FELchat benchmark otázky

PVTY1 - Práce v týmu

Kryštof Vilímek

29. května 2025

1 Úvod

Účel tohoto dokumentu je shromáždit relevantní informace o benchmarkovacích otázkách projektu FELchat. Vše od postupu vytvoření otázek až po statistická data lze najít v tomto dokumentu. Sada obsahuje 100 otázek, které se vstahují k jedinému dokumentu *Study and Examination Rules for Students at CTU*. Tyto otázky byly vytvořeny za účelem testování RAG implementace studentského chatbota.

2 Způsob vytvoření otázek

Prvotní implementace chatbota je v angličtině pro odstranění potenciální překážky, kterou je překlad y angličtiny do češtiny. Proto je dataset v anglickém jazyce.

Pro generování otázek byly použity dva přístupy. Jedním z přístupů bylo použití umělé neuronové sítě - *DeepSeek R1* LLM - k vygenerování otázek. Druhý přístup byl použití biologické neuronové sítě - autora tohoto dokumentu - k vygenerování otázek. Většina otázek byla vygenerována LLM *DeepSeek R1* a zkontrolována autorem.

2.1 LLM

Hlavním důvodem zvolení LLM *DeepSeek R1* byl fakt, že umožňuje velké množství výstupu bez jakékoli peněžní investice. Otázky byly generovány iteračně. Největším problémem bylo donutit LLM, aby opravdu vygeneroval 100 otázek a nezkracoval výstup. Níže je v bodech uvedena sekvence promptů použitých při komunikace s LLM. Sekvence promptů je reprezentativní zjednodušení skutečné konverzace ze strany uživatele a zachytává překážky, které se při generování vyskytly.

- *"Hi, please help me with generating 100 benchmark questions for our RAG implementation of a university chatbot called FELchat. Analyze the imported PDF file and generate the 100 question-answer pairs so that: * questions 1-50 are answered with Yes or No and the Yes/No answers are split evenly * questions 51-100 are answered with a sentence. Please put all of the questions and answers into a json file with the following format*

  id: question: answer: page:"
- "*Please continue and finish the list by making additional question-answer pairs 80-100.*"
- "*Please continue. Finish th elist with question-answer pairs 90-100*"
- "*What is the page number of question 79?*"
- "*Please add the relevant articles which answer the questions.*"
- "*These are my current benchmark questions... Please add the article identifier to all of them similarly to how you did it previously. Please also update the pages when they are written as zero.*"
- "*Now please make a list of questions 81 to 85 in the same format for the following questions... if the information cannot be found in the document, indicate it in the "article"section*

  "Where can the dean get their lunch?"

  "How many students can the university accept?"

"How many toilets does the university have?"

"Can i utilize a CTU printer for free?"

"Does the university offer employment to students?""

- "*please make entries in the same format to the following questions...*

  "How many times can i retake a failed exam?"

  "What happens if I fail a subject?""

- "*This is my current benchmark set. Please replace the "page"entries by an out_of_scope flag which is True if the question cannot be explicitly answered by the document*"

- "*Thank You.*"

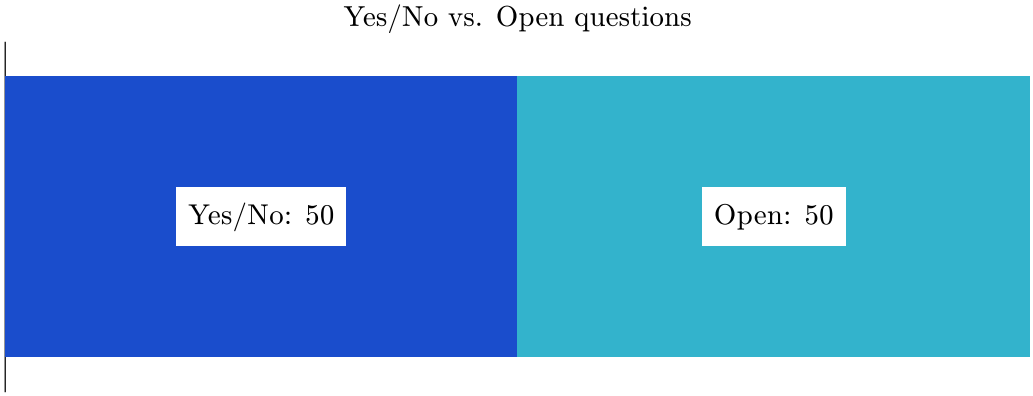
Použití LLM se obešlo bez větších chyb. Model měl sem tam problém s identifikací stránek, relevantní články identifikoval bezchybně. Pouze 1-2 vygenerované otázky byly ve finální verzi upraveny autorem.

2.2 Člověk

Autor zkontroloval vygenerované otázky, odpovědi, články a relevanci. Několik otázek sám vytvořil, aby v datasetu byly zastoupeny i otázky, které by studenti reálně zadali do chatbota. Jedná se o otázky 81 - 100. Některé z těchto otázek byly navrženy tak, aby je nebylo možno odpovědět na základě poskytnutého dokumentu. Takové otázky mají flag "out_of_scope: true"

3 Statistiky datasetu

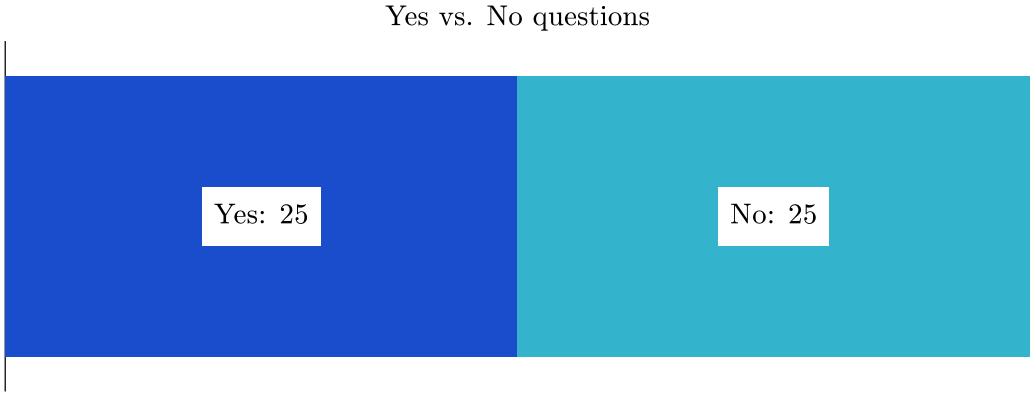
Jelikož je v datasetu 100 otázek, metrika ohledně počtu zastoupení je ekvivalentní s procentuální metrikou. Dataset je rozdělený na uzavřené otázky (ano/ne) a otevřené otázky (odpověď celou větou).



Obrázek 1: Počet uzavřených a otevřených otázek.

3.1 Uzavřené otázky

Uzavřené otázky jsou rozdělené dle odpovědí, Ano nebo Ne.

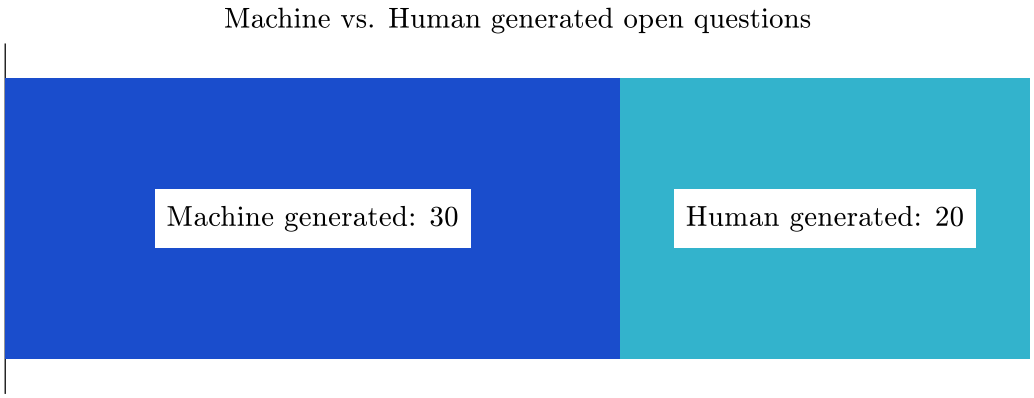


Obrázek 2: Počet otázek s příslušnou odpovědí Ano/Ne.

Všechny uzavřené otázky byly vygenerovány LLM a jejich odpovědi jsou dohledatelné v dokumentu.

3.2 Otevřené otázky

Otevřené otázky se dělí na strojem-generované a člověkem generované. Po přičtení uzavřených otázek činí celkový



Obrázek 3: Počet strojem/člověkem-generovaných otázek.

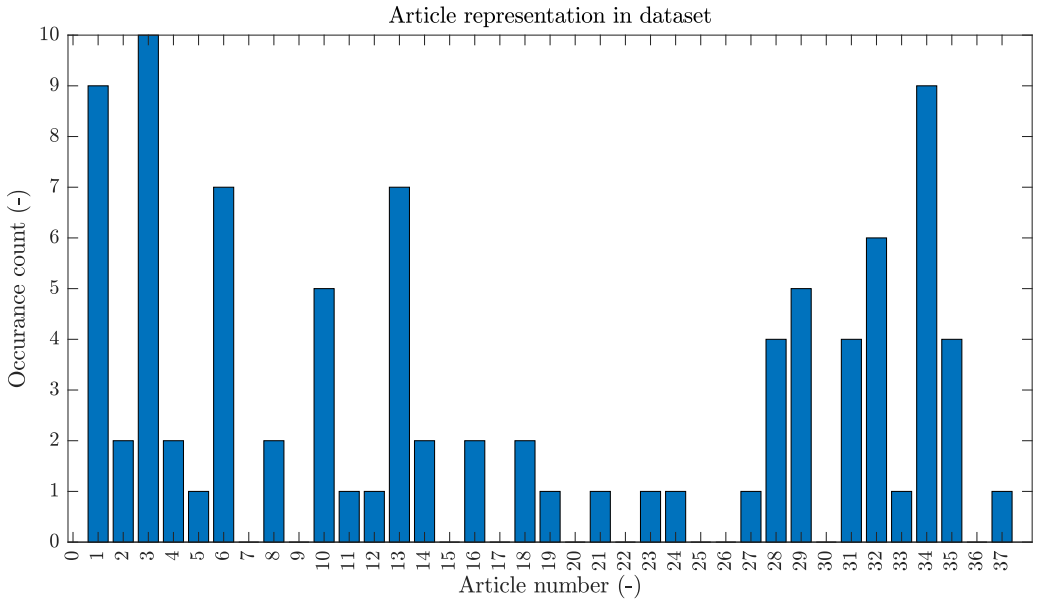
počet strojem-generovaných otázek 80.

Do datasetu byly vloženy otázky, které nelzou zodpovědět na základě poskytnutého dokumentu ("Out of scope"). Po přičtení uzavřených otázek činí celkový počet zodpověditelných otázek 92.

Obrázek 4: Počet zodpověditelných otázek.

3.3 Zastoupení článků

Zajímavou metrikou v kontextu tohoto datasetu je počet zastoupení jednotlivých článků v otázkách.



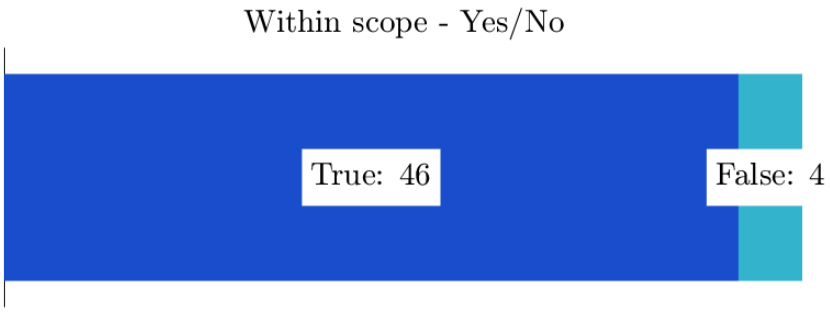
Obrázek 5: Zastoupení jednotilvých článků v datasetu.

4 Vyhodnocení člověkem

Naší implementací byly vygenerovány odpovědi na všech 100 testovacích otázek. Tyto odpovědi byly zkontrolovány člověkem a vyhodnoceny jako správné (true) nebo nesprávné (false).

4.1 Uzavřené otázky

Uzavřené otázky byly správně zodpovězené ve 46 případech z 50, úspěšnost 92%.



Obrázek 6: Úspěšnost odpovědí na uzavřené otázky.

4.2 Otevřené otázky

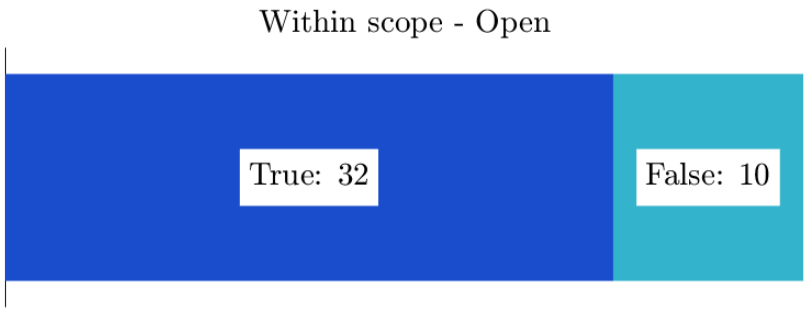
Jak již bylo zmíněno, otevřené otázky jsou rozděleny na zodpověditelné (within scope) a nezodpověditelné (out of scope).

4.2.1 Zodpověditelné

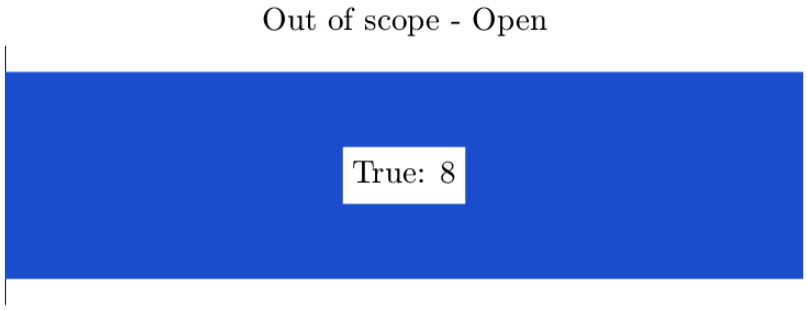
Zodpověditelné otevřené otázky byly správně zodpovězené v 32 ze 42 případů, úspěšnost 76%.

4.2.2 Nezodpověditelné

Nezodpověditelné otevřené otázky byly správně zodpovězeny 8 z 8 případů, úspěšnost 100%.



Obrázek 7: Úspěšnost odpovědí na zodpověditelné otevřené otázky.



Obrázek 8: Úspěšnost odpovědí na nezodpověditelné otevřené otázky.

5 Závěr

Byla vytvořena benchmarkovací sada 100 otázek, jejíž cílem je rozsáhle otestovat opdovídání na otázky související s dokumentem *Study and Examination Rules for Students at CTU*. Odpovědi chatbota na otázky byly následně vyhodnoceny jako správné/nesprávné. Chatbot se projevil jako funkční s pár nedostatky, jejichž vyřešení by znamenalo větší procento úspěšnosti a lepší user experience. Aktuální implementace chatbota k datu 29. května 2025 má problém se skládáním kontextu ze vzdálených míst v dokumentu, což negativně ovlivňuje správnost či kompletnost některých odpovědí. Dalším nedostatkem je vracení identifikace článku (např. *Article 5(c)*) uživateli v opdovědi. Implementace tohoto vracení je bohužel závislá na struktuře dokumentu. Časová náročnost vyhodnocení odpovědi ("rag_total_time_sec")se pohybovala mezi 3 až 6 sekundami, když byl chatbot zprovozněn na ČVUT GPU serveru. Chatbot je důkazem potenciálu nasazení podobné implementace jako oficiálního chatbota FEL ČVUT. Podnětem pro další práci může být vylepšení přesnosti a zprovoznění chatbota pro testovací skupinu studentů/zaměstnanců FEL ČVUT.