Генериране на тестови въпроси с множествен избор върху зададен текст от голям езиков модел-система Lecturio и примери за локално и онлайн приложение с модели BGGPT и Мистрал2

Пенчо Тончев, Сергей Илиев, Кирил Статев, Добромир Димитров

Generating multiple-choice test questions on a given text from a large language model in Lecturio system and examples of local and online applications with BGGPT and Mistral2 models.

Pencho Tonchev, Sergey Iliev, Kiril Statev, Dobromir Dimitrov

Abstract:
The use of Artificial Intelligence in education includes a variety of applications, among which the simplest to implement is the creation of test questions. The report discusses the realization of this task at MU-Pleven, as an important part of the digitalization of the educational process. The large language models BGGPT and Mistral7B, with local installation with Koboldcpp and LMStudio, as well as the implementation of a virtual server and LangChain application are used for demonstration. The new functions for generating MCQ from AI in Lecutrio are also demonstrated.
Keywords: Retrieval-Augmented Generation, Large Language Models, Mistral2, BgGPT,

For contacts: Assoc. prof. Dr. Pencho Tonchev, PhD, Medical University – Pleven pencho.tonchev@mu-pleven.bg

ВЪВЕДЕНИЕ

Използването на Изкуствен интелект в обучението включва разнообразни приложения, между които най-просто за реализация е създаването на тестови въпроси. Генеративният Изкуствен Интелект може да се използва за обобщаване на текстове, генериране на въпроси, сценарии, симулиране на пациенти и пр. Какво е Retrieval-Augmented Generation? Генериране с допълнено извличане (RAG) е процес на оптимизиране на изхода на голям езиков модел, така че той препраща към авторитетна база знания извън своите източници на данни за обучение, преди да генерира отговор. Големите езикови модели (LLM) се обучават върху огромни обеми от данни и използват милиарди параметри за генериране на оригинален резултат за задачи като отговаряне на въпроси, превод на езици и довършване на изречения. RAG разширява вече мощните възможности на LLMs до специфични домейни или вътрешната база знания на организацията, всичко това без необходимост от повторно обучение на модела. Това е рентабилен подход за подобряване на резултатите от LLM, така че да остане подходящ, точен и полезен в различни контексти. Как работи Retrieval-Augmented Generation? Без RAG, LLM взема задачата от потребителя и създава отговор въз основа на информацията, върху която е бил обучен или това, което вече знае. С RAG се въвежда компонент за извличане на информация, който използва входа на потребителя, за да изтегли първо информация от нов източник на данни. Техника, наречена вграждане на езикови модели(embedding models) , преобразува данни в числени представяния и ги съхранява във векторна база данни (напр. QDrant,

Chroma, Pinecone).Този процес създава библиотека със знания, която генеративните AI модели могат да разберат. Следващата стъпка е търсене по релевантност. Потребителската заявка се преобразува във векторно представяне и се съпоставя с векторните бази данни. Най-подходящият отговор се установява с помощта на математически векторни изчисления и представяния (cosine similarity, etc).[1][2]

Големите комерсиални модели като ChatGPT4, Claude2, Gemmini позволяват вграждане на собствени файлове с информация и дават отлични отговори на много задачи. На разположение са и модели за локална инсталация и вграждане в web приложения. Те имат предимството да са „безплатни“ и с „отворен лиценз“.

Тестови въпроси могат да се генерират от онлайн приложения на RAG концепцията -след съответния абонамент и качване на файла с учебното съдържание системата генерира разнообразни въпроси.

В доклада се обсъжда реализацията на тази задача в МУ-Плевен, като важна част от дигитализацията на учебния процес.

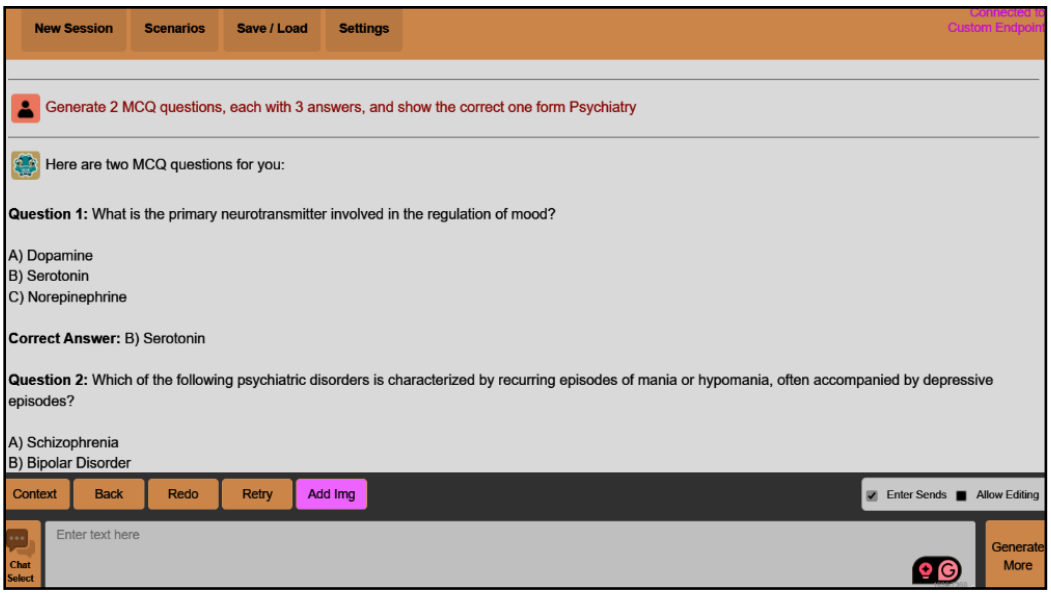
ИЗЛОЖЕНИЕ

Цел: да се направи обзор на възможостите за генериране на тестови въпроси с множествен избор с един верен отговор(MCQ), върху конкретен текст, да се изпробва реализация на тази задача на персонален компютър, на специално създаден сървър с web приложение и в платформата Lecturio

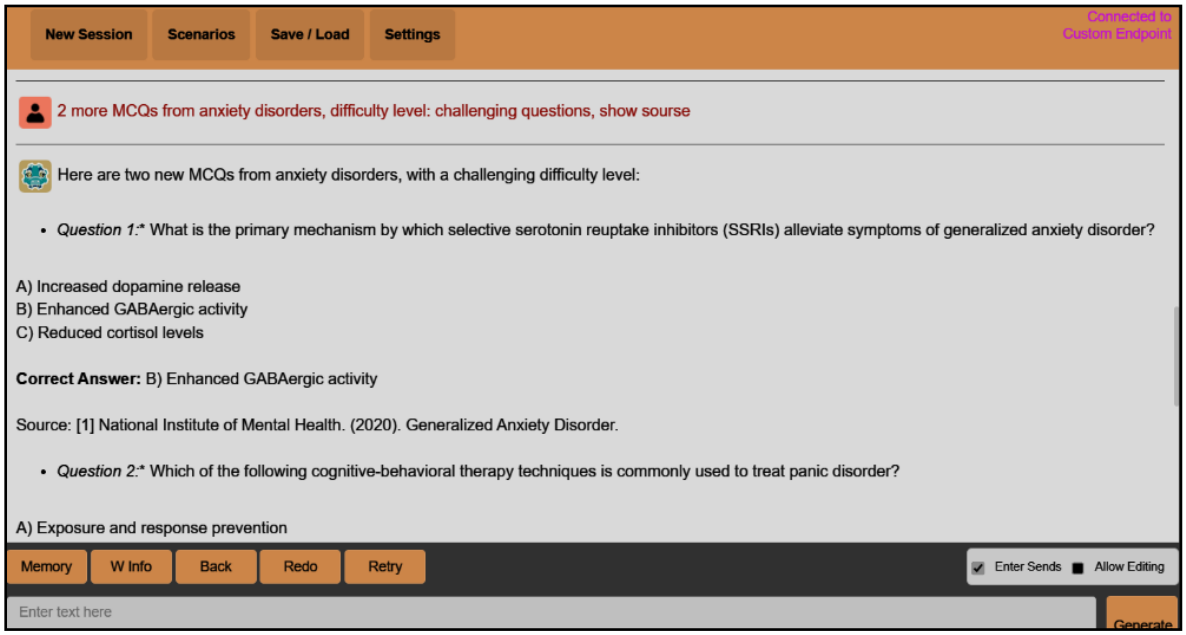
Методи: За демонстрация са използвани големите езикови модели BGGPT[3][4], Llama3[5] и Mistral7B[6] с локална инсталация с Koboldcpp[7], LMStudio[8] и GPT4All[9], Демонстрирани са и новите функции за генериране на MCQ от AI в Lecutrio[10]. За примерите са ползван учебниците по Психиатрия на Саксби Придмор-елактронно издание на МУ-Варна.[11] и на Burns[12]

Резултати:

1. Ползвахме Koboldcpp-NoCuda, на лаптоп с процесор Intel i5 и 32 GB Ram, със зареден модел Meta-Llama-3-8B-Instruct-Q4_K_M и без зареден файл на учебник по психиатрия.

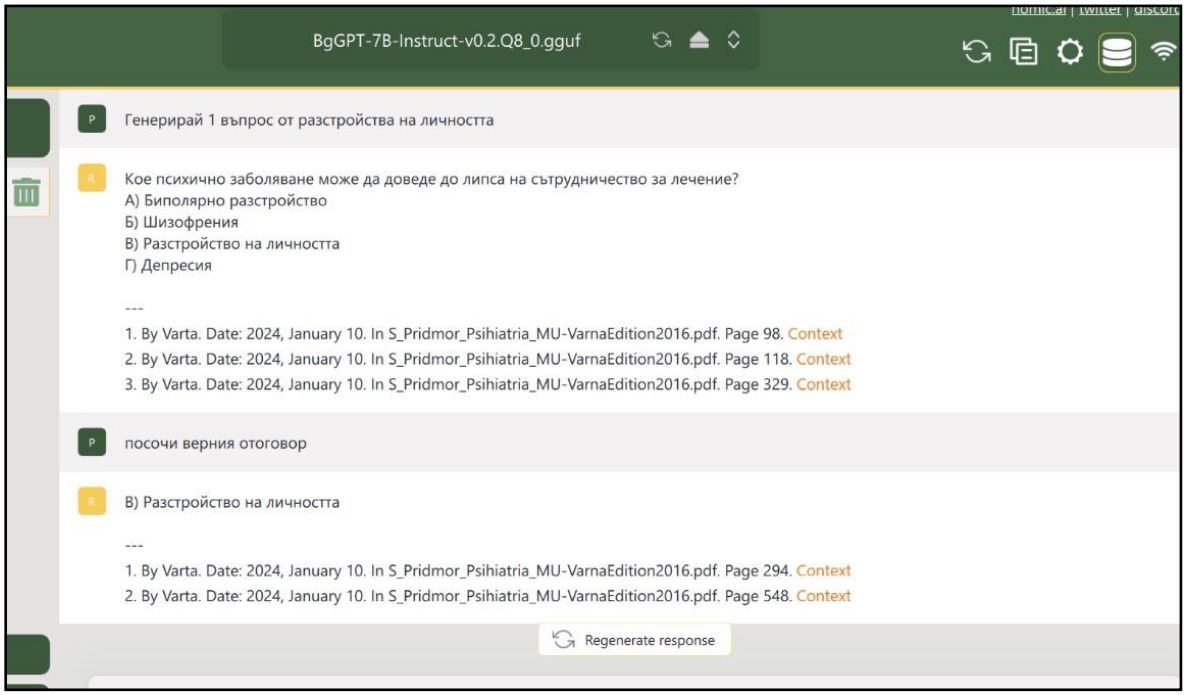


Фигура 1: MCQ генерирано при заявка на английски език. Не е зададено изискване за трудност на въпроса, или посочване на източник. Llama3 модела е отлично обучен по психиатрия



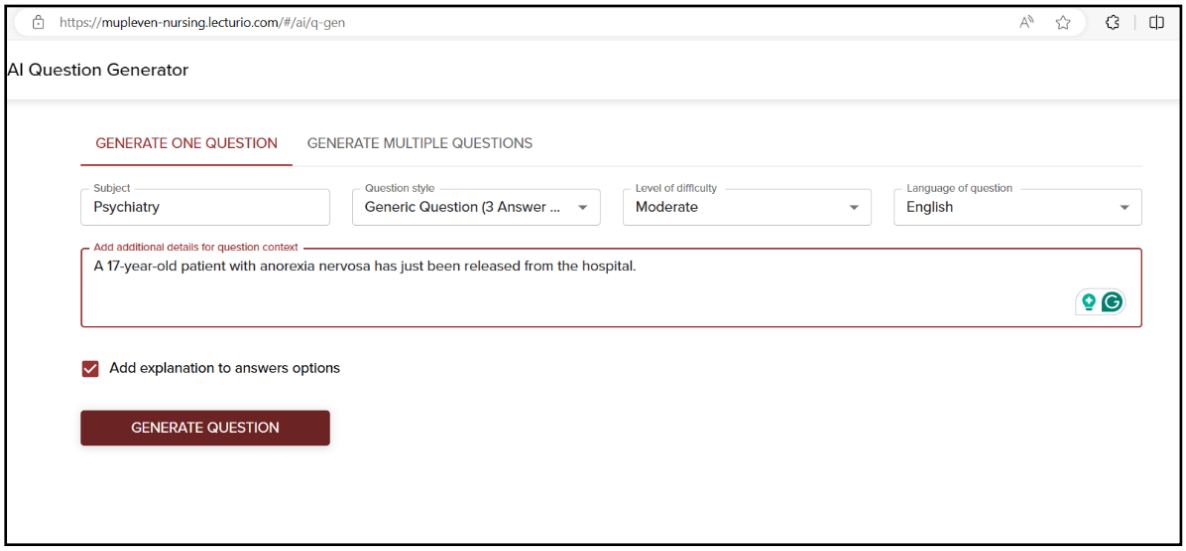
Фигура 2: MCQ генерирано след зададено изискване за трудност на въпроса и посочване на източник.

2. Ползвахме GPT4All със зареден модел BgGPT-7B-Instruct-v0.2.Q8_0 и зареден учебника на Придмор на български език. Това е реализация на RAG концепцията. В промпта указахме да задава MCQ въпроси, като посочва верния отговор и източника.

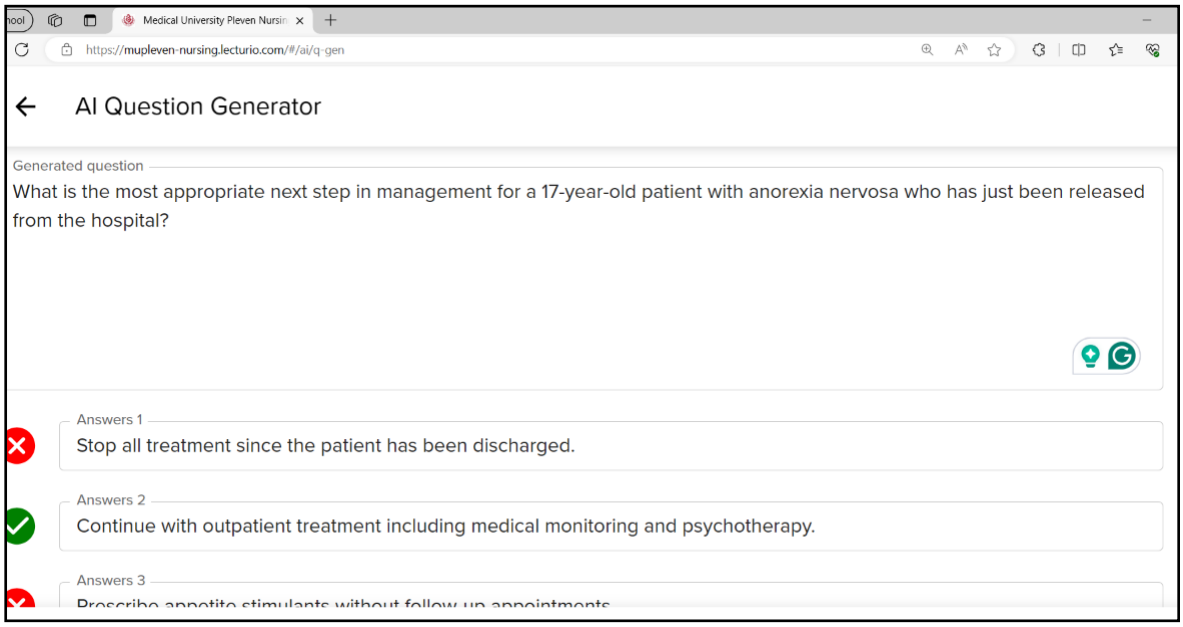


Фигура 3. MCQ генерирано от BGGPT модел, с извличане и включване от локален файл.

3. Ползвахме AI генератора на тестови въпроси, вграден в системата Lecturio –Nursing на МУ-Плевен. Има възможност да се избере тежест на въпроса, предметна област и допълнителни указания



Фигура 4. "Настройки" на генератора на въпроси.



Фигура 5 Генериран въпрос с посочване на верния отговор и обяснения.

ЗАКЛЮЧЕНИЕ

Възможностите на Големите езикови модели (LLM) да разбират задачи на различни езици и да генерират смислен текст се увеличават с всеки ден. Наличието на LLM с отворен достъп и разработените готови приложения за локално инсталиране на тези модели правят възможна реализацията на RAG концепцията (Генериране с допълнено извличане), както за обобщаване на документи, така и за генериране на тестови въпроси. Към момента модела Llama3 и BGGPT и локалното приложение GPT4All са най-подходящи за генериране на тестове върху представен материал на български език.

ЛИТЕРАТУРА

[1] What is RAG? - Retrieval-Augmented Generation Explained - AWS n.d. https://aws.amazon.com/what-is/retrieval-augmented-generation/ (accessed May 11, 2024).

[2] Mastering RAG: Choosing the Perfect Vector Database - Galileo n.d. https://www.rungalileo.io/blog/mastering-rag-choosing-the-perfect-vector-database

(accessed May 11, 2024).

[3] BgGPT n.d. https://bggpt.ai/ (accessed May 11, 2024).

[4] INSAIT-Institute/BgGPT-7B-Instruct-v0.2-GGUF · Hugging Face n.d. https://huggingface.co/INSAIT-Institute/BgGPT-7B-Instruct-v0.2-GGUF (accessed May 11, 2024).

[5] lmstudio-community/Meta-Llama-3-8B-Instruct-GGUF · Hugging Face n.d. https://huggingface.co/lmstudio-community/Meta-Llama-3-8B-Instruct-GGUF (accessed May 11, 2024).

[6] TheBloke/Mistral-7B-Instruct-v0.2-GGUF · Hugging Face n.d. https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-GGUF (accessed May 11, 2024).

[7] GitHub - LostRuins/koboldcpp: A simple one-file way to run various GGML and GGUF models with KoboldAI's UI n.d. https://github.com/LostRuins/koboldcpp (accessed May 11, 2024).

[8] LM Studio - Discover, download, and run local LLMs n.d. https://lmstudio.ai/ (accessed May 11, 2024).

[9] GPT4All n.d. https://gpt4all.io/index.html (accessed May 11, 2024).

[10] Lecturio Nursing -Medical University Pleven n.d. https://mupleven-nursing.lecturio.com/ (accessed May 10, 2024).

[11] Придмор С. Психиатрия. Варна: МУ-Варна; 2016.

[12] Burns T. Psychiatry: A Very Short Introduction 2018. https://doi.org/10.1093/actrade/9780198826200.001.0001.