Инфраструктура за създаване на чатбот с големи езикови модели и контекстно разширяване на инструкциите

Представяне на резултати от проекта
„Инфраструктура за фина настройка на предварително обучени големи езикови модели“

Йордан Кралев
Секция по компютърна лингвистика
Институт за български език, Българска академия на науките

29.05.2026

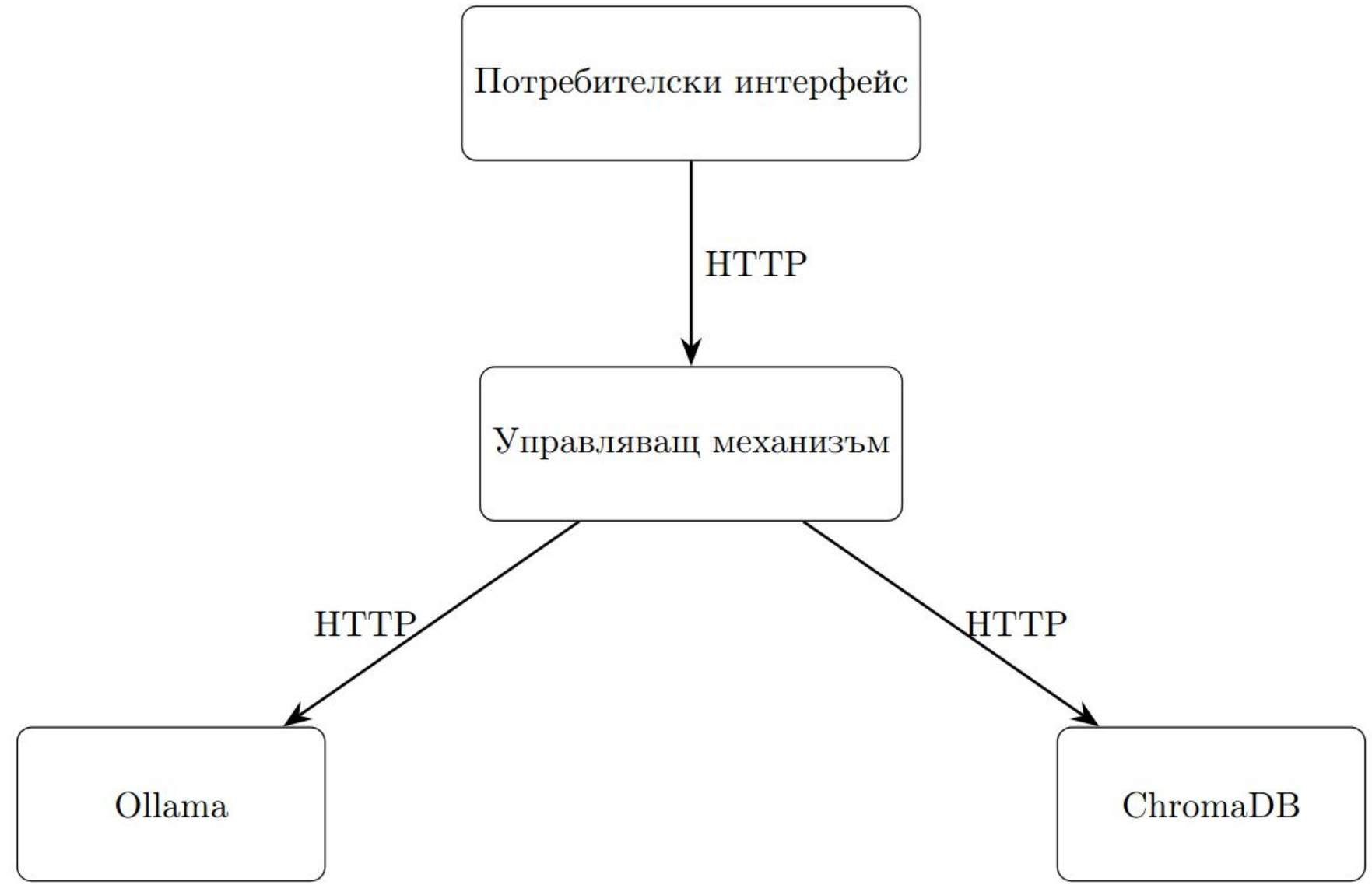
Описание на проблема

- Недостатъчна база от LLM решения за български език.
- Ограничена точност на моделите без външен контекст.
- Обобщаване на документи в тясно специализирана научно-образователна среда.
- Нужда от работеща архитектура с достъпен хардуер.

Налични решения

- LangChain
  - подходящ за разработка на големи езикови системи
- LlamaIndex
  - лесно захранване с разнородни документи
- Haystack
  - специализирано решение за RAG системи
- DSPy
- Pathway

Структура на системата



Векторна база данни

- Вграждане на речеви единици в метрично пространство
  - на ниво дума
  - на ниво изречение
  - на ниво контекст

$$D = (u_1, u_2, \ldots, u_n)$$

$$\varphi(u) \in \mathbb{R}^d \qquad \varphi(D) \in \mathbb{R}^{d \times n}$$

- Индексиране и търсене по критерий за семантична близост

$$\mathrm{Top}_k(q) = \mathrm{argmax}_{u,k} s(q, u)$$

$$s(u_i, u_j) = \frac{\varphi(u_i)^T \varphi(u_j)}{||\varphi(u_i) - \varphi(u_j)||}$$

$$s_d(u_i, u_j) = ||\varphi(u_i) - \varphi(u_j)||_2$$

Настройка на езикови модели

- Постановка
  - корпус на български език

$$\mathcal{D} = \{x^{(i)}\}_{i=1}^{N} \qquad x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_T^{(i)}) \qquad x_t^{(i)} \in \mathbb{R}^d$$

  - авторегресионен подход за генерация

$$x_t \sim p(z|x_{<t}, \theta)$$

- Критерий за оптимизация

$$p(x) = \prod_{t=1}^{T} p_\theta(x_t|x_{<t}, \theta)$$

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N}\sum_{t=1}^{T_i} \log p(x_t^{(i)}|x_{<t}^{(i)}, \theta)$$

- Подходи за редукция на реда на модела (LoRA)

$$W = W_0 + \Delta W, \quad \Delta W = BA$$

$$W_0 \in \mathbb{R}^{d\times d}, \quad B \in \mathbb{R}^{d\times r}, \quad A \in \mathbb{R}^{r\times d}$$

$$\mathcal{L}(A, B) = -\sum_{i=1}^{N}\sum_{t=1}^{T_i} \log p(x_t^{(i)}|x_{<t}^{(i)}, \theta_0, A, B)$$

Управляващ механизъм

- Потребителска заявка

$$\varphi(q) \in \mathbb{R}^{d \times n_q}$$

- База данни с документи

$$D = \bigcup_{i=1}^{N} D_i$$

- Семантично търсене и извличане на сегменти

$$R(q) = \mathrm{Top}_k(q, D)$$

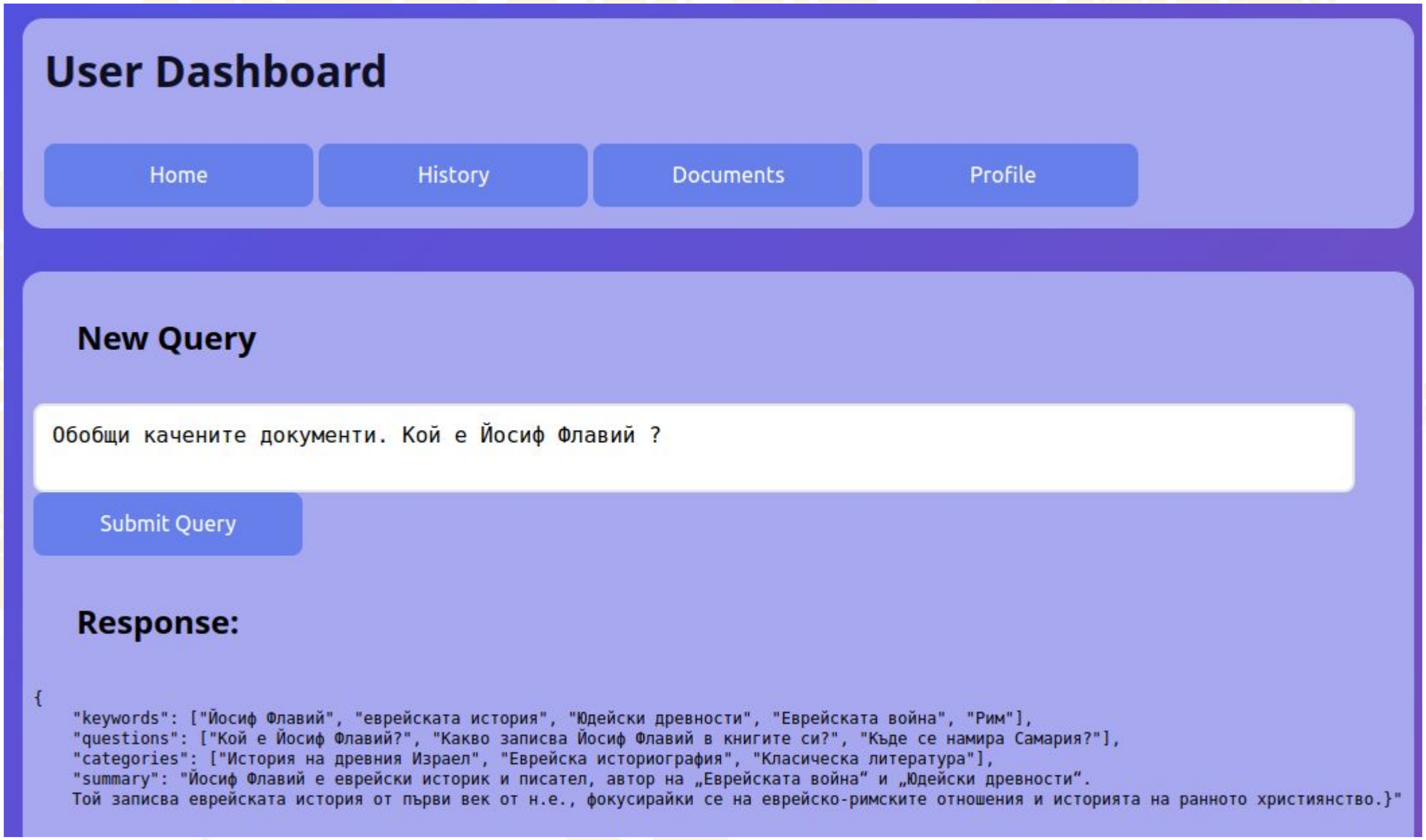
- Авторегресионно генериране на изходен израз

$$y_t \sim p(z|y_{<t}, R(q), q, p)$$

- Управляващи инструкции

$$y \in \mathcal{Y}_{target}$$

Резултати


User Dashboard
Home
History
Documents
Profile
New Query
Обобщи качените документи. Кой е Йосиф Флавий ?
Submit Query
Response:
{
"keywords": ["Йосиф Флавий", "еврейската история", "Юдейски древности", "Еврейската война", "Рим"],
"questions": ["Кой е Йосиф Флавий?", "Какво записва Йосиф Флавий в книгите си?", "Къде се намира Самария?"],
"categories": ["История на древния Израел", "Еврейска историография", "Класическа литература"],
"summary": "Йосиф Флавий е еврейски историк и писател, автор на „Еврейската война“ и „Юдейски древности“.
Той записва еврейската история от първи век от н.е., фокусирайки се на еврейско-римските отношения и историята на ранното християнство.}"

Заключение

- Реализирана е RAG система за български език с отворени средства.
- Демонстрирана е ефективност при обобщаване на документи с външен контекст.
- Потвърдена е приложимост при достъпен хардуер.
- Бъдещо развитие:
  - Системи за отговор на въпроси
  - Системи за класификация
  - Фина настройка на модели по тематични области

Благодарност

- Работата е част от проекта „Инфраструктура за фина настройка на предварително обучени големи езикови модели“ по договор номер ПВУ – 55 от 12.12.2024 /BG-RRP-2.017-0030-C01/.

Funded by the European Union
NextGenerationEU

National Recovery and Resilience Plan of the Republic of Bulgaria

- http://ifgpt.dcl.bas.bg