Public Summary of Training Content for General-Purpose AI models

Version of the Summary: *1.0*
Last update: *None*

1. General information

1.1. Provider identification

Provider name and contact details: *FUNDACJA SPEAKLEASH*
KRS 0001099568
NIP 8992990977
REGON 528375240
Niemczańska 33 lok 4, 50-561 Wrocław, Poland

Authorised representative name and contact details: *N/A*

1.2. Model identification

Versioned model name(s): *Bielik v3 11B Instruct*

Model dependencies: *Mistral 7B v0.2*

Date of placement of the model on the Union market: *31.12.2025*

1.3 Modalities, overall training data size and other characteristics

Modality	Training data size	Types of content
X Text	☐ Less than 1 billion tokens X 1billion to 10 trillions tokens ☐ More than 10 trillions tokens	The training data includes a wide variety of content types, such as: legal and official documents (e.g., court rulings, legal acts, regulations), scientific texts (from the Science Library), press publications (from commercially licensed sources), web content from public domains and thematic forums, parliamentary discourse, and multilingual resources from public repositories like Wikipedia and Europeana. It also contains

		synthetic data generated for training purposes.

Latest date of data acquisition/collection for model training:	*May 2025*
Description of the linguistic characteristics of the overall training data:	*The linguistic composition of the training data is predominantly Polish and English, which together constitute approximately 90% of the corpus. The remaining 10% consists of an admixture of other languages, with a focus on official languages of the European Union.*
Other relevant characteristics of the overall training data:	*The dataset has a strong focus on content from Polish and European sources, particularly within the legal, administrative, and public discourse domains. Furthermore, specific efforts were made to include data reflecting regional specificities within Poland, such as the inclusion of content from the Silesian and Kashubian language versions of Wikipedia.*

2. List of data sources

2.1. Publicly available datasets

Have you used publicly available datasets to train the model? X Yes ☐ No

If yes, specify the modality(ies) of the content covered by the datasets concerned:

X Text ☐ Image ☐ Video ☐ Audio

☐ Other *If so, please specify...*

List of large publicly available datasets:

The publicly available datasets used for training were collected and processed over a defined time range, covering data acquired between 2022 and November 2025. Where applicable, fixed temporal snapshots of large-scale datasets were used to ensure consistency and traceability of the training data.

The categorisation of datasets, including their thematic and linguistic characteristics, as well as the detailed preprocessing steps, are described in the corresponding technical reports (see: https://arxiv.org/abs/2505.02410). These reports document the data curation pipeline applied prior to training.

As part of the preprocessing pipeline, deduplication and quality assessment of textual content were performed to reduce redundancy and low-quality data. The methods used for deduplication and quality

	filtering are described in detail in the technical reports (see: https://arxiv.org/abs/2505.02410). *In addition, the data curation process includes systematic checks of robots.txt files, explicit Text and Data Mining (TDM) reservations expressed via meta tags or equivalent mechanisms, as well as reviews of terms of service or usage policies available on website footers or main domain pages. Data from sources expressing such reservations are excluded from the training datasets.* *Anonymisation is applied as a mandatory step in the preprocessing pipeline to remove personally identifiable information prior to training.* *The data exclusion mechanisms described above are applied in a retroactive manner. If a source subsequently expresses a reservation of rights, the corresponding data is removed from the training datasets used for all subsequent model versions.* *https://hplt-project.org/datasets/v2.0* *https://huggingface.co/datasets/uonlp/CulturaX* *https://huggingface.co/datasets/HuggingFaceFW/fineweb-2* *https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu* *http://commoncrawl.org/* *https://huggingface.co/datasets/cerebras/SlimPajama-627B* *A mix of web domains was used, which was then subjected to a mixing and deduplication process.*
General description of other publicly available datasets not listed above:	*This category includes a variety of smaller, publicly available text datasets. The content comprises translation datasets, resources from publicly available repositories such as Wikipedia, the Science Library (Biblioteka Nauki, accessed via API), the Parliamentary Discourse Corpus (Korpus Dyskursu Parlamentarnego), and Europeana collections. Additionally, it includes a collection of anonymized, publicly available legal and administrative documents such as court rulings, legal acts, resolutions, court gazettes, and journals of laws.* *In terms of the nature of the content, the datasets include copyright-protected materials (e.g. Wikipedia, legal acts) and data that may contain personal information. An anonymization mechanism is applied as part of the standard preprocessing pipeline to remove sensitive data such as phone numbers, email addresses, URLs, and PESEL numbers. Deduplication and quality filtering are also applied to reduce redundancy and low-quality content. The scope and strictness of these preprocessing measures are applied in a proportionate manner, taking into account the characteristics of the data sources, including their licensing terms and mode of access (e.g. open licenses, access via official APIs, or public data dumps). This proportional approach does not affect the application of anonymization measures with respect to personal data.*

Linguistically, the datasets are primarily in Polish and English, with an admixture of other European languages. The data collection for these datasets began in November 2022 and is an ongoing process.

2.2 Private non-publicly available datasets obtained from third parties

2.2.1. Datasets commercially licensed by rightsholders or their representatives

Have you concluded transactional commercial licensing agreement(s) with rightsholder(s) or with their representatives?	X Yes ☐ No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	X Text ☐ Image ☐ Video ☐ Audio ☐ Other *If so, please specify...* *Agreements were concluded with the itwiz.pl editorial office. The rightsholder, represented by the owner and editor-in-chief of ITWiz, provided a licensed corpus of articles published over a period of approximately ten years. The content was delivered directly by the rightsholder in a structured document format (DOCX) for the purpose of model training.*

2.2.2. Private datasets obtained from other third parties

Have you obtained private datasets from third parties that are not licensed as described in Section 2.2.1, such as data obtained from providers of private databases, or data intermediaries?	☐ Yes X No
If yes, specify the modality(ies) of the content covered by the datasets concerned:	☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other *If so, please specify...*

2.3 Data crawled and scraped from online sources

Were crawlers used by the provider or on behalf of?	X Yes ☐ No
If yes, specify crawler name(s)/identifier(s):	*Speakleash*
Purposes of the crawler(s):	*The purpose of the crawlers was to collect text data for model training. This included gathering content from publicly available PDF files and texts from online forums, sourced from open sources.*
General description of crawler behaviour:	*The crawlers were designed to respect the robots.txt protocol. In the case of online forums, content was collected only if there was no explicit prohibition against such activity.*
Period of data collection:	*From 11/2022 to 05/2025*
Comprehensive description of the type of content and online sources crawled:	*The crawled online sources primarily included two categories: public sector websites and thematic online forums. The public sector websites (e.g., government and municipal portals) were targeted for their content rich in legal and administrative documents, such as resolutions, regulations, and ordinances. Thematic online forums covered a range of topics, for example, automotive, electronics, gardening etc.. In terms of linguistic and geographical characteristics, the content was primarily Polish.*
Type of modality covered:	X Text ☐ Image ☐ Video ☐ Audio ☐ Other *If so, please specify...*
Summary of the most relevant domain names crawled:	*The most relevant internet domain names contributing to the crawled data are characterised by their functional categories rather than disclosed individually. The majority of the crawled content originates from public sector domains, including domains operated under governmental and public information bulletin top-level domains (such as .gov and .bip). These sources primarily provided open public documents and official attachments, including resolutions, regulations, and other administrative or legal materials.* *The remaining portion of the crawled data originates from thematic online forums covering a wide range of non-sensitive subject areas (e.g. automotive, electronics, gardening). These forums were selected as dedicated crawl targets due to their publicly accessible nature and relevance for general-language modeling, and were processed in accordance with the same data governance and opt-out procedures as other crawled sources.*
Additional comments (optional):	*Data collected through the provider's own crawlers is subject to the same preprocessing and governance pipeline as the large publicly available datasets used for training.*

Prior to any model training, crawled data undergoes deduplication, quality classification and filtering, thematic categorisation, anonymization of personal data, and systematic checks for text and data mining (TDM) reservations, including the analysis of robots.txt files, relevant meta tags, and applicable terms of service or usage policies.

These preprocessing steps are applied before the data is included in any training dataset. In addition, the exclusion mechanisms related to TDM reservations and other opt-out signals are applied retroactively. Where a source subsequently expresses a reservation of rights, the corresponding data is removed from the training datasets used for all subsequent model versions.

2.4 User data

Was data from user interactions with the AI model (e.g. user input and prompts) used to train the model?	☐ Yes X No
Was data collected from user interactions with the provider's other services or products used to train the model?	☐ Yes X No
If yes, provide a general description of the provider's services or products that were used to collect the user data:	*(Not applicable)*
Type of modality covered:	☐ Text ☐ Image ☐ Video ☐ Audio ☐ Other *If so, please specify...*

2.5 Synthetic data

Was synthetic AI-generated data created by the provider or on their behalf to train the model?	X Yes ☐ No
If yes, modality of the synthetic data:	X Text ☐ Image ☐ Video ☐ Audio ☐ Other *If so, please specify...*
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	*Name of the models: Deepseek v3, Bielik v2.3 11B*
Information about other AI models, including provider's own AI model(s) not available on the	

market, used to generate synthetic data to train the model to which this Summary applies:	*Not applicable.*

2.6 Other sources of data

Have data sources other than those described in Sections 2.1 to 2.5 been used to train the model?	☐ Yes X No

3. Data processing aspects

3.1. Respect of reservation of rights from text and data mining exception or limitation

Are you a Signatory to the Code of Practice for general-purpose AI models that includes commitments to respect reservations of rights from the TDM exception or limitation?	☐ Yes X No
Describe the measures implemented before model training to respect reservations of rights from the TDM exception or limitation before and during data collection, including the opt-out protocols and solutions honoured by the provider or, as applicable, by third parties from which datasets have been obtained:	*Although we are not formal signatories to a Code of Practice, we are publicly committed to respecting the rights of rightsholders. The following measures have been implemented before and during data collection to comply with the reservation of rights from the TDM exception:* *Automated Protocol Compliance: Our data collection tools, including crawlers, are designed to honour standard opt-out protocols. This primarily involves strict adherence to the robots.txt file, which prevents access to parts of a website that a publisher does not wish to be crawled.* *TDM Reservation Checks: Before processing data from a source, we conduct checks for explicit Text and Data Mining (TDM) reservations, including machine-readable signals and equivalent declarations. Data from sources that have expressed such a reservation are excluded from the training datasets.*

Terms of Service and Community Guidelines Review: For sources such as online forums and community websites, we review applicable terms of service or community guidelines. Content is collected only where no explicit prohibition against data collection, scraping, or use for AI training is identified.

In addition, these exclusion mechanisms are applied retroactively. Where a source subsequently expresses a reservation of rights, the corresponding data is removed from the training datasets used for all subsequent model versions.

Decisions related to the inclusion or exclusion of data sources, including opt-out signals and TDM reservations, are logged and documented as part of the data governance process, enabling traceability and review of data sourcing decisions over time.

3.2 Removal of illegal content

General description of measures taken:

Our approach to preventing and removing illegal and harmful content from the training data is multi-layered and implemented as a mandatory, end-to-end preprocessing pipeline applied prior to model training. The process combines the use of pre-filtered external datasets with proprietary automated filters and safety models developed and maintained by the provider.

Reliance on Pre-filtered Public Datasets: As an initial step, we leverage the data cleaning and filtering measures already implemented by the providers of large, publicly available datasets used for training, such as HPLT, FineWeb-2, and CulturaX. These datasets undergo additional internal processing before inclusion in the training data.

Deduplication and Quality Assessment: To reduce redundancy and limit the risk of memorization and exposure to low-quality or problematic content, the preprocessing pipeline includes dataset-level and document-level deduplication. In parallel, a proprietary quality classification model is applied to assess and filter out low-quality text, a category into which harmful content (e.g. pornographic or spam-like material) frequently falls. This classifier includes detection and scoring of vulgar and sexually explicit language as part of the automated filtering process.

Dedicated Guardrail Model: To further strengthen safety controls, we deploy a dedicated, small-scale guardrail model trained specifically to identify and categorise harmful content into predefined risk categories, including self-harm, hate, sexual content, crime-related content, and vulgar language. Outputs flagged by this model are excluded from the training datasets.

Anonymization of Personal Data: In parallel with content-based filtering, anonymization is applied as a standard step of the preprocessing pipeline to remove personally identifiable information

(PII), including contact details and unique personal identifiers, using automated detection mechanisms.

Continuous Improvement and Oversight: The data filtering and safety pipeline is continuously reviewed and improved by a multidisciplinary team. Additional auxiliary classifiers are under development to further enhance detection capabilities and to expand coverage to additional categories of potentially harmful content, such as gambling-related texts. Despite these measures, a residual risk of inappropriate content remaining in the data is acknowledged.

3.3. Other information (optional)

Other relevant information about data processing (optional):	*We acknowledge that despite the comprehensive measures implemented, no filtering system is infallible, and a residual risk of some inappropriate content remaining in the dataset always exists. We are committed to continuously improving our data processing and safety pipelines to identify and address such content as effectively as possible*