22B-Instruct-v0.1	141.0	3.56	3.67	3.78	3.68	3.55	3.24
Qwen2.5-14B-Instruct	14.8	3.55	3.62	3.91	3.57	3.37	3.34
Llama-PLLuM-70B-chat	70.6	3.53	3.63	3.94	3.61	3.35	3.21
Bielik-4.5B-v3.0-Instruct	4.8	3.38	3.68	3.76	3.61	3.67	2.46
phi-4	14.7	3.30	3.50	3.72	3.54	3.24	2.72
PLLuM-12B-chat	12.2	3.14	3.32	3.32	3.21	3.43	2.59
PLLuM-8x7B-nc-instruct[†]	46.7	3.11	3.56	3.88	3.59	3.22	1.76
Qwen2.5-7B-Instruct	7.62	3.07	3.23	3.56	3.03	3.10	2.58
EuroLLM-9B-Instruct	9.0	3.15	3.28	3.37	3.30	3.17	2.75
PLLuM-8x7B-nc-chat[†]	46.7	3.03	3.44	3.76	3.48	3.08	1.80
Meta-Llama-3.1-8B-Instruct	8.0	3.01	3.31	3.97	3.38	2.58	2.11
PLLuM-8x7B-chat	46.7	3.01	3.41	3.44	3.45	3.35	1.78
Meta-Llama-3-8B-Instruct	8.0	3.00	3.17	3.33	3.15	3.04	2.48
Llama-PLLuM-8B-chat	8.0	2.92	3.14	3.13	2.93	3.36	2.25
Bielik-7B-Instruct-v0.1	7.2	2.88	3.13	3.59	3.48	2.32	2.16

[†]Models with a non-commercial license.

Table 7: Complex Polish Text Understanding Benchmark (CPTUB) results across different evaluation categories

On CPTUB (Table 7), Bielik-11B-v3.0-Instruct achieves an overall average of 3.73, ranking competitively among models evaluated. The model performs particularly well on implicature understanding with an average of 3.92, demonstrating strong capabilities in language understanding (3.91), sentiment analysis (3.88), and phraseology (3.96). The tricky questions component yields a score of 3.19, reflecting the challenging nature of these adversarial queries. This performance places Bielik-11B-v3.0-Instruct ahead of several larger models including Qwen2.5-14B-Instruct and Mixtral-8x22B-Instruct-v0.1, while approaching the performance of frontier models with significantly higher parameter counts.

5.4. Polish Medical Leaderboard

The Polish Medical Leaderboard provides a domain-specific assessment of language models using authentic questions from the Polish State Specialization Examination (Państwowy Egzamin Specjalizacyjny, PES) spanning 2018-2022. This benchmark measures both medical domain knowledge and clinical reasoning abilities within the Polish healthcare context. The evaluation employs the speakleash/PES-2018-2022 dataset, derived from amu-cai/PES-2018-2022 [32], and tests models' capacity to apply medical knowledge in scenarios similar to those encountered by medical professionals seeking board certification. Results are shown in Table 8.

On the Polish Medical Leaderboard (Table 8), Bielik-11B-v3.0-Instruct achieves 50.21%, demonstrating substantial medical knowledge and clinical reasoning capabilities. This represents a significant improvement over the base model Bielik-11B-v3-Base-20250730 (45.86%), highlighting the effectiveness of instruction tuning for specialized domain tasks. These results demonstrate Bielik's capability to handle domain-specific knowledge in the medical field when evaluated in Polish.

5.5. Polish Linguistic and Cultural Competency Benchmark (PLCC)

PLCC [7] extends evaluation beyond linguistic proficiency to encompass cultural knowledge deeply embedded in Polish society. This benchmark features 600 manually curated questions spanning six thematic domains: historical events, geographical knowledge, cultural traditions, artistic and entertainment heritage, grammatical correctness, and lexical richness.

The evaluation probes understanding of culturally-specific references, including historical milestones, traditional customs, folklore narratives, literary works, and contemporary popular culture–all elements crucial for authentic communication and comprehension in Polish contexts. Question difficulty varies from widely recognized facts to regionally-specific knowledge, employing both multiple-choice and open-ended response formats that demand precise factual recall of dates, proper nouns, and conceptual relationships. Performance results are presented in Table 9.

Bielik-11B-v3.0-Instruct achieves an exceptional score of 71.83% on PLCC (Table 9), demonstrating superior cultural and linguistic knowledge compared to other open-source models evaluated. This represents a substantial improvement over the previous Bielik-11B-v2.6-Instruct (65.50%) and significantly surpasses much larger models including DeepSeek-V3-0324 (71.00%, 685B parameters), Llama-3.1-Tulu-3-405B (63.83%). These results underscore

Model	Parameters (B)	Average (%)
Meta-Llama-3.1-405B-Instruct-FP8	405.0	69.20
Mistral-Large-Instruct-2407[†]	123.0	64.28
Qwen2.5-72B-Instruct	72.7	63.89
Meta-Llama-3.1-70B-Instruct	70.6	61.75
Qwen2-72B-Instruct	72.7	61.35
Meta-Llama-3-70B-Instruct	70.6	57.51
Qwen2.5-32B	32.8	55.69
Qwen2.5-32B-Instruct	32.8	54.52
Bielik-11B-v3.0-Instruct	**11.2**	**50.21**
Qwen2.5-14B-Instruct	14.8	49.60
Bielik-11B-v3-Base-20250730	**11.2**	**45.86**
Bielik-11B-v2.6-Instruct	11.2	44.88
Bielik-11B-v2.5-Instruct	11.2	44.85
GLM-4-9b-chat	9.0	44.54
Mistral-Small-Instruct-2409	22.2	43.60
Bielik-4.5B-v3.0-Instruct	4.8	43.55
Bielik-11B-v2.3-Instruct	11.2	43.26
Bielik-11B-v2.1-Instruct	11.2	43.16
Bielik-11B-v2.2-Instruct	11.2	43.05
Qwen2.5-7B-Instruct	7.6	42.69
Bielik-11B-v2.0-Instruct	11.2	41.53
Meta-Llama-3.1-8B-Instruct	8.0	40.60
Mistral-Nemo-Instruct-2407	12.2	40.36
Bielik-11B-v2	11.2	39.98
PLLuM-12B-nc-chat-250715[†]	12.2	38.53
PLLuM-12B-chat	12.2	36.51
EuroLLM-9B-Instruct	9.0	35.96
Mistral-7B-Instruct-v0.3	7.0	31.24
Bielik-7B-Instruct-v0.1	7.2	29.74

[†]Models with a non-commercial license.

Table 8: Polish Medical Leaderboard results (5-shot setting) showing model performance on Polish Board Certification Examinations.

Bielik's particular strength in capturing Polish-specific cultural knowledge, a critical capability for applications requiring deep understanding of Polish context.

5.6. Open LLM Leaderboard

The Open LLM Leaderboard [3] serves as a comprehensive English-language evaluation suite, assessing models across diverse tasks including commonsense reasoning (ARC challenge, HellaSwag, WinoGrande), factual accuracy (TruthfulQA), broad knowledge (MMLU), and mathematical reasoning (GSM8K). This benchmark provides crucial insights into multilingual models' English language capabilities, which is particularly important for European models

like Bielik that aim to balance strong native language performance with English proficiency. Tables 10 and 11 present results for base and instruction-tuned models, respectively.

Model	Parameters (B)	Average Score (%)
Gemini-3.0-Pro-Preview	Unknown	95.83
Gemini-2.5-Pro-Exp-03-25	Unknown	89.50
Claude-3.7-Sonnet	Unknown	81.50
GPT-5.1-2025-11-13 (default reasoning)	Unknown	77.83
Gemini-2.0-Flash-Experimental	Unknown	74.17
O4-Mini-2025-04-16	Unknown	72.83
Bielik-11B-v3.0-Instruct	**11.2**	**71.83**
Kimi-K2-Thinking	1000	71.67
Claude-Sonnet-4.5	Unknown	71.00
DeepSeek-v3-0324	685.0	71.00
GLM-4.6	355	70.67
Grok-4-Fast	Unknown	70.17
Gemini-Pro-1.5	Unknown	69.67
PLLuM-12B-nc-chat-250715[†]	12.2	69.67
DeepSeek-v3	685.0	69.17
PLLuM-8x7B-nc-chat[†]	46.7	68.17
Bielik-11B-v2.6-Instruct	11.2	65.50
GPT-4.1-mini-2025-04-14	Unknown	62.17
Llama-3.1-405B	405.0	60.00
PLLuM-12B-nc-chat[†]	12.2	59.50
Llama-PLLuM-70B-chat	70.6	58.50
Llama-4-Maverick	402.0	58.17
Command-A-03-2025[†]	111.0	56.17
Mistral-Large-2407[†]	123.0	54.17
PLLuM-8x7B-chat	46.7	54.17
WizardLM-2-8x22B	141.0	51.50
Qwen-Max	Unknown	50.83
Command-R-Plus-08-2024[†]	Unknown	50.17
Mixtral-8x22B	141.0	49.83
Llama-3.3-70B	70.6	48.83
Gemma-3-27B	27.4	47.33
PLLuM-12B-chat	12.2	47.00
Bielik-7B-Instruct-v0.1	7.2	46.67
Mistral-Small-3.1-24B-2503	24.0	43.33
Llama-3.0-70B	70.0	43.00
Gemma-2-27B	27.0	42.67
Llama-4-Scout	109.0	41.50
Bielik-4.5B-v3.0-Instruct	4.8	41.47
EuroLLM-9B	9.0	41.00
Qwen-2.5-72B	72.7	39.17
Mistral-Small-24B-2501	24.0	39.00
Llama-PLLuM-8B-chat	8.0	38.50
Qwen3-32B	32.8	37.67
Mixtral-8x7B	46.7	35.33
GPT-OSS-20b	20	32.3
Qwen-2.5-32B	32.8	30.50
Gemma-2-9B	9.0	29.17
Phi-4	14.7	29.17
Bielik-1.5B-v3.0-Instruct	1.6	27.50
Qwen-2.5-14B	14.8	26.67
Mistral-Nemo	12.2	23.00
Mistral-7B-v0.3	7.2	21.83

[†]Models with a non-commercial license.

Table 9: Polish Linguistic and Cultural Competency Benchmark (PLCC) results for open-source models. Closed proprietary models have been excluded from this comparison.

Model	AVG	arc_challenge	hellaswag	truthfulqa_mc2	mmlu	winogrande	gsm8k
Bielik-11B-v3	**68.45**	**61.43**	**81.38**	**47.65**	**67.55**	**78.53**	**74.15**
Qwen1.5-14B	66.70	56.57	81.08	52.06	69.36	73.48	67.63
Bielik-11B-v2	65.87	60.58	79.84	46.13	63.06	77.82	67.78
Qwen-14B	65.86	58.28	83.99	49.43	67.70	76.80	58.98
Meta-Llama-3-8B	62.62	60.24	82.23	42.93	66.70	78.45	45.19
Bielik-4.5B-v3	61.02	51.19	73.01	45.63	61.32	71.35	63.61
Mistral-7B-v0.1	60.97	59.98	83.31	42.15	64.16	78.37	37.83
Mistral-7B-v0.2	60.37	60.84	83.08	41.76	63.62	78.22	34.72
Bielik-7B-v0.1	49.98	45.22	67.92	47.16	43.20	66.85	29.49

Table 10: Open LLM Leaderboard results for base models

Model	AVG	arc_challenge	hellaswag	truthfulqa_mc2	mmlu	winogrande	gsm8k
SOLAR-10.7B-Instruct-v1.0	74.20	71.08	88.16	71.43	66.21	83.58	64.75
Phi-3-medium-4k-instruct	73.45	67.32	85.76	57.71	77.83	72.69	79.38
Bielik-11B-v3.0-Instruct	**72.45**	**64.59**	**81.96**	**54.25**	**71.11**	**77.19**	**85.60**
Bielik-11B-v2.5-Instruct	71.42	61.95	80.71	53.17	67.44	79.72	85.52
Bielik-11B-v2.6-Instruct	71.10	62.54	80.56	53.43	67.53	78.77	83.78
Bielik-11B-v2.2-Instruct	69.86	59.90	80.16	58.34	64.34	75.30	81.12
Bielik-11B-v2.3-Instruct	69.82	59.30	80.11	57.42	64.57	76.24	81.27
Bielik-11B-v2.1-Instruct	69.82	59.56	80.20	59.35	64.18	75.06	80.59
openchat-3.5-0106-gemma	69.42	64.68	81.08	54.93	64.69	78.30	72.86
Bielik-11B-v2.0-Instruct	68.04	58.62	78.65	54.65	63.71	76.32	76.27
Meta-Llama-3-8B-Instruct	66.87	60.75	78.55	51.65	67.07	74.51	68.69
Mistral-7B-Instruct-v0.2	65.71	63.14	84.88	68.26	60.78	77.19	40.03
Bielik-4.5B-v3-Instruct	64.89	56.06	73.90	50.79	63.66	71.19	73.69
gemma-7b	64.29	61.09	82.47	44.91	66.03	78.45	52.77
Qwen1.5-32B-Chat	62.95	66.04	85.49	66.95	74.99	77.19	7.05
Qwen1.5-14B-Chat	62.27	58.70	82.27	60.36	68.57	73.09	30.63
Qwen1.5-7B-Chat	55.15	55.89	78.56	53.54	61.65	67.72	13.57
Mistral-7B-Instruct-v0.1	54.96	54.52	75.63	56.28	55.38	73.72	14.25
Bielik-7B-Instruct-v0.1	51.26	47.53	68.91	46.18	49.47	65.51	29.95

Table 11: Open LLM Leaderboard results for selected instruction-tuned models

On English language tasks (Tables 10 and 11), Bielik models demonstrate solid cross-lingual capabilities. The base Bielik-11B-v2 reaches 65.87 average score. For instruction-tuned models, Bielik-11B-v3.0-Instruct scores 72.45 average, with particularly strong performance on GSM8K (85.60) and ARC challenge (64.59), indicating robust mathematical and reasoning capabilities.

5.7. INCLUDE-base-44

INCLUDE [34] is a comprehensive knowledge- and reasoning-centric benchmark designed to evaluate multilingual language models across 44 languages in realistic deployment scenarios. The benchmark comprises 22,637 four-option multiple-choice questions extracted from academic and professional examinations, covering 57 topics across diverse domains including STEM (Biology, Chemistry, Physics, Mathematics, Computer Science), Arts & Humanities (History, Philosophy, Literature, Visual Arts, Law), Social Sciences (Sociology, Economics, Psychology), Health-oriented Education (Medicine), and professional certifications (driving licenses, medical licenses, professional certifications).

A distinguishing feature of INCLUDE is its emphasis on regional knowledge and cultural context. Questions are categorized as either "agnostic" (universally applicable) or "region implicit/explicit" (requiring cultural or geographical knowledge specific to particular regions). This design enables assessment of models' ability to handle not only universal knowledge but also culturally-specific content essential for deployment in diverse linguistic communities. For our evaluation, we focus on a subset of 20 European languages from the full benchmark to assess Bielik's performance across its target linguistic region. The benchmark evaluation is presented in Table 12.

On INCLUDE-base-44 (Table 12), Bielik-11B-v3-Instruct achieves the highest scores among all evaluated models, with 64.8 overall average across European languages and 69.0 on Polish-specific tasks. This demonstrates superior balanced multilingual performance, surpassing Qwen2.5-14B-Instruct (61.7 average, 58.9 Polish) despite having fewer parameters. Notably, Bielik's Polish-specific score (69.0) substantially exceeds its multilingual average (64.8), reflecting the model's particular strength in its primary target language while maintaining robust cross-lingual capabilities. The base model Bielik-11B-v3 also shows strong performance (60.6 average, 63.9 Polish), outperforming several instruction-tuned models including Llama-3.1-8B-Instruct and EuroLLM-9B-Instruct. Compared to the previous version, Bielik-11B-v2 achieved 44.8 average with 53.5 on Polish tasks, showing significant improvement in v3.

5.8. Belebele Reading Comprehension

Belebele [2] is a massively multilingual reading comprehension benchmark spanning 122 language variants. The benchmark consists of multiple-choice reading comprehension questions derived from the FLORES-200 dataset, where models must demonstrate understanding of short passages by correctly answering questions about their content. For our evaluation, we assess performance across 28 European language variants to evaluate Bielik's reading comprehension capabilities across its target linguistic region. Results are presented in Table 13.

Model	Params (B)	AVG	Polish
Bielik-11B-v3-Instruct	**11**	**64.8**	**69.0**
Qwen2.5-14B-Instruct	14	61.7	58.9
Bielik-11B-v3	**11**	**60.6**	**63.9**
phi-4	15	58.8	49.6
Apertus-8B-Instruct-2509	8	57.9	49.6
Llama-3.1-8B-Instruct	8	55.3	53.8
EuroLLM-9B-Instruct	9	55.1	52.0
Qwen2.5-7B-Instruct	7	54.4	52.2
Mistral-Nemo-Instruct-2407	12	53.2	48.4
<u>Bielik-11B-v2.6-Instruct</u>	<u>11</u>	<u>51.5</u>	<u>59.3</u>
Mistral-Nemo-Base-2407	12	51.2	44.9
EuroLLM-9B	9	49.2	45.6
aya-expanse-8b	8	45.3	46.4
Mistral-7B-Instruct-v0.2	7	45.3	44.7
<u>Bielik-11B-v2</u>	<u>11</u>	<u>44.8</u>	<u>53.5</u>
pllum-12b-nc-chat-250715	12	44.2	60.6
Mistral-7B-v0.2	7	41.8	37.2
pllum-12b-nc-base-250715	12	37.8	52.7
<u>Bielik-4.5B-v3</u>	<u>4.5</u>	<u>35.9</u>	<u>48.7</u>
PLLuM-12B-base-250801	12	35.5	44.5
Llama-PLLuM-8B-base-250801	8	30.0	37.2

Table 12: INCLUDE-base-44 benchmark results showing average performance across European languages (20 language subset) and Polish-specific scores.

On Belebele (Table 13), Bielik-11B-v3.0-Instruct achieves 82.98 average across European languages, representing a substantial improvement over the previous version Bielik-11B-v2.6-Instruct (68.67). This score places Bielik second overall, closely following Qwen2.5-14B-Instruct (85.91) and ahead of the larger phi-4 15B model (81.71). The strong performance on this benchmark confirms Bielik's robust reading comprehension abilities across multiple European languages.

5.9. FLORES Machine Translation

FLORES (FLORES-200) is a widely-used machine translation benchmark covering 200 languages, designed to evaluate translation quality across diverse linguistic families. The benchmark measures translation performance using BLEU scores, which assess n-gram overlap between model-generated translations and human reference translations. For Bielik evaluation, we assess translation performance across 20 European language pairs, focusing on bidirectional translations between Polish and other European languages, as well as translations among European languages. This evaluation provides insights into the model's multilingual translation capabilities across its target linguistic region. Results are shown in Table 14.

Model	Params (B)	Score
Qwen2.5-14B-Instruct	14	85.91
Bielik-11B-v3.0-Instruct	**11**	**82.98**
phi-4	15	81.71
Qwen2.5-7B	7	74.60
Mistral-Nemo-Instruct-2407	12	74.14
cjvt/GaMS-9B-Instruct	9	72.40
Apertus-8B-Instruct-2509	8	69.58
EuroLLM-9B-Instruct	9	69.05
Bielik-11B-v2.6-Instruct	11	68.67
Apertus-8B-2509	8	59.04

Table 13: Belebele benchmark results showing model performance on reading comprehension across European languages (28 language subset).

Model	Params (B)	AVG	to Polish	from Polish
EuroLLM-9B-Instruct[*]	9	20.61	19.28	21.95
Bielik-11B-v3.0-Instruct	**11**	**19.22**	**18.54**	**19.91**
Bielik-11B-v3	**11**	**17.85**	**17.60**	**18.11**
phi-4 (15B)	15	15.58	14.55	16.61
Mistral-Nemo-Instruct-2407	12	14.35	13.37	15.33
Bielik-11B-v2.6-Instruct	11	13.58	15.77	11.38
Qwen2.5-14B-Instruct	14	13.24	12.55	13.93
Bielik-11B-v2	11	11.25	14.86	7.64
Qwen2.5-7B-Instruct	7	11.34	10.43	12.26

[*] EuroLLM was trained on FLORES dataset

Table 14: FLORES machine translation benchmark results showing translation performance across European languages (20 language pairs) measured by BLEU scores.

On FLORES translation tasks (Table 14), Bielik-11B-v3.0-Instruct achieves an average BLEU score of 19.22 across European language pairs, ranking second only to EuroLLM-9B-Instruct (20.61), which was trained on FLORES data. Notably, Bielik demonstrates balanced bidirectional translation capabilities with 18.54 BLEU for translation to Polish and 19.91 for translation from Polish. This represents a significant improvement over Bielik-11B-v2.6-Instruct (13.58 average), particularly in the from-Polish direction where v3.0 achieves 19.91 versus v2.6's 11.38. The base model Bielik-11B-v3 also shows strong translation performance (17.85 average), substantially outperforming larger models like phi-4 15B (15.58) and Qwen2.5-14B-Instruct (13.24).

5.10. Summary of Evaluation Results

The comprehensive evaluation demonstrates that Bielik-11B-v3.0-Instruct achieves state-of-the-art performance among open-source models of comparable size, with particularly

exceptional results on Polish-specific benchmarks. The model excels in cultural knowledge (PLCC: 71.83%, ranking first among all open-source models), general Polish language understanding (Open PL LLM Leaderboard: 65.93), and reading comprehension (Belebele: 82.98). Strong performance is also observed in specialized domains including medical knowledge (50.21%), complex text understanding (CPTUB: 3.73), and emotional intelligence (EQ-Bench: 71.20).

Importantly, Bielik maintains competitive English language capabilities (Open LLM Leaderboard: 72.45) and demonstrates robust machine translation performance (FLORES: 19.22 BLEU), confirming its effectiveness as a multilingual European model. Across benchmarks, Bielik-11B-v3.0-Instruct frequently outperforms models with 2-6× more parameters, demonstrating exceptional parameter efficiency.

6. Limitations and Biases

The Bielik v3 series of models may produce factually incorrect output and should not be relied upon to generate completely accurate information in all contexts. These models were trained on diverse public datasets, and despite our extensive efforts to clean and filter the training data, they may occasionally generate content that is biased, offensive, or factually inaccurate. Users should apply appropriate caution and verification when deploying these models, particularly in sensitive or high-stakes applications.

7. Conclusion

We have presented Bielik 11B v3, a state-of-the-art multilingual language model optimized for Polish while maintaining strong capabilities across European languages. This work represents a significant advancement in developing high-performance language models for underrepresented languages through efficient scaling and comprehensive training methodologies.

Our main contributions include: (1) a depth up-scaling approach that achieves 11 billion parameters while remaining deployable on consumer-grade hardware; (2) a substantially expanded multilingual pre-training corpus spanning 32 languages with over 1.1 trillion tokens, representing a fivefold increase over our previous version; (3) a sophisticated four-stage training pipeline incorporating continuous pre-training, supervised fine-tuning, Direct Preference Optimization with the DPO-Positive variant, and reinforcement learning through Group Relative Policy Optimization; and (4) comprehensive evaluation across nine diverse benchmarks demonstrating exceptional performance on Polish-specific tasks while maintaining competitive multilingual capabilities.

The evaluation results demonstrate that Bielik 11B v3 achieves exceptional parameter efficiency, frequently outperforming models with 2-6 times more parameters. On Polish-specific benchmarks, the model establishes new standards: it ranks first among open-source models on the Polish Linguistic and Cultural Competency Benchmark (71.83%), achieves 65.93 on the Open PL LLM Leaderboard, and scores 82.98 on Belebele reading comprehension across European languages. Notably, the model demonstrates balanced multilingual capabilities, scoring 64.8 on INCLUDE-base-44 across European languages while excelling on English

benchmarks (72.45 average on Open LLM Leaderboard, with particularly strong mathematical reasoning performance of 85.60 on GSM8K).

Compared to the previous Bielik 11B v2 series, version 3 shows substantial improvements across most benchmarks, particularly in multilingual translation capabilities (FLORES: 19.22 vs. 11.25 BLEU) and reading comprehension (Belebele: 82.98 vs. 68.67). The model maintains competitive performance on specialized tasks including medical knowledge assessment (50.21%) and complex text understanding (CPTUB: 3.73), demonstrating its versatility across diverse domains.

These results confirm that high-quality, resource-efficient language models can be developed for less-represented languages without sacrificing performance. The model's design prioritizes accessibility, with quantization support enabling deployment across diverse hardware configurations, from enterprise servers to consumer GPUs. This democratization of access is crucial for advancing AI capabilities beyond dominant languages.

Future research directions include exploring larger model variants to further push performance boundaries, extending language coverage to additional European and global languages, developing domain-specific adaptations for legal, medical, and scientific applications, and investigating multimodal extensions incorporating vision and speech capabilities. Additionally, we aim to conduct more extensive ablation studies to better understand the individual contributions of each training stage and to optimize the training pipeline further.

Acknowledgements

References

[1] Ainslie J., Lee-Thorp J., de Jong M., Zemlyanskiy Y., Lebron F., Sanghai S.: GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. In: H. Bouamor, J. Pino, K. Bali, eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901. Association for Computational Linguistics, Singapore, 2023. URL http://dx.doi.org/10.18653/v1/2023.emnlp-main.298.

[2] Bandarkar L., Liang D., Muller B., Artetxe M., Shukla S.N., Husa D., Goyal N., Krishnan A., Zettlemoyer L., Khabsa M.: The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775. Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 2024. URL `https://aclanthology.org/2024.acl-long.44`.

[3] Beeching E., Fourrier C., Habib N., Han S., Lambert N., Rajani N., Sanseviero O., Tunstall L., Wolf T.: Open LLM Leaderboard. `https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard`, 2023.

[4] Beeching E., Fourrier C., Habib N., Han S., Lambert N., Rajani N., Sanseviero O., Tunstall L., Wolf T.: Open LLM Leaderboard (2023-2024). `https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard`, 2023.

[5] Burchell L., de Gibert O., Arefyev N., Aulamo M., Bañón M., Chen P., Fedorova M., Guillou L., Haddow B., Hajič J., Helcl J., Henriksson E., Klimaszewski M., Komulainen V., Kutuzov A., Kytöniemi J., Laippala V., Mæhlum P., Malik B., Mehryary F., Mikhailov V., Moghe N., Myntti A., O'Brien D., Oepen S., Pal P., Piha J., Pyysalo S., Ramírez-Sánchez G., Samuel D., Stepachev P., Tiedemann J., Variš D., Vojtěchová T., Zaragoza-Bernabeu J.: An Expanded Massive Multilingual Dataset for High-Performance Language Technologies (HPLT), 2025. URL `https://arxiv.org/abs/2503.10267`.

[6] Dadas S.: Training Effective Neural Sentence Encoders from Automatically Mined Paraphrases. In: *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 371–378. 2022. URL `http://dx.doi.org/10.1109/SMC53654.2022.9945218`.

[7] Dadas S., Grębowiec M., Perełkiewicz M., Poświata R.: Evaluating Polish linguistic and cultural competency in large language models, 2025. URL `https://arxiv.org/abs/2503.00995`.

[8] Dadas S., Perełkiewicz M., Poświata R.: Evaluation of Sentence Representations in Polish. In: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, eds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1674–1680. European Language Resources Association, Marseille, France, 2020. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.207`.

[9] Dauphin Y.N., Fan A., Auli M., Grangier D.: Language Modeling with Gated Convolutional Networks. In: D. Precup, Y.W. Teh, eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, *Proceedings of Machine Learning Research*, vol. 70, pp. 933–941. PMLR, 2017. URL `http://proceedings.mlr.press/v70/dauphin17a.html`.

[10] Dong J., Feng B., Guessous D., Liang Y., He H.: Flex Attention: A Programming Model for Generating Optimized Attention Kernels, 2024. URL `https://arxiv.org/abs/2412.05496`.

[11] Gao L., Tow J., Abbasi B., Biderman S., Black S., DiPofi A., Foster C., Golding L., Hsu J., Le Noac'h A., Li H., McDonell K., Muennighoff N., Ociepa C., Phang J., Reynolds L., Schoelkopf H., Skowron A., Sutawika L., Tang E., Thite A., Wang B., Wang K., Zou A.: A framework for few-shot language model evaluation, 2024. URL `http://dx.doi.org/10.5281/zenodo.12608602`.

[12] Hernández-Cano A., Hägele A., Huang A.H., Romanou A., Solergibert A.J., Pasztor B., Messmer B., Garbaya D., Ďurech E.F., Hakimi I., Giraldo J.G., Ismayilzada M., Foroutan N., Moalla S., Chen T., Sabolčec V., Xu Y., Aerni M., AlKhamissi B., Marinas I.A., Amani M.H., Ansaripour M., Badanin I., Benoit H., Boros E., Browning N., Bösch F., Böther M., Canova N., Challier C., Charmillot C., Coles J., Deriu J., Devos A., Drescher L., Dzenhaliou D., Ehrmann M., Fan D., Fan S., Gao S., Gila M., Grandury M., Hashemi D., Hoyle A., Jiang J., Klein M., Kucharavy A., Kucherenko A., Lübeck F., Machacek R., Manitaras T., Marfurt A., Matoba K., Matrenok S., Mendonça H., Mohamed F.R., Montariol S., Mouchel L., Najem-Meyer S., Ni J., Oliva G., Pagliardini M., Palme E., Panferov A., Paoletti L., Passerini M., Pavlov I., Poiroux A., Ponkshe K., Ranchin N., Rando J., Sauser M., Saydaliev J., Sayfiddinov M.A., Schneider M., Schuppli S., Scialanga M., Semenov A., Shridhar K., Singhal R., Sotnikova A., Sternfeld A., Tarun A.K., Teiletche P., Vamvas J., Yao X., Ilic H.Z.A., Klimovic A., Krause A., Gulcehre C., Rosenthal D., Ash E., Tramèr F., VandeVondele J., Veraldi L., Rajman M., Schulthess T., Hoefler T., Bosselut A., Jaggi M., Schlag I.: Apertus: Democratizing Open and Compliant LLMs for Global Language Environments. `https://arxiv.org/abs/2509.14233`, 2025.

[13] Jiang A.Q., Sablayrolles A., Mensch A., Bamford C., Chaplot D.S., de las Casas D., Bressand F., Lengyel G., Lample G., Saulnier L., Lavaud L.R., Lachaux M.A., Stock P., Scao T.L., Lavril T., Wang T., Lacroix T., Sayed W.E.: Mistral 7B, 2023. URL `https://arxiv.org/abs/2310.06825`.

[14] Jiang Z., Gu J., Zhu H., Pan D.Z.: Pre-RMSNorm and Pre-CRMSNorm Transformers: Equivalent and Efficient Pre-LN Transformers. In: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine, eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. 2023. URL `http://papers.nips.cc/paper_files/paper/2023/hash/8f1bacee31caf990a4f08d84f0ccb322-Abstract-Conference.html`.

[15] Kim S., Kim D., Park C., Lee W., Song W., Kim Y., Kim H., Kim Y., Lee H., Kim J., Ahn C., Yang S., Lee S., Park H., Gim G., Cha M., Lee H., Kim S.: SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling. In: Y. Yang, A. Davani, A. Sil, A. Kumar, eds., *Proceedings of the 2024 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track), pp. 23–35. Association for Computational Linguistics, Mexico City, Mexico, 2024. URL `https://aclanthology.org/2024.naacl-industry.3`.

[16] Kocoń J., Miłkowski P., Zaśko-Zielińska M.: Multi-Level Sentiment Analysis of PolEmo 2.0: Extended Corpus of Multi-Domain Consumer Reviews. In: M. Bansal, A. Villavicencio, eds., *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 980–991. Association for Computational Linguistics, Hong Kong, China, 2019. URL `http://dx.doi.org/10.18653/v1/K19-1092`.

[17] Kocoń J., Piasecki M., Janz A., Ferdinan T., Radliński Ł., Koptyra B., Oleksy M., Woźniak S., Walkowiak P., Wojtasik K., Moska J., Naskręt T., Walkowiak B., Gniewkowski M., Szyc K., Motyka D., Banach D., Dalasiński J., Rudnicka E., Alberski B., Walkowiak T., Szczęsny A., Markiewicz M., Bernaś T., Mazur H., Żyta K., Tykierko M., Chodak G., Kajdanowicz T., Kazienko P., Karlińska A., Seweryn K., Kołos A., Chrabąszcz M., Lorenc K., Krasnodębska A., Wilczek A., Dziewulska K., Betscher P., Cieślińska Z., Kowol K., Mikoś D., Trzciński M., Krutul D., Kozłowski M., Dadas S., Poświata R., Perełkiewicz M., Grębowiec M., Kazuła M., Białas M., Roszko R., Roszko D., Vaičenonėnienė J., Utka A., Levchuk P., Kowalski P., Prawdzic-Jankowska I., Ogrodniczuk M., Borys M., Bulińska A., Gumienna W., Kieraś W., Komosińska D., Krasnowska-Kieraś K., Kobyliński Ł., Lewandowska M., Łaziński M., Łątkowski M., Mastalerz D., Milewicz B., Mykowiecka A.A., Peljak-Łapińska A., Penno S., Przybysz Z., Rudolf M., Rybak P., Saputa K., Tomaszewska A., Wawer A., Woliński M., Wołoszyn J., Wróblewska A., Żuk B., Żarnecki F., Kaczyński K., Cichosz A., Deckert Z., Garnys M., Grabarczyk I., Janowski W., Karasińska S., Kujawiak A., Misztela P., Szymańska M., Walkusz K., Siek I., Kwiatkowski J., Pęzik P.: PLLuM: A Family of Polish Large Language Models, 2025. URL `https://arxiv.org/abs/2511.03823`.

[18] Liu Z., Chen C., Li W., Qi P., Pang T., Du C., Lee W.S., Lin M.: Understanding R1-Zero-Like Training: A Critical Perspective, 2025. URL `https://arxiv.org/abs/2503.20783`.

[19] Loshchilov I., Hutter F.: Decoupled Weight Decay Regularization. In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

[20] Lozhkov A., Ben Allal L., von Werra L., Wolf T.: FineWeb-Edu: the Finest Collection of Educational Content, 2024. URL `http://dx.doi.org/10.57967/hf/2497`.

[21] Marcinczuk M., Ptak M., Radziszewski A., Piasecki M.: Open dataset for development of Polish Question Answering systems. In: *Proceedings of the 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Wydawnictwo Poznanskie, Fundacja Uniwersytetu im. Adama Mickiewicza*. 2013.

[22] Martins P.H., Fernandes P., Alves J., Guerreiro N.M., Rei R., Alves D.M., Pombal J., Farajian A., Faysse M., Klimaszewski M., Colombo P., Haddow B., de Souza J.G., Birch A., Martins A.F.: EuroLLM: Multilingual Language Models for Europe. In: *Procedia Computer Science*, vol. 255, pp. 53–62, 2025. ISSN 1877-0509. URL `http://dx.doi.org/https://doi.org/10.1016/j.procs.2025.02.260`. Proceedings of the Second EuroHPC user day.

[23] Nguyen T., Nguyen C.V., Lai V.D., Man H., Ngo N.T., Dernoncourt F., Rossi R.A., Nguyen T.H.: CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages. In: N. Calzolari, M.Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue, eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4226–4237. ELRA and ICCL, Torino, Italia, 2024. URL `https://aclanthology.org/2024.lrec-main.377`.

[24] Ociepa K., Flis Ł., Kinas R., Gwoździej A., Wróbel K., SpeakLeash Team, Cyfronet Team: Bielik-11B-v2.6-Instruct model card, 2025. URL `https://huggingface.co/speakleash/Bielik-11B-v2.6-Instruct`. Accessed: 2025-12-30.

[25] Ociepa K., Flis Ł., Kinas R., Wróbel K., Gwoździej A.: Bielik v3 Small: Technical Report. URL `https://arxiv.org/abs/2505.02550`.

[26] Ociepa K., Flis Ł., Wróbel K., Gwoździej A., Kinas R.: Bielik 11B v2 Technical Report, 2025. URL `https://arxiv.org/abs/2505.02410`.

[27] Ociepa K., Flis Ł., Wróbel K., Gwoździej A., Kinas R.: BIELIK 7B V0.1: POLISH LANGUAGE MODEL - DEVELOPMENT, INSIGHTS, AND EVALUATION. In: *Computer Science*, vol. 26(4), 2025. URL `http://dx.doi.org/10.7494/csci.2025.26.4.7689`.

[28] Ogrodniczuk M., Kopeć M.: The Polish Summaries Corpus. In: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3712–3715. European Language Resources Association (ELRA), Reykjavik, Iceland, 2014. URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/1211_Paper.pdf`.

[29] Paech S.J.: EQ-Bench: An Emotional Intelligence Benchmark for Large Language Models, 2023. URL `https://arxiv.org/abs/2312.06281`.

[30] Penedo G., Kydlíček H., Sabolčec V., Messmer B., Foroutan N., Kargaran A.H., Raffel C., Jaggi M., Werra L.V., Wolf T.: FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language, 2025. URL `https://arxiv.org/abs/2506.20920`.

[31] Peng B., Quesnelle J., Fan H., Shippole E.: YaRN: Efficient Context Window Extension of Large Language Models. In: B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan,

Y. Sun, eds., *International Conference on Representation Learning*, vol. 2024, pp. 31932–31951. 2024. URL `https://proceedings.iclr.cc/paper_files/paper/2024/file/874a4d89f2d04b4bcf9a2c19545cf040-Paper-Conference.pdf`.

[32] Pokrywka J., Kaczmarek J., Gorzelańczyk E.: GPT-4 passes most of the 297 written Polish Board Certification Examinations, 2024. URL `https://arxiv.org/abs/2405.01589`.

[33] Ptaszynski M., Pieciukiewicz A., Dybala P., Skrzek P., Soliwoda K., Fortuna M., Leliwa G., Wroczynski M.: Expert-Annotated Dataset to Study Cyberbullying in Polish Language. In: *Data*, vol. 9(1), p. 1, 2023.

[34] Romanou A., Foroutan N., Sotnikova A., Chen Z., Nelaturu S.H., Singh S., Maheshwary R., Altomare M., Haggag M.A., Amayuelas A., et al.: INCLUDE: Evaluating Multilingual Language Understanding with Regional Knowledge. In: *arXiv preprint arXiv:2411.19799*, 2024.

[35] Rybak P., Mroczkowski R., Tracz J., Gawlik I.: KLEJ: Comprehensive Benchmark for Polish Language Understanding. In: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault, eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1191–1201. Association for Computational Linguistics, Online, 2020. URL `http://dx.doi.org/10.18653/v1/2020.acl-main.111`.

[36] Rybak P., Przybyła P., Ogrodniczuk M.: PolQA: Polish Question Answering Dataset. In: N. Calzolari, M.Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue, eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 12846–12855. ELRA and ICCL, Torino, Italia, 2024. URL `https://aclanthology.org/2024.lrec-main.1125`.

[37] Shao Z., Wang P., Zhu Q., Xu R., Song J., Zhang M., Li Y., Wu Y., Guo D.: DeepSeek-Math: Pushing the Limits of Mathematical Reasoning in Open Language Models. In: *arXiv preprint arXiv:2402.03300*, 2024.

[38] Shazeer N.: GLU Variants Improve Transformer, 2020. URL `https://arxiv.org/abs/2002.05202`.

[39] Sheng G., Zhang C., Ye Z., Wu X., Zhang W., Zhang R., Peng Y., Lin H., Wu C.: HybridFlow: A Flexible and Efficient RLHF Framework. In: *arXiv preprint arXiv:2409.19256*, 2024.

[40] Shi Z., Yang A.X., Wu B., Aitchison L., Yilmaz E., Lipani A.: Instruction Tuning With Loss Over Instructions. In: A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J.M. Tomczak, C. Zhang, eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*. 2024. URL `http://papers.nip`

s.cc/paper_files/paper/2024/hash/7ffb43adf37b3eeaba559098b
c084cc6-Abstract-Conference.html.

[41] Soboleva D., Al-Khateeb F., Myers R., Steeves J.R., Hestness J., Dey N.: SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. `https://www.cerebr as.net/blog/slimpajama-a-627b-token-cleaned-and-deduplica ted-version-of-redpajama`, 2023. URL `https://huggingface.co/d atasets/cerebras/SlimPajama-627B`.

[42] Sowa J., Krawczyk M., Nadolna N., Zielińska A., Filipkowska M., Kosiak A., Kania M., Wróbel K., Kinas R., Baczyński S., SpeakLeash Team, Cyfronet Team: Complex Polish Text Understanding Benchmark. `https://huggingface.co/spaces/speakl eash/cptu_bench`, 2024.

[43] Su J., Ahmed M., Lu Y., Pan S., Bo W., Liu Y.: RoFormer: Enhanced transformer with Rotary Position Embedding. In: *Neurocomputing*, vol. 568, p. 127063, 2024. ISSN 0925-2312. URL `http://dx.doi.org/https://doi.org/10.1016/j.ne ucom.2023.127063`.

[44] Tuora R., Zwierzchowska A., Zawadzka-Paluektau N., Klamra C., Kobyliński Ł.: PoQuAD-The Polish Question Answering Dataset-Description and Analysis. In: *Proceedings of the 12th Knowledge Capture Conference 2023*, pp. 105–113. 2023.

[45] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I.: Attention is All you Need. In: I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan, R. Garnett, eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008. 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547 dee91fbd053c1c4a845aa-Abstract.html`.

[46] Wang S., Wang G., Wang Y., Li J., Hovy E., Guo C.: Packing Analysis: Packing Is More Appropriate for Large Models or Datasets in Supervised Fine-tuning, 2024. URL `https://arxiv.org/abs/2410.08081`.

[47] Wróbel K., SpeakLeash Team, Cyfronet Team: Open PL LLM Leaderboard. `https: //huggingface.co/spaces/speakleash/open_pl_llm_leaderboard`, 2024.

Affiliations

Krzysztof Ociepa
    SpeakLeash, Azurro, krzysztof.ociepa@bielik.ai
Łukasz Flis
    SpeakLeash, ACK Cyfronet AGH, lukasz.flis@cyfronet.pl
Remigiusz Kinas
    SpeakLeash, remigiusz.kinas@bielik.ai

Krzysztof Wróbel

 SpeakLeash, Jagiellonian University, Enelpol, krzysztof.pawel.wrobel@uj.edu.pl

Adrian Gwoździej

 SpeakLeash, ACK Cyfronet AGH, adrian.gwozdziej@cyfronet.pl