Финансирано от Европейския съюз

СледващоПоколениеЕС

План за възстановяване и устойчивост

Република България

# ПРОТОКОЛ ЗА РАЗРАБОТВАНЕ НА ЛАБОРАТОРЕН ПРОТОТИП

на **дигитален продукт (езиков ресурс)**

# ГОЛЯМ НАБОР ОТ ЕЗИКОВИ ДАННИ IfGPT ЗА БЪЛГАРСКИ ЕЗИК,

планиран резултат от
**Работен пакет 4. Колекциониране и предварителна обработка на данни за фина настройка на големи езикови модели**
по проекта „Инфраструктура за фина настройка на предварително обучени големи езикови модели“, договор No ПВУ – 55 от 12.12.2024 г. /BG-RRP-2.017-0030-C01/
от Механизма за възстановяване и устойчивост за изпълнение на инвестиция по C2I2 „Повишаване на иновационния капацитет на Българската академия на науките (БАН) в сферата на зелените и цифровите технологии“ от Плана за възстановяване и устойчивост

**Протокол: IFGPT-2025-DATA-001**
**Версия на протокола: 1.0**

Име и версия на продукта: **IfGPT version 1.0**
Описание на продукта: **Набор от езикови данни IfGPT за български език за фина настройка на големи езикови модели**
Ключови думи: **големи езикови модели, текстови данни, фина настройка, набор от данни**
Keywords: **LLMs, text data, fine-tuning, dataset**

Ръководител: **Светла Коева**
Автори: **Светла Коева, Тинко Тинчев, Ивелина Стоянова, Валентина Стефанова, Светлозара Лесева, Цветана Димитрова, Мария Тодорова, Христина Кукова, Михаела Москова**

Текущо ниво на техническа готовност на продукта: **TRL 4**
Дата: **15.09.2025 г.**

## Общо описание на набора от данни IfGPT

При събирането и предварителната обработка на данни за фина настройка на LLM (материали, свързани с правото на учене), целта е да се съберат възможно най-много разнообразни

човешки данни на български език, с високо качество, чисти от чувствителна, невярна или етично неприемлива информация и придружени от точно описание на техния източник и лицензa за използването им.

**Наборът от данни IfGPT** включва компоненти, които могат условно да бъдат категоризирани в три основни групи в зависимост от вида на текстовете, техния състав и възможните приложения: 1) колекции от текстове (корпуси), които вече са налични, обработени и достъпни, 2) други съществуващи набори от български текстове, които трябва да бъдат прегледани, изтеглени и, ако е необходимо, форматът на текстовете и метаданните да бъдат конвертирани във формата и метаданните на набора от данни IfGPT, 3) компилиране на нови множества от данни от данни чрез целенасочено обхождане и обработка на идентифицираните текстове – филтриране, почистване, дедупликиране и добавяне на метаданни.

Към момента **IfGPT** обхваща 390 хил. документа и 227 млн. токъна, които са основно извлечени от съществуващите текстови колекции. Те включват корпуси, създадени за лингвистични и корпусни изследвания, както и корпуси, създадени по различни проекти, свързани с обработката на български език, например за обучение на системи за машинен превод.

**Организация на данните в IfGPT**

Текущата версия на IfGPT съдържа 390 хил. документа и 227 млн. токъна и включва следните големи корпуси на български език, чиито метаданни са приведени във формата на IfGPT:

- MARCELL (25 хил. текста и 45 млн. токъна),
- CURLICAT (113 хил. текста и 35 млн. токъна),
- Административен корпус на БНК (17 хил. текста и 79 млн. токъна),
- Уикипедия статии от БНК (89 хил. текста и 41 млн. токъна),
- Субтитри от БНК (146 хил. текста и 27 млн. токъна).

Текстовете и метаданните се съхраняват отделно. Метаданните обхващат 23 категории, свързани както с администрирането на набора от текстове (име, източник и др.), така и с възможностите за прецизно извличане на подмножества от данните за различни задачи, свързани с фината настройка (лиценз, тематична област, ключови думи, възможни приложения и др.).

Метаданните са организирани в графова база данни Neo4j за лесно и бързо обхождане и селектиране на подмножества. Базата данни отразява и релациите между различните категории (например, Domain / Subdomain).

**Необходимо оборудване и среда за разработване на прототип на набора от данни IfGPT**

Обработването на данните се извършва локално.

Данните (текстовите данни и метаданните) са предоставени за свободно сваляне в различни публични хранилища като GitHub, HuggingFace.

Предоставено е и онлайн демо за достъп до метаданните и демонстриране на възможностите за извличане на подмножества от данни по избрани параметри.

**Процедура за разработване на прототип на набора от данни IfGPT**

Протоколът за разработване на прототип на набора от данни **IfGPT** включва следните процедури и правила за работа:

**1.** Като изходни данни са използвани вече налични множества от данни, разработени по различни проекти на Института за български език – Българския национален корпус, MARCELL, CURLICAT за български език и др.

**2.** Изработен е модел за представяне на текстовите данни – те са предоставени като списък от отделни изречения без допълнителна анотация.

**3.** Създадени са и са приложени процедури за изчистване на текстовите данни в рамките на всяка текстова единица:

- Отстраняване на изречения, които не са на български език, изречения, които не са пунктуационно оформени, стандартни елементи от оформлението или навигацията на уебстраниците и др.
- Отстранени са много кратки (с дължина под 10 символа) и много дълги (с дължина над 500 символа) изречения.
- Отстранени са многократно повтарящи се изречения в рамките на даден текст.

**4.** Допълнителни процедури са разработени и се прилагат за целия набор от данни:

- Процедури за идентифициране и отстраняване на дубликати и текстове много близки по съдържание (почти дубликати).
- Отстранени са много кратки текстове (под 3 изречения).

Процедурите за осигуряване на качеството са приложени върху всички текстове в настоящата версия на IfGPT. Процедурите са прилагани и за всички новодобавяни текстови единици. Предстои интегриране на процедурите в цялостен компонент на инфраструктурата, който да се занимава с обработката на данните.

**5.** За описанието на всяка текстова единица се прилага подробна система от метаданни, която включва 23 категории (Приложение 1).

- Ако текстовата единица вече е снабдена с метаданни (идва от корпус с описание), тези метаданни се пренасят към формата и типовете на метаданните на IfGPT.
- Ако текстовата единица е извлечена от интернет, от уебстраницата се извличат възможно най-много метаданни за текстовата единица.
- Ако дадени метаданни са валидни за всички текстове от даден източник (например лиценз), те се задават за всички текстове от този източник. Същото може да се приложи и за част от текстовете, дефинирани по определени критерии.
- В редки случаи е възможно метаданните да се допълват и ръчно.
- Допълнителни процедури за допълване на метаданните могат да се извършват преди публикуването на нови версии на набора от данни.

Метаданните се разпространяват заедно с текстовите данни, както и самостоятелно, като позволяват търсене и извличане на подмножества (за които след това могат да бъдат свалени и

текстовите данни, само селектираните текстови данни за по-лесно и бързо обработване на данните).

**6.** Предоставя се онлайн демо версия на системата за търсене в метаданните, която демонстрира основни параметри и възможности за търсене и извличане на подмножества от данните. Същевременно се работи върху функционална система за извличане на подкорпуси, която да се интегрира в инфраструктурата.

**7.** Предвижда се регулярно публикуване на нови версии с разширения на набора от данни IfGPT. Данните се публикуват с изрично посочване на лицензите на текстовите единици, за да се осигури спазването на Закона за авторското право и сродните му права. Ползването на данните от потребителите трябва да е съобразено с предоставената информация за лицензите на индивидуалните текстови единици.

**Приложение 1.**

**Списък на метаданните**
Категории на метаданните:

**Identifier** – уникален идентификатор на документа във всички колекции, създаден с езиков код **bg** като префикс;

**Licence** – условията за използване, например лиценз CC BY-SA 4.0;

**PublicationDate** – датата на оригиналната публикация на документа (ако е налична) в ISO 8601 формат;

**DocumentTitle** – четимо за човек заглавие на документа;

**Source** – името на организацията, която е публикувала изходния документ, т.е. списание, издател, блог, уебсайт и т.н.;

**Medium** – дали документът е текст, звуков файл, изображение или видео;

**Url** – оригиналният интернет адрес, от който е извлечен документът, ако е приложимо;

**Domain** – класификация на документите по предварително зададени (24) специфични тематични области; всеки текст може да се класифицира в до шест домейна;

**Keywords** – извлечени ключови думи, които са специфични за документа; могат да бъдат изброени до шест ключови думи;

**NumberWords** – общият брой думи в документа;

**NumberSentences** – общият брой изречения в документа;

**NumberTokens** – общият брой токъни в документа;

**PersonallyIdentifiableInformation** – масив от стойности, показващи изреченията, които съдържат чувствителна информация, и процента на токъните, съдържащи такава информация, спрямо общия брой токъни в тези изречения;

**BiasedInformation** – масив от стойности, показващи изреченията, които съдържат потенциална информация, показваща пристрастия, и процента на токъните, съдържащи такава информация, спрямо общия брой токъни в тези изречения;

Следните метаданни са незадължителни за документите:

**Author** – име на автора на текста в изходния документ;

**Style** – литературният стил на текста в документа, избран от предварително дефиниран списък: Художествен, Административен, Правен, Журналистически и т.н.;

**Type** – уточнява типа на изходния документ (напр. книга, глава, есе, вестникарска статия, блог публикация и т.н.);

**Subdomain** – допълнителна класификация на документите в по-тесни категории, напр. научни домейни за областта на науката или културни домейни за областта на културата; поддомейнът е свързан с конкретен домейн;

**TranslatedDocument** – дали документът е оригинално създаден на български или е бил преведен;

**CollectionDate** – датата на събиране на документа в ISO 8601 формат;

**LicenseLink** – връзката към лицензa на уебсайта на източника, ако е налична;

**NumberParagraph** – общият брой абзаци в документа;

**TaskCategories** – приложенията (избрани от предварително дефиниран списък), за които данните в документа са разработени или са подходящи, напр. за въпроси и отговори.

# История на редакциите на протокола

| Версия | Дата | Автор | Промени |
|---|---|---|---|
| 1.0 | 15.09.2025 | С. Коева и екип | Първоначален протокол |