



Финансирано от
Европейския съюз
Следващо Поколение ЕС



План за възстановяване и
устойчивост



Република България

ВЪТРЕШЕН ДОКЛАД

от гл. ас. д-р Йордан Кралев
до проф. д-р Светла Коева, ръководител на проекта,

за текуща самооценка на техническата готовност на софтуер

Разработване на чатбот с големи езикови модели и контекстно разширяване на инструкциите

планиран резултат от **Работен пакет 5. Фина настройка на големи езикови модели** по проекта „Инфраструктура за фина настройка на предварително обучени големи езикови модели“, договор No ПБУ – 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

Вътрешен доклад ID: IFGPT-2025-REPORT-005

Име и версия на продукта: **Чатбот с големи езикови модели version 0.1**

Тип на продукта: **Софтуер**

Текущо ниво на техническа готовност: **TRL 4**

Дата: **15.09.2025 г.**

☒ **TRL 3: Начални скриптове и функции (Септември 2025)**

Изисквания:

- Разработени са базови алгоритми, модули или скриптове, които решават ядрото на поставения проблем.
- Не се изисква цялостен интерфейс или архитектура.
- Липсва интеграция между отделни модули.

Доказателства:

- Прототипен код (<https://github.com/DCL-IBL/dcl-summarizer>) на базова архитектура и **интеграция между отделните модули**:
 - Конфигурация на Docker контейнер за векторна база данни ChromaDB за съхранение на векторно-матрично представяне на сегментираните документи
 - Конфигурация на Docker контейнер за сървър на езикови модели Ollama с достъп до графични карти
 - NodeJS Backend сървър за осигуряване на отдалечен програмен интерфейс REST API позволяващ:
 - Зареждане на нови документи във векторната базата данни
 - Семантично търсене по качен документ
 - Обръщение към езиков модел с контекстно разширение на инструкциите
 - HTML5/JS базиран потребителски интерфейс достъпен през NGINX сървър с базова функционалност за качване на документи и получаване на резултат от търсенето в базата данни и обобщението му с езиковия модел
- Документация на архитектурата на системата и алгоритмите:
 - публикация на тема: **“Инфраструктура за създаване на чатбот с големи езикови модели и контекстно разширяване на инструкциите”, сп. Математика и Информатика, ноември, 2025г.**
 - Инструкции за свободно инсталиране и настройка на системата: <https://github.com/DCL-IBL/dcl-summarizer/blob/master/README.md>
- Proof-of-Concept демо (REST API):
 - Интерфейс за изпращане на текстови PDF или TXT документи за обработка (сегментация, векторизация) и съхранение в базата данни
`curl -X POST -F "txtFile=@/home/jordan/curlicat2/export/7fc200cafa637d6890f1729a57cf78f3.txt" http://localhost:8000/api/process-txt`
 - Интерфейс за извършване на контекстно търсене и обобщаване с JSON структуриран вход:
`curl -X POST -H "Content-Type: application/json" -d '{"RAGQuery":"Кой е фараон на Египет?"}' http://localhost:8000/api/process-rag`
 - Интерфейс за изтриване/нулиране на базата данни (с цел провеждане на тестове):
`curl -X POST -H "Content-Type: application/json" -d '{"collection":"text_docs"}' http://localhost:1316/api/clear-db`
- Proof-of-Concept демо (минимален GUI)

RAG Query System

Моделирането е познавателен процес, чиято цел е изучаването на изследваните обекти. Това се постига с изграждането на модели – приближения, отразяващи определени аспекти на обектите. Моделите са фундаментални за човешкото познание. От раждането си човек се учи да интерпретира потока от сетивни данни от околния свят, като с времето формира все по-сложни и по-прецизни представи за света. Така той изгражда своята познавателна система [1], с която осъзнава света.

Например, ако човек попадне в стая, в която не е влизал, с бърз поглед той мигновено „подрежда“ конкретната реалност. Не е нужно да се взира във всеки предмет – неговата познавателна система услужливо попълва всички липси в потока от сетивни данни и така се

Submit Query

Response:

Here is the expected output in JSON format:

```
{
  "keywords": [
    "моделиране",
    "познавателна система",
    "сетивни данни",
    "реална среда",
    "практическо обучение"
  ],
}
```

- o Web базиран потребителски интерфейс - HTML страница, достъпна през изолиран сървър от Backend услугата
- o Форма за въвеждане на заявки включващо текстово поле и бутон за изпращане на формата
- o Динамичен HTML елемент за визуализиране на резултата от обработката като структуриран текст
- o JS функции за изпращане на заявки към управляващата Backend услуга, получаване, обработка и визуализация на получения отговор
- o Стилистични дефиниции на елементите на потребителския интерфейс за постигане на каноничен дизайн с унифицирано поведение върху стационарни и мобилни устройства.
- Unit тестове
 - o Тестове на ChromaDB контейнерът като отделна единица.
 - Осъществяване на достъп

- Създаване на документ
- Изтриване на документ
- Семантично търсене
- Групиране на документи
- Тестове на сървърът на езикови модели с различни езикови модели.
 - Изпращане на chat заявки към инсталирани езикови модели през REST API
 - Генериране на структуриран JSON изход с явно оказване на redundant структура на данните
- Тестове на Backend услугата през командния интерпретатор CLI
 - изпращане на заявките според спецификацията на интерфейса
 - проследяване на логовете от работата на услугата
- Тестове на потребителския интерфейс

✓ TRL 4: Алфа версия – вътрешно тестване (Ноември 2025)

Изисквания:

- Алфа версия със заданието на основната архитектура и функционалност.
- Тествана вътрешно от разработчиците или QA екипа.
- Има базов интерфейс и ограничена стабилност.

Примерни доказателства:

- Вътрешни тестови доклади
- Скриншоти/видео от интерфейса
- Release notes на Alpha версия
- Документирана архитектура

✓ TRL 5: Алфа версия – външно тестване (Предстои: Януари/Февруари 2026)

Изисквания:

- Същата Alpha версия е предоставена на ограничен брой външни потребители (напр. партньори, тестери).
- Събират се обратна връзка и бъгове от реална употреба.
- Тестовата среда е частично контролирана.

Примерни доказателства:

- Резултати от потребителски тестове (UX reports)
- Доклади от външни бета тестери

- Журнал на грешки (issue tracker)
- Споразумения с външни тестови партньори

Предал: Йордан Кралев
Дата: 15.09.2025

Приел: Светла Коева