IfGPT: A Dataset in Bulgarian for Large Language Models

Svetla Koeva
DCL – IBL,
Bulgarian Academy of Sciences, Sofia, Bulgaria
svetla@dcl.bas.bg

Ivelina Stoyanova
DCL – IBL,
Bulgarian Academy of Sciences, Sofia, Bulgaria
iva@dcl.bas.bg

Jordan Kralev
DCL – IBL;
Technological University
Sofia, Bulgaria
jkralev@dcl.bas.bg

Abstract

The paper presents the large dataset **IfGPT**, which contains available corpora and datasets for Bulgarian, and describes methods to continuously expand it with unduplicated and unbiased Bulgarian data. The samples in the dataset are annotated with metadata that enable effective extraction of domain- and application-oriented datasets for fine-tuning or Retrieval Augmented Generation (RAG) of large language models (LLMs). The paper focuses on the description of the extended metadata of the **IfGPT** dataset and its management in a graph database.

1 Introduction

The large-scale transformer-based models (Vaswani et al., 2017) have significantly changed the state of the art in language processing. There are two basic steps in the development of LLMs, both of which have to do with datasets: Pre-training on large text data and subsequent fine-tuning for a specific task with suitable data.

Developing datasets for LLMs is a major challenge for languages with limited resources. These include:

Data scarcity There are few sources for compiling large datasets for pre-training and fine-tuning LLMs for languages such as Bulgarian, whose relatively low production of authentic digital texts is predetermined by the relatively small number of its speakers.[1]

Copyright restrictions It is even more difficult to find datasets that do not raise copyright issues and are available for both non-commercial and commercial use.[2]

Quality of the data Freely accessible data is often noisy and inhomogeneous and can therefore cause problems or lead to distortions. Procedures for data cleansing and selecting only high-quality texts further limit the scope of the data.

In this paper, we present the **large dataset IfGPT**,[3] which contains some already available corpora and datasets for Bulgarian, as well as methods for its continuous expansion with non-duplicated, clean Bulgarian data. The samples in the dataset are annotated with metadata that enable effective extraction of domain- and application-oriented datasets. The paper focuses on the description of the extended metadata of the **IfGPT** dataset and its management in a graph-based database.

The aim is to avoid the redundant compilation of datasets by different users and the multiple efforts for cleaning the data and to facilitate the reuse of the data for solving different application tasks. The main contribution of our work can be summarised as follows:

(a) Merging several relatively large text collections for Bulgarian into one dataset with standardised metadata description and document formats.

(b) Adding new texts to the dataset in a standardised way.

(c) Deploying and customising a set of tools in a chain for text cleaning, deduplication, detection of sensitive and biassed information to ensure the quality of the data.

(d) Providing a uniform metadata description for all documents in the datasets and organising the metadata categories in a graph representation, originally proposed for the Bulgarian National Corpus (Koeva et al., 2012) and extended to the present **IfGPT** dataset.

(e) Providing means to efficiently query metadata to find suitable text documents for a given

[1] https://datareportal.com/reports/digital-2025-bulgaria

[2] The Bulgarian Intellectual Property Rights Act of 2023 liberalises the use of texts that are accessible digitally or in digital form for automatic analysis, but some proprietary collections that are protected by copyright and are not accessible.

[3] https://ifgpt.dcl.bas.bg/en/

Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languages associated with RANLP 2025, pages 65–75, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-100-4-007

LLM fine-tuning or Retrieval Augmented Generation (RAG) task.

2 Large text datasets

Recent advances in the development of LLMs have demonstrated the effectiveness of their pre-training on large text datasets. Despite the fact that some technologies enable shorter parts of training datasets for specific domains and/or languages, the growing demand for language modelling data for most languages, including Bulgarian, remains a challenge. Here we will briefly present some of the widely used and recently created large text datasets used for pre-training.

CommonCrawl creates and maintains an open web crawl dataset. Since 2008, CommonCrawl has collected petabytes of data, including raw web page data, metadata, and text extractions. CommonCrawl is typically used to retrieve subsets of websites during a specific time period. Due to the noisy and low-quality information in web data (Luccioni and Viviano, 2021), it is necessary to clean and filter the data before using it. There are a number of filtered datasets based on CommonCrawl including Bulgarian. **OSCAR** (Open Super-large Crawled Aggregated coRpus) is a large multilingual corpus created by language classification and filtering of the CommonCrawl dataset (Abadji et al., 2022). It covers 152 languages and offers both original and deduplicated versions of the data. Similarly, the **C4** (Raffel et al., 2020) and **mC4** (Xue et al., 2020) datasets were derived from Common Crawl. These corpora were created using heuristic methods to filter out non-linguistic content (such as boilerplate or noise) and underwent extensive deduplication. While **C4** was developed for English only, **mC4** covers over 100 languages. Another related resource is **CC100** (Conneau et al., 2020), which provides monolingual data for more than 100 languages. It was created by processing CommonCrawl snapshots collected between January and December 2018.

Many of the large datasets do not contain Bulgarian, e.g. **Pile**, an 825 GB English text corpus developed for large-scale language model training (Gao et al., 2020); **MassiveText**, a collection of large English language text datasets from various sources, including websites, books, news articles and code (Rae et al., 2022), etc.

There are several studies that present available datasets and categorise them under different aspects: (1) Pre-training Corpora; (2) Instruction Fine-tuning Datasets; (3) Preference Datasets; (4) Evaluation Datasets; (5) Traditional Natural Language Processing Corpora (Liu et al., 2024; Lu et al., 2024). The **IfGPT** dataset presented here can be used as (part of) a pre-training dataset, a fine-tuning dataset (with some modifications), an evaluation dataset (with some modifications), and a traditional natural language processing dataset. However, our motivation for its compilation, management and extension is the fine-tuning of LLMs or RAG applications.

3 Data sources for IfGPT dataset

When collecting and pre-processing data for fine-tuning LLMs, the aim is to collect as much diverse Bulgarian language data as possible that is human-generated, of high quality, does not contain sensitive, false or ethically unacceptable information, is not repetitive and is accompanied by accurate information about its source and the licence for its use.

The components of the **IfGPT** dataset can be categorised into three main groups depending on the type of text, its composition and its possible uses: 1) collections of texts (corpora) that have already been created and processed and are available to us, 2) other existing datasets of Bulgarian texts that need to be reviewed, downloaded and, if necessary, the format of the texts and metadata converted to the format and metadata of the **IfGPT** dataset, 3) compilation of new datasets through targeted crawling and processing of the identified texts for filtering, cleaning, deduplicating and adding metadata.

3.1 Brief description of existing text collections (corpora)

The existing text collections include corpora created for linguistic and corpus-related studies and corpora created for various NLP projects, e.g. for training machine translation systems. The **Bulgarian National Corpus (BulNC)** contains a wide range of texts of different sizes, different media types (written and spoken), different styles, different time periods (synchronous and diachronic) and different licences. Each text in the collection is labelled with metadata (Koeva et al., 2012). BulNC was originally compiled from the Bulgarian Lexicographical Archive and the Text Archive for Written Bulgarian, which make up 55.95% of the corpus.

Later, the EMEA corpus (medical administrative texts) and the OpenSubtitles corpus (film subtitles) were added, accounting for 1.27% and 8.61% of BulNC respectively. The remaining texts were automatically crawled and include a large number of administrative texts, news from monolingual and multilingual sources, scientific texts and popular science texts. The BulNC currently contains around 420,000,000 words and more than 10,000 text samples. Each text sample is provided with a detailed metadata description in a separate file, which makes it possible to extract subcorpora from specific domains and, if permitted, to distribute them with original licences. The texts are stored both in a word-per-line format (or 'vertical' format, in which each line contains a token, its lemma, the part of speech and grammatical features) and in a raw text format.

The dataset **General News in Bulgarian** contains news from different thematic domains. The news items and their metadata were collected automatically from various (mainly Bulgarian) Internet sources: 11,840 web domains and 2,116,739 web pages. The total number of words in the collected general news in Bulgarian language amounts to 601,330,975 words, spread over 33,375,366 sentences and about 28,000 texts. A crawling platform was used for the identification and collection of monolingual data from web pages, the removal of near-duplicates at the document level, and text normalisation and cleaning (Koeva et al., 2020). The extracted texts were structured into JSON files containing extracted metadata and an automatic categorisation of the content into 185 thematic domains (ordered by probability). The main domains with the largest number of documents are: Economics; Sociology; Politics; Law; Business; Commerce; Education; Administration; School; Leisure; and History. The links to the original sources and the distribution licences (if indicated in the sources) are part of the metadata.

The corpus **Bulgarian CURLICAT (Curated Multilingual Language Resources for CEF.AT)** consists of texts from various sources (Váradi et al., 2022). The collection comprises 113,087 documents divided into seven thematic domains: Culture, Education, European Union, Finance, Politics, Economy and Science. All documents are licenced under CC-BY, CC-BY-SA and CC-BY-NC. The texts are linguistically annotated and are available in CoNLL-U Plus format.[4]

The corpus **Bulgarian MARCELL (Multilingual resources for CEF.AT in the legal domain)** consists of legislative documents divided into fifteen types (Váradi et al., 2020). The time span of the documents ranges from 1946 to 2023 and the texts were extracted from the Bulgarian State Gazette, the official gazette of the Bulgarian government, in which documents from official institutions such as the government, the Bulgarian National Assembly, the Constitutional Court, etc. are published. The Bulgarian corpus consists of 25,283 documents categorised into eleven types: Administrative Court; Agreements; Amendments, legal acts; Conventions; Decrees; Decrees of the Council of Ministers; Directives; Instructions; Laws (legal acts); Memoranda; Resolutions. The documents were annotated in CoNLL-U Plus format. The dataset comprises around 45,000,000 tokens and 3,281,000 sentences.

Our work on the datasets already available to us is currently focused on three directions: Identifying texts that are suitable for distribution (with appropriate licences and not duplicated in other selected parts); standardising the format of the texts provided in addition to the original formats, raw text format and JSONL format;[5] and, where necessary, harmonising metadata (categories and values).

3.2 Use of other available datasets

In recent years, many large datasets have been created and gradually expanded with new data, with a focus on open datasets without usage restrictions. These include CommonCrawl;[6] and its cleaned derivatives such as C4[7] and CC-100;[8] OPUS Corpora;[9] etc.

Datasets are also distributed via well-known language repositories such as **ELG**, **CLARIN**, **GitHub**, **HuggingFace**, etc. For example, at the time of writing, HuggingFace has 258 text datasets containing Bulgarian; the ELG catalogue has 388 corpora containing Bulgarian; etc.

The main problems with these are that: (a) Bulgarian and other low-resource languages are rarely included; (b) if they are, they are only a small part

[4] https://universaldependencies.org/ext-format.html

[5] https://jsonlines.org/

[6] https://commoncrawl.org

[7] https://github.com/google-research/text-to-text-transfer-transformer#c4

[8] https://data.statmt.org/cc-100/

[9] https://opus.nlpl.eu/

of the data; (c) they may already be included in the datasets available to us; (d) they may not fulfil the quality requirements both in terms of overall data quality and suitability for training; (e) their availability on the web often means that they have already been included in the LLMs.

The aim here is to avoid overlaps with texts that have already been collected and to carry out a massive textual clean-up in order to filter out malformed texts and irrelevant data. The next step is to assign as many metadata as possible and convert the documents into the standardised format.

3.3 Compilation of new datasets through targeted crawling

A regularly updated source for the provision of new text data has been identified:

(a) Repositories for scientific papers, dissertations and other research publications such as: **Bulgarian Portal for Open Science**, a platform providing free access to full texts of articles published in Bulgarian scientific journals, selected scientific books together with extensive bibliographic metadata, etc.; scientific and popular science journals and blogs; websites of universities and research institutions publishing scientific papers, dissertations, etc. from various domains.

(b) **Public administrative data** provided by the Bulgarian National Assembly (parliamentary minutes and the Government Gazette), ministries, agencies and municipalities.

(c) Data from **websites and technical documentation of companies** from various domains and with appropriate licences.

(d) **Websites of media**: newspapers, television and radio stations that publish news from various domains and have appropriate licences.

Sources that have already been used for the collection of resources (see 3.1) can be monitored and crawled to update the datasets. When adding new text samples, the same format of the text, metadata and annotations is used to ensure compatibility with the procedures for validation, data enrichment and extraction of subsets of the data. In addition, the metadata provides a reliable means of filtering data (by source, year, domain, etc.) for more efficient deduplication (see 3.4.1).

To this end, we need reliable means to assess data diversity and techniques to improve it. Particular attention should be paid to less frequent linguistic phenomena, which firstly are not well captured in smaller datasets and secondly are crucial for ensuring and maintaining linguistic diversity.

One of the biggest challenges is to find and use data with suitable licences that allow sharing of the data (as part of the dataset). Many existing datasets disregard the restrictions on sharing and consider it sufficient to provide appropriate references to the source and authorship of the text samples.

3.4 Procedures for improving the quality of the dataset

Any application that needs to reliably represent a domain requires diverse, balanced and unbiased data. The following techniques are important to provide high quality data.

3.4.1 Removing duplicates

Deduplication has been shown to improve the quality of data and the performance of LLMs, in particular by removing overlap between training and test data, allowing for more reliable evaluation (Lee et al., 2022).

The first pre-filtering step relies on metadata and involves matching texts by source, year, domain, title, author, etc. to quickly identify and remove identical text samples which come from different dataset sources. This significantly improves the efficiency of further deduplication.

The main deduplication method we implement is based on the MinHash and Locality Sensitive Hashing (LSH) algorithm, which is widely used for this purpose (Leskovec et al., 2020; Lee et al., 2022; Albalak et al., 2024) and which provides an efficient way to identify even near-duplicates. The algorithm estimates the n-gram similarity between all pairs of text samples and identifies those with high n-gram overlap.

The deduplication procedures are implemented in a pipeline to facilitate ongoing deduplication as the dataset is regularly updated with new texts.

3.4.2 Handling formatting, boilerplate, web navigation elements from texts

For the extraction of raw text from HTML documents, we used CSS selectors to mark the elements we wanted to extract. In addition, various techniques are used for raw text extraction (Koeva et al., 2020): automatic correction of hyphenated words based on vocabulary, regular expressions to filter out metadata, sentence tokenisation and language detection to filter out non-Bulgarian sentences, etc. Since there are also PDF documents, we used a

PDF to text converter to extract text data. Additional scripts were written to remove headers and footers from the PDF documents. The extracted paragraphs were merged based on a heuristic analysis of capitalisation and lexical content when a sentence crossed a paragraph boundary. Scanned and OCR-recognised PDF files were not processed due to their lower quality, and text and paragraphs written in languages other than Bulgarian (mostly English) were removed.

3.4.3 Identification of sensitive personal data, biases, etc.

A pressing ethical issue that is essential in the development of large datasets with diverse sources is the identification and removal (e.g. masking) of personally identifiable information (Kober et al., 2023). However, the identification and removal of personally identifiable information is a difficult task due to the different types and forms as well as the inconsistent definitions, especially in various data protection laws (Song et al., 2025). A number of methods have been developed, including those based on machine learning techniques (Kulkarni and Cauvery, 2021; Shahriar et al., 2024), Transformers (Johnson et al., 2020; Shahriar et al., 2024) and rule-based identification (Jaikumar et al., 2023) as well as masking or tokenisation to remove personally identifiable information. In our approach, we experimented with the MAPA anonymisation package for Bulgarian[10] and with some naive rule-based methods to detect sentences of the document with potentially sensitive information and mark their number per document in the metadata.

The increasing development of LLMs has led to consideration of the biases inherent in them, resulting in the development of a range of techniques to measure and eliminate bias, particularly in relation to social issues. The main groups of techniques that address bias include: (a) the introduction of metrics to assess and identify bias in datasets; (b) techniques to reduce bias in the pre-processing, training and post-processing stages. Gallegos et al. (2024) summarises a wide range of current research focused on better understanding and preventing the propagation of bias in LLMs. Our goal is to score the documents in our dataset according to the percentage of potentially biased or abusive sentences and include this information in the metadata for further use, text filtering, etc. In this way, we can make a selection of documents for fine-tuning without sensitive and biased content, but we can also use the data for further research on bias. Currently, the classification of potentially biased sentences is being developed.

[10] https://mapa-project.eu/

3.5 Current structure of the IfGPT dataset

The current structure of the **IfGPT** dataset in terms of the source datasets of the text samples, the domain distribution and the size is shown in Table 1. The newly compiled dataset has a standardised representation of the metadata and text formats and was subjected to the data quality improvement procedures (see 3.4). The process of expanding **IfGPT** dataset with clean data is ongoing.

Source	# texts	# tokens	Licence
MARCELL	25K	45M	PD
CURLICAT	113K	35M	CC
BulNC Admin	17K	79M	PD
BulNC Wikipedia	89K	41M	CC/GNU
BulNC Subtitles	146K	27M	OPUS

Table 1: Current structure of IfGPT (August 2025). Licences: PD – public domain, CC – Creative Commons (various), GNU – GNU Free Documentation License, other open or restrictive licenses.

The metadata description of the texts within the **IfGPT** dataset is available to search and extract subsets.[11]

4 File format

Some of the documents in the **IfGPT** dataset are already available in vertical format, in CoNLL-U Plus format or in JSON format. The metadata is included in both the CoNLL-U Plus and JSON format, while in the vertical format the metadata is available in separate associated files. All documents are also saved in raw text format before being annotated and converted to either CoNLL-U Plus or JSON format.

The metadata descriptions are in the form of attribute-value pairs. For some categories, the values are predefined, e.g. for the media type, for others, e.g. the title of the document, any value is permitted.

The IfGPT dataset is provided in JSONL format for the LLM tasks, but the other available format versions can be requested if required.

[11] https://ifgpt.dcl.bas.bg/ifgpt-dataset/

5 Metadata categories

Metadata is essential to ensure efficient and effective selection of datasets for fine-tuning and RAG for specific domains and applications. Fine-tuning of LLMs is performed as a language-dependent task, focusing on a specific language, in our case Bulgarian, and further reducing the scope to a specific domain, task, etc. This requires the selection of a dataset with relevant data to ensure successful fine-tuning. On the other hand, metadata can be used not only for the selection of datasets suitable for RAG, but also for more effective methods of filtering information in RAG based on metadata (Bruni et al., 2025). Even though we emphasise the importance of metadata, we must point out that the empirical evaluation of the efficiency of metadata descriptions is beyond the scope of this study.

All four text collections (described in 3.1) have been supplied with metadata. The metadata of the Bulgarian National Corpus is aimed at searching and retrieving information for the needs of corpus and language research in general and therefore has a complex graph-based structure of related categories (Koeva et al., 2016). The metadata for the resource General News in Bulgarian is simply a categorisation into up to six most likely thematic domains (sports, politics, history, etc.). The metadata for the other two multilingual resources that also contain Bulgarian (MARCELL and CURLICAT) are synchronised between the different languages to form a single subset of categories. All four resources have overlapping metadata, and based on our task we have defined a set of metadata that is mandatory for each document (regardless of whether there are categories with a null value) and metadata that is optional. Optional metadata is metadata that is already assigned to the document but is not part of the mandatory metadata.

The following mandatory metadata is defined for the documents:

Identifier – unique identifier of the document in all collections, created with the language code `bg` as a prefix;

Licence – the conditions for use, i.e. CC BY-SA 4.0 licence;

PublicationDate – the date of original publication of the document (if available) in ISO 8601 format;

DocumentTitle – human-readable title (name) of the document;

Source – the name of the organisation that published the source document, i.e. journal, publisher, blog, website, etc.;

Medium – whether the document is text, audio, image or video;

Url – the original individual address where the document was retrieved from, if applicable;

Domain – classification of a specific thematic domain selected from a predefined list of 24 domains; up to six domains can be listed;

Keywords – extracted terms that specify the document; up to six keywords can be listed;

NumberWords – the total number of words in the document;

NumberSentences – the total number of sentences in the document;

NumberTokens – the total number of tokens in the document;

PersonallyIdentifiableInformation – the percentage of tokens in the total number of tokens in the document;

BiasedInformation – the percentage of tokens in the total number of tokens in the document.

The following metadata is optional for the documents:

Author – name(s) of the person(s) who created the text in the source document;

Style – the literary style of the text in the document, selected from a predefined list: Fantasy, Administrative, Legal, Journalism, etc.;

Type – specifies the type of the source document (e.g. book, chapter, essay, newspaper article, blog post, etc.);

Subdomain – a further classification of documents into narrower categories, e.g. scientific domains for the field of science or cultural domains for the field of culture; a subdomain is linked to a specific domain;

TranslatedDocument – whether the document was originally created in Bulgarian or whether it has been translated;

CollectionDate – the date of collection of the document in ISO 8601 format;

LicenseLink – the link to the licence on the source's website, if available;

NumberParagraph – the total number of paragraphs in the document;

TaskCategories - the applications (selected from a predefined list) for which the template was developed or is suitable, e.g. for question-answering.

Some of the metadata values are extracted automatically. The main techniques for automatically extracting metadata are: (a) metatextual techniques, which consist of extracting information from the HTML markup of the original files; and (b) textual techniques, which consist of text analysis and heuristics using a set of language resources. The following metadata values are automatically extracted from the HTML sources: Author, DocumentTitle, PublicationDate. The classification information includes the thematic domain of the texts, their genre and type as well as the results of the text analysis. In some cases, the source may contain classification labels according to an assumed domain and/or genre classification of the source, e.g. texts on a news website may be divided into editorials and articles of different domains – business, sports, etc.

Some metadata values are generated automatically. These are statistical information resulting from the processing of the text that includes the number of words, tokens, sentences, etc. Administrative metadata such as the document identifier, language code and source are also generated.

In order to improve the quality and quantity of the metadata used to describe each text entry, several procedures are defined. These procedures aim to identify contextually relevant descriptors to fill in missing values in the metadata. The reasons for incomplete data are manifold: in some cases, the data is not collected (by the users/authors), the website where the text is stored does not store certain types of data, it may be difficult or impossible to extract it from the online source, or it may be the result of data integration errors.

The task can be performed as a multi-class classification using heuristics, statistical methods or machine learning. It has been pointed out that traditional statistical methods for data imputation often do not provide an accurate and comprehensive description as they do not analyse the semantic context and relationships within the data (Jin et al., 2025). Mei et al. (2021) propose the use of a pretrained language model to assign metadata based on semantic features of the text and its description. Alyafeai et al. (2025) uses LLMs to automatically extract metadata from scientific articles by analysing context length and few-shot learning.

So far, we use more traditional methods for extracting metadata – statistical and rule-based, depending on the source of the document, the original format of the document (PDF, HTML, etc.) and the structure of the document itself. As we want to harmonise the metadata of existing datasets and new incoming texts, extending and standardising the metadata of available documents may require re-crawling the sources and repeating the text extraction process. We will upgrade the methods and tolls we use (Koeva et al., 2020) with the functionalities of applications like Trafilatura (Alyafeai et al., 2025), Maker,[12] etc.

6 Metadata management

Graph databases are designed to efficiently process large amounts of interconnected data. They can be scaled horizontally by adding more nodes to the database, while maintaining performance even for complex queries. The most commonly used graph databases are Neo4J, Microsoft Azure Cosmos DB, ArangoDB, TigerGraph and Amazon Neptune. Neo4J[13] is one of the most popular graph databases due to its high performance, support for the Cypher query language (Francis et al., 2018) and strong community support.

To effectively utilise the properties of a graph database when storing metadata, a schema is designed that captures the most important entities and their connections. The nodes of the metadata schema are defined as follows:

Document nodes with the properties: **Identifier, Title, Source, Domain, Author, Licence**, etc.;

Domain nodes with the properties **Name** and **Parent_category**;

Author nodes with the properties **Name** and optional details such as **Biography**;

Source nodes with the properties **Name** and **Url**;

Licence nodes with a single property **Type**.

The graph edges, which represent the relations between the nodes, are defined as follows:

Document-Domain of type BELONGS_TO;

Domain-Domain of type SUBCATEGORY_OF;

Document-License of type LICENSED_WITH;

Document-Author of type WRITTEN_BY;

Document-Source of type PUBLISHED_IN.

[12] https://github.com/datalab-to/marker
[13] https://neo4j.com/

```
CREATE (d:Document {id: "bg-bnc-2011040848215", title: "Ото
Варбуpг", author: "ChuispastonBot", source:
"bg.wikipedia.org", publication_date: "2011-04-08", domain:
"SCIENCE", subdomain: "BIOLOGY", license: "CC-BY-SA"})

CREATE (c1:Domain {name: "SCIENCE"})
CREATE (c2:Domain {name: "BIOLOGY"})

CREATE (a:Author {name: "ChuispastonBot"})
CREATE (s:Source {name: "bg.wikipedia.org", url:
"http://bg.wikipedia.org/"})
CREATE (l:License {type: "CC-BY-SA"})

// Create relationships
MATCH (d:Document {id: "bg-bnc-2011040848215"}), (c2:Domain
{name: "BIOLOGY"})
CREATE (d)-[:BELONGS_TO]->(c2)

MATCH (c2:Category {name: "BIOLOGY"}), (c1:Domain {name:
"SCIENCE"})
CREATE (c2)-[:SUBCATEGORY_OF]->(c1)

MATCH (d:Document {id: "bg-bnc-2011040848215"}), (a:Author
{name: "ChuispastonBot"})
CREATE (d)-[:WRITTEN_BY]->(a)

MATCH (d:Document {id: "bg-bnc-2011040848215"}), (s:Source
{name: "bg.wikipedia.org"})
CREATE (d)-[:PUBLISHED_IN]->(s)

MATCH (d:Document {id: "bg-bnc-2011040848215"}), (l:License
{type: "CC-BY-SA"})
CREATE (d)-[:LICENSED_UNDER]->(l)
```

Example 1: Processing a document with Cypher QL

Integrating the datasets into a vector database such as ChromaDB[14] can improve the efficiency of storing and querying vector representations of text data, which is critical for RAG technology. The conversion of text data into vector representations can be done using embeddings generated by specialised models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) or other transformer-based architectures. The purpose of these vectors is to capture semantic information about the text and use it for similarity searches.

To integrate ChromaDB into a graph database, the unique identifier is stored for each document or vector representation, depending on the granularity required for a particular task. An example workflow for processing Bulgarian texts with ChromaDB is presented below:

→ Vectorise the text using a selected embedding model. Save these vectors in ChromaDB with unique IDs.

→ Store metadata in the Neo4J graph database by creating nodes for sentences or documents and storing metadata such as Author, Source and Domain.

→ Also save the vector ID from ChromaDB as a property of the node.

→ Query similar documents using ChromaDB to find the closest vectors to a given query vector and retrieve the IDs of these vectors. Use this to query the graph database for the corresponding metadata.

[14] https://www.trychroma.com/

7 Conclusion

The most important results reported in this paper include the compilation of the **IfGPT** dataset for Bulgarian and the development of a metadata schema with graph-structured categories that enables efficient searching in the metadata. We also provide an online search interface in the metadata that enables the identification of smaller datasets tailored to specific domains and applications.

The metadata description of the **IfGPT** dataset contains a large number of categories that describe the text samples on different levels. Some of the most important metadata categories for the compilation of domain- and application-specific datasets are the following:

Domain information: A set of characteristics used to comprehensively describe the domain of the text was produced, including style, domain, subdomain. The source can also provide information about the domain, e.g. scientific journals in different domains or subsections of a news source.

Keywords: A schematic description of the content of the text sample can be created automatically based on the title, abstract (if available) or full text.

Sensitive personal data and biases: The parts containing sensitive personal data and biases are not removed or replaced by neutral data, but the percentage of such content in a document is calculated and can thus vary the strictness of the criteria for exclusion from certain datasets.

Using a graph database to store metadata offers several advantages over traditional relational databases or file-based systems. One of the main advantages is the ability to effectively model complex relationships between linguistic entities.

To summarise, a suitable dataset such as **IfGPT** – as large as possible, equipped with rich metadata for efficient search and retrieval of suitable documents, clearly defined tasks and thematic domains, and adequately managed with a graph database integrated with a database of embeddings – will enable fast and efficient fine-tuning of LLMs and Retrieval Augmented Generation.

Acknowledgment

The present study is carried out within 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/.

Limitations

The main practical constraints involve the lack of extensive and diverse sources for collecting texts from specialised domains in Bulgarian. Additionally, specialised texts are often distributed in PDF format, which presents challenges for maintaining high text quality in the data.

For metadata, automatic collection may be inadequate, as online sources often provide limited information about the text. Conversely, manual metadata description is inefficient in terms of human effort and time. As high-quality metadata is important for correct dataset selection, some evaluation metrics for automatically assigned metadata, ensuring its completeness and consistency, need to be developed.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv e-prints*, page arXiv:2201.06642.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A Survey on Data Selection for Language Models. *Transactions on Machine Learning Research*.

Zaid Alyafeai, Maged S. Al-Shaibani, and Bernard Ghanem. 2025. MOLE: Metadata extraction and validation in scientific papers using llms. *arXiv e-prints*, page page arXiv: 2505.19800.

Davide Bruni, Marco Avvenuti, Nicola Tonellotto, and Maurizio Tesconi. 2025. AMAQA: A metadata-based qa dataset for rag systems. *arXiv e-prints*, page page arXiv: 2505.13557.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nadime Francis, Alastair Green, Paolo Guagliardo, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Stefan Plantikow, Mats Rydberg, Petra Selmer, and Andrés Taylor. 2018. Cypher: An Evolving Query Language for Property Graphs. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, page 1433–1445, New York, NY, USA. Association for Computing Machinery.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling.

Jaikishan Jaikumar, Mohana, and Pavankumar Suresh. 2023. Privacy-Preserving Personal Identifiable Information (PII) Label Detection Using Machine Learning. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5.

Can Jin, Tong Che, Hongwu Peng, Yiyuan Li, Dimitris N. Metaxas, and Marco Pavone. 2025. Learning from teaching regularization: generalizable correlations should be easy to imitate. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Alastair E.W. Johnson, Lucas Bulgarelli, and Tom J. Pollard. 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference and Learning, Toronto, Ontario, Canada 2020*, pages 214–221.

Maria Kober, Jordan Samhi, Steven Arzt, Tegawendé F. Bissyandé, and Jacques Klein. 2023. Sensitive and personal data: What exactly are you talking about? In *MOBILESoft*, pages 70–74.

Svetla Koeva, Nikola Obreshkov, and Martin Yalamov. 2020. Natural language processing pipeline to annotate Bulgarian legislative documents. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6988–6994, Marseille, France. European Language Resources Association.

Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova, and Ekaterina Tarpomanova. 2012. The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, (1):65–110.

Svetla Koeva, Ivelina Stoyanova, Maria Todorova, Svetlozara Leseva, and Tsvetana Dimitrova. 2016. Metadata extraction, representation and management within the Bulgarian National Corpus. In *Priceedings of the 4th Workshop on Challenges in the Management of Large Corpora*, pages 33–39. ELDA.

Poornima Kulkarni and N. K. Cauvery. 2021. Personally Identifiable Information PII Detection in the Unstructured Large Text Corpus using Natural Language Processing and Unsupervised Learning Technique. *International Journal of Advanced Computer Science and Applications*, 12(9).

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2020. *Mining of Massive Datasets*, 3rd edition. Cambridge University Press.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. 2024. Datasets for Large Language Models: A Comprehensive Survey. *arXiv e-prints*, page arXiv:2402.18041.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, page arXiv:1907.11692.

Yuting Lu, Chao Sun, Yuchao Yan, Hegong Zhu, Dongdong Song, Qing Peng, Li Yu, Xiaozheng Wang, Jian Jiang, and Xiaolong Ye. 2024. A Comprehensive Survey of Datasets for Large Language Model Evaluation. In *2024 5th Information Communication Technologies Conference (ICTC)*, pages 330–336.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

Yinan Mei, Shaoxu Song, Chenguang Fang, Haifeng Yang, Jingyun Fang, and Jiang Long. 2021. Capturing Semantics for Imputation with Pre-trained Language Models. In *Proceedings of the 2021 IEEE 37th International Conference on Data Engineering (ICDE), Chania, Greece, 2021*, pages 61–72.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis & insights from training gopher. *arXiv e-prints*, page arXiv:2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Md Hasan Shahriar, Anne V. D. M. Kayem, David Reich, and Christoph Meinel. 2024. Identifying personal identifiable information (PII) in unstructured text: A comparative study on transformers. In *Database and Expert Systems Applications: 35th International Conference, DEXA 2024, Naples, Italy, August 26–28, 2024, Proceedings, Part II*, page 174–181, Berlin, Heidelberg. Springer-Verlag.

Qiurong Song, Yanlai Wu, Rie Helene (Lindy) Hernandez, Yao Li, Yubo Kou, and Xinning Gui. 2025. Understanding Users' Perception of Personally Identifiable Information. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 26 April - 1 May 2025*. Association for Computing Machinery, New York.

Tamás Váradi, Svetla Koeva, Martin Yamalov, Marko Tadić, Bálint Sass, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pęzik, Verginica Barbu Mititelu, Radu Ion, Elena Irimia, Maria Mitrofan, Vasile Păiș, Dan Tufiș, Radovan Garabík, Simon Krek, Andraz Repar, Matjaž Rihtar, and Janez Brank. 2020. The MARCELL Legislative Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3761–3768, Marseille, France. European Language Resources Association.

Tamás Váradi, Bence Nyéki, Svetla Koeva, Marko Tadić, Vanja Štefanec, Maciej Ogrodniczuk, Bartłomiej Nitoń, Piotr Pęzik, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan, Dan Tufiș,

Radovan Garabík, Simon Krek, and Andraž Repar. 2022. Introducing the CURLICAT Corpora: Seven-language Domain Specific Annotated Corpora from Curated Sources. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 100–108, Marseille, France. European Language Resources Association.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.