Infrastructure for Fine-tuning Pre-trained Large Language Models (IfGPT)

International Forum on Advanced ICT Research and Innovation (30.09-01.10.2025)

Svetla Koeva
Institute for Bulgarian Language, Bulgarian Academy of Sciences

01.10.2025

Main goal

Selection and **pre-processing** of **big data for Bulgarian**, including **company-** or **industry-specific data**, and **fine-tuning** of suitable freely available **large language models** to solve **specific tasks**.

Specific tasks

A detailed description of the **distinguishing traits** of **large language models** and the specification of criteria for their **assessment, comparison, and selection**, which will facilitate the choice of free-to-use large language models that meet predefined criteria for purpose and mode of operation.

Specific tasks

Collecting, filtering, anonymising, and deduplicating large, diverse, high-quality **datasets for Bulgarian**, and defining a graph-based model for their **metadata** to enable the extraction of thematically oriented or specialised datasets.

Specific tasks

Fine-tuning pre-trained **large language models for Bulgarian** facilitated by a protocol that outlines **effective modern techniques,** depending on the purpose and domain.

Specific tasks

Evaluation of fine-tuned large language models for Bulgarian using a protocol that includes both automatic and human assessments.

Expected results

IfGPT will enhance freely available large language models and chat models to better reflect the Bulgarian language and culture, taking into account context and information reliability.

Expected results

IfGPT will improve technologies for extracting clean data without duplication or violation of content integrity. Methods for selecting non-toxic data and for data anonymisation will be expanded.

Expected results

IfGPT will create infrastructure for fine-tuning large language models in Bulgarian, as well as for specific domains or tasks. The infrastructure will include tools for developing high-quality datasets and for evaluating the fine-tuned large language models.

Selection of LLMs

When selecting suitable large language models for fine-tuning, our methodology relies on a comparative analysis of the technical descriptions of existing models.

Currently, there are many open large language and multilingual models, released under various access types, each offering different levels of openness and usability.

Selection of LLMs

- API-access models: no control over or modification of the underlying weights (OpenAI's GPT-4 model family, Anthropic's Claude model family).

- Open-weights (restricted-use) models: the weights are publicly released, but under conditions that limit their use (e.g., no commercial use, no redistribution) (DeepSeek V3, Meta's LLaMA 2 model family).

Selection of LLMs

- Open weights (non-commercial) models: weights are shared openly, but only for research or educational use. (Mistral Large 2, BigScience BLOOM).

- Open weights (unrestricted) models: weights are open, with licences allowing free use, modification, and redistribution. (AlphaGeometry, GPT-NeoX, Falcon 180B).

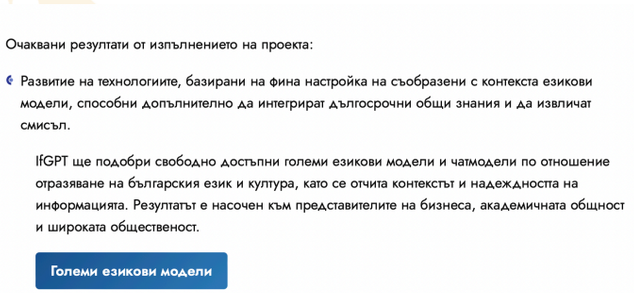
Selection of LLMs

- **Purpose**: the intended use of the models
- **Access**: whether the models are free or require payment
- **Size:** number of parameters
- **Support** for fine-tuning
- **Training data:** size, quality, and languages

Selection of LLMs

- **Speed**: the time required for the model to process a request and return a response
- **Performance**: assessment of the model's accuracy, semantic and grammatical correctness, and relevance of results
- **Risk assessment**: tendency towards hallucinations and data biases
- **Supported languages:** including Bulgarian

Selection of LLMs



- A **selection protocol,** based on criteria for assessing and comparing large language models, has been developed.
- An **online search** within a database containing large language model descriptions (approximately **150 models** so far) is available on the project website.
- Using this search, the large language models most suitable for fine-tuning experiments in Bulgarian and for a specific task within a particular thematic area can be easily selected.

Selection of datasets

The aim is to gather as much **diverse, high-quality Bulgarian language data** as possible, created by people, that **does not contain sensitive, incorrect, or ethically unacceptable information, avoids repetition**, and is accompanied by accurate information about its source and usage licence.

Selection of datasets

- **Merging several relatively large Bulgarian text collections** into a single dataset with standardised metadata descriptions and document formats.
- **Adding new texts** to the dataset in a standardised manner.
- Deploying and customising a **set of tools in a pipeline** for text cleaning, deduplication, and detection of sensitive and biased information to ensure data quality.

Selection of datasets

- Providing a **uniform metadata description** for all documents in the datasets and organising the metadata categories in a graph representation, as originally proposed for the Bulgarian National Corpus.

- Providing **means to efficiently query metadata** to find suitable text documents for a given domain or task.

IfGPT dataset

The components of the IfGPT are:

1) collections of texts that have already been created, processed, and are available to us;

2) other existing datasets of Bulgarian texts that need to be reviewed, downloaded, and have their formats converted to those of the IfGPT dataset;

3) compilation of new datasets through targeted crawling and processing of identified texts for filtering, cleaning, deduplication, and the addition of metadata.

IfGPT dataset

Mandatory metadata for each document includes: Identifier, Licence, Document Title, Medium – text, audio, image or video; URL, Domain, Keywords, Number of words, Number of sentences, Number of tokens, Personally identifiable information – the percentage of tokens in the total number of tokens in the document, Biased information – the percentage of tokens in the total number of tokens in the document.

IfGPT dataset

Optional metadata for each document includes: Author, Style, Type – specifies the type of the source document (e.g. book, chapter, essay, newspaper article, blog post, etc.), Subdomain, Translated document, Collection date, Number of paragraphs, Task categories.

IfGPT dataset as of 01.01.2025

Source	# texts	# words	License
MARCELL	25K	45M	Public domain
CURLICAT	113K	35M	Creative Commons (CC)
BulNC Administrative	17K	79M	Public domain
BulNC Wikipedia	89K	41M	CC / GNU
BulNC Subtitles	146K	27M	OPUS

01.10.2025

Selection of datasets


```
{
  "searchResults": {
    "statistics": {
      "numberOfTextSamples": 4570,
      "numberOfSentences": 272195,
      "numberOfTokens": 6560464
    },
    "selectedDocuments": [
      {
```


- A **selection protocol**, based on texts metadata, has been developed.
- An **online search** within a database containing metadata is available on the project website.
- Using this search, the datasets suitable for fine-tuning experiments in Bulgarian and for a specific task within a particular thematic area can be easily selected.

Acknowledgments

- The work is part of the project *Infrastructure for Fine-tuning Pre-trained Large Language Models*, Grant Agreement No. ПВУ – 55 from 12.12.2024 /BG-RRP-2.017-0030-C01/.

- *The International Forum on Advanced ICT Research and Innovation is organized by the Science and Innovation Council on Information and Communication Technologies, and Bulgarian Academy of Sciences under Investment C2.I2 "Increasing the innovation capacity of the Bulgarian Academy of Sciences in the field of green and digital technologies", National Recovery and Resilience Plan.*

Funded by the European Union
NextGenerationEU

National Recovery and Resilience Plan
of the Republic of Bulgaria

http://ifgpt.dcl.bas.bg

svetla@dcl.bas.bg

01.10.2025