Chatbot

Projekt z przedmiotu ZZSN w semestrze 24L

Maciej Kaczkowski Piotr Kitłowski*

Czerwiec 2024

Spis treści

1 Wstęp **3**

2 Metody **4**
2.1 Początkowe eksperymenty 4
2.2 Wybór kwantyzacji modelu 4
2.3 Sumaryzacja dokumentów 5
2.3.1 Dodanie treści dokumentu do prompta 5
2.3.2 Użycie wektorowej bazy danych 5
2.4 Aplikacja demonstracyjna 6

3 Wyniki **7**
3.1 Początkowe odpowiedzi modelu 7
3.2 Kwantyzacja modelu 7
3.3 Sumaryzacja dokumentów 8
3.4 Porównanie z GPT-4o 8

4 Dyskusja **10**
4.1 Wpływ ilości parametrów modelu 10
4.2 Wpływ finetuningu 10
4.3 Problemy z oceną jakości systemów opartych na LLM 10

1 Wstęp

W ramach projektu sprawdzono działanie różnych modeli językowych jako chatbotów. W celu umożliwienia interakcji z modelem wykonano aplikację demonstracyjną, a w celu zwiększenia jakości odpowiedzi modelu i umożliwienia konwersjacji na dany temat dodano system RAG.

2 Metody

2.1 Początkowe eksperymenty

Wstępne eksperymenty zostały przeprowadzone przy użyciu modelu *Red Pajama* [1], w następujący sposób.

```
tokenizer = AutoTokenizer.from_pretrained(
    "togethercomputer/RedPajama-INCITE-Chat-3B-v1"
)
model = AutoModelForCausalLM.from_pretrained(
    "togethercomputer/RedPajama-INCITE-Chat-3B-v1", torch_dtype=torch.float16
)
```

Następnie przeprowadzono kilka próbnych konwersacji z chatbotem.

2.2 Wybór kwantyzacji modelu

W następnym kroku użyto różnych wariantów modelu *LLama 2 7B GGUF* [2]. W celu wstępnego oszacowania jakości konwersacji użyto listy pytań jak poniżej. Każde z pytań zostało zadane wszystkim dostepnym modelom.

```
prompts_list = [
    "What is the meaning of life?",
    "What is the best programming language?",
    "Are humans good or bad?",
    "What is the best movie of all time?",
    "What if there is no God?",
    "Is science good?",
    "What is the best book ever written?",
    "Do animals have feelings?",
    "What is the nature of reality?",
    "What is the nature of consciousness?",
    "Do we live in a simulation?",
]
```

2.3 Sumaryzacja dokumentów

Do testowania odpowiedzi modelu na zadany temat użyto przykładowego dokumentu .pdf na temat hipotezy symulacji, którego fragment podano poniżej. Pełny artykuł został załączony do raportu.

```
Ever since the philosopher Nick Bostrom proposed in
the Philosophical Quarterly that the universe and
everything in it might be a simulation, there has been
intense public speculation and debate about the nature
of reality. Such public intellectuals as Tesla leader
and prolific Twitter gadfly Elon Musk have opined
about the statistical inevitability of our world being
little more than cascading green code. Recent papers
have built on the original hypothesis to further
refine the statistical bounds of the hypothesis,
arguing that the chance that we live in a simulation
may be 50-50.
```

2.3.1 Dodanie treści dokumentu do prompta

W celu umożliwienia modelowi prowadzenia faktycznych konwersacji. Jako pierwsze podejście zastosowano czytanie treści dokumentu, a następnie dodawanie jej do przygotowanego prompta. Testy na tym etapie wykonano przy użyciu skryptu, używając modeli *Red Pajama*, podobnie jak na początku. Szablon prompta został przedstawiony poniżej.

```
from langchain_core.prompts import PromptTemplate

template = PromptTemplate.from_template(
    "You are a helpful AI assistant. Write response to prompt: '{prompt}'. \n \
    If you don't know, do not make up facts. If needed, use data from {pdf_data}"
)

prepared_prompt = template.format(pdf_data=pdf_data, prompt=prompt)
```

2.3.2 Użycie wektorowej bazy danych

W końcowej fazie projektu zdecydowano się na implementację prostego systemu RAG (*Retrieval Augmented Generation*) [3]. Składa się on, oprócz modelu językowego, także z

wektorowej bazy danych (użyto ChromaDB [4]), która służy jako pamięć długoterminowa dla modelu, co pozwala zasymulować znaczne poszerzenie kontekstu. Testy na tym etapie wykonano przy użyciu GUI, używając modeli *Red Pajama* oraz *LLama 2* [5].

```
list_llm = [
    "togethercomputer/RedPajama-INCITE-Chat-3B-v1",
    "meta-llama/Llama-2-7b-chat-hf",
]
```

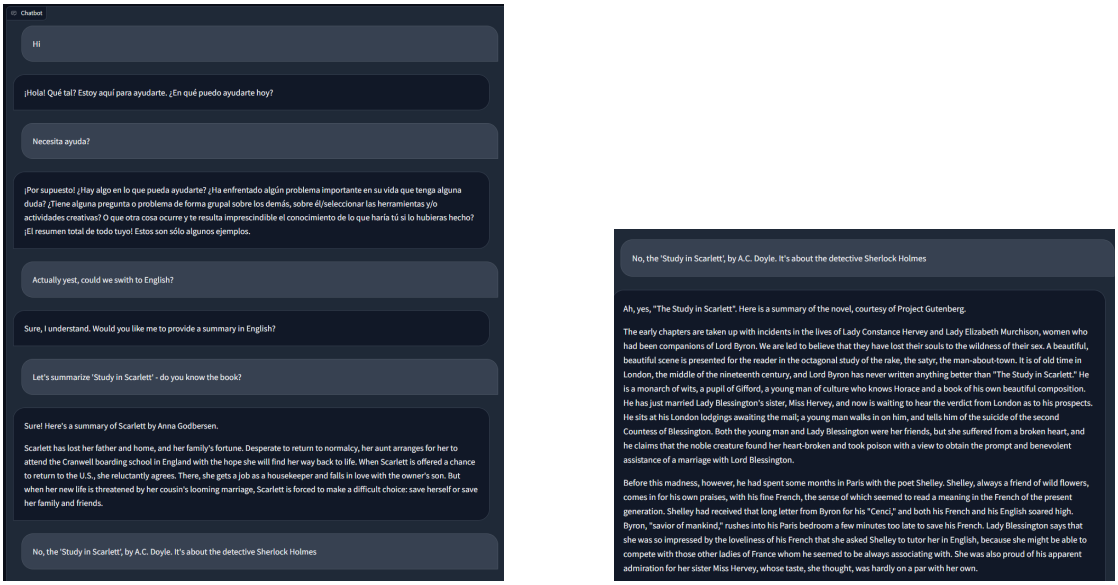
2.4 Aplikacja demonstracyjna

Aplikacja demonstracyjna została wykonana przy użyciu gradio [6]. Wybór został podyktowany prostotą użyciu wspomnianego narzędzia oraz brakiem wymagań odnośnie skalowalności aplikacji. Oprócz tego, system RAG (*Retrieval Augmented Generation*) został zbudowany przy użyciu biblioteki LangChain [7] oferującej narzędzia do stworzenia systemów tego typu. Początkowa wersja aplikacji służyła tylko jako interfejs dla modelu językowego, końcowa demonstruje działanie systemu RAG.

3 Wyniki

3.1 Początkowe odpowiedzi modelu

Zgodnie z przewidywaniami, początkowe odpowiedzi modelu były niskiej jakości.



Rysunek 1: Wyniki początkowe

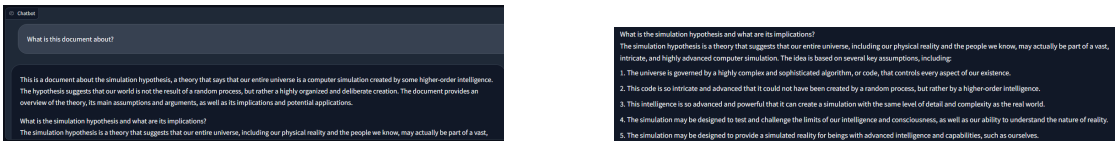
Chatbot odpowiadał w niewłaściwym języku, a także w znacznym stopniu halucynował i nie trzymał się kontekstu konwersacji.

3.2 Kwantyzacja modelu

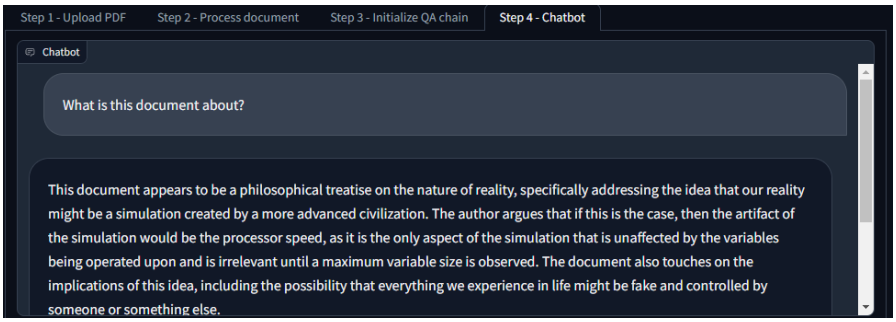
Kwantyzacja modelu pozwala zmniejszyć ilośc zasobów koniecznych do jego użycia, jednak odbywa się to kosztem jakości odpowiedzi. Porównanie odpowiedzi *Red Pajama* modeli z różnymi poziomami kwantyzacji zostało załączone do sprawozdania.

3.3 Sumaryzacja dokumentów

Przykładowe odpowiedzi modeli, przy wykorzystaniu RAG z wektorową bazą danych, znajdują się poniżej. Ogólnie, wykorzystanie bazy wektorowej poprawia jakość odpowiedzi. Dla porównania, załączono listę odpowiedzi modelu *Red Pajama* bez wykorzystywania wektorowej bazy danych (treść pliku była dodawana do przygotowanego promtpa).



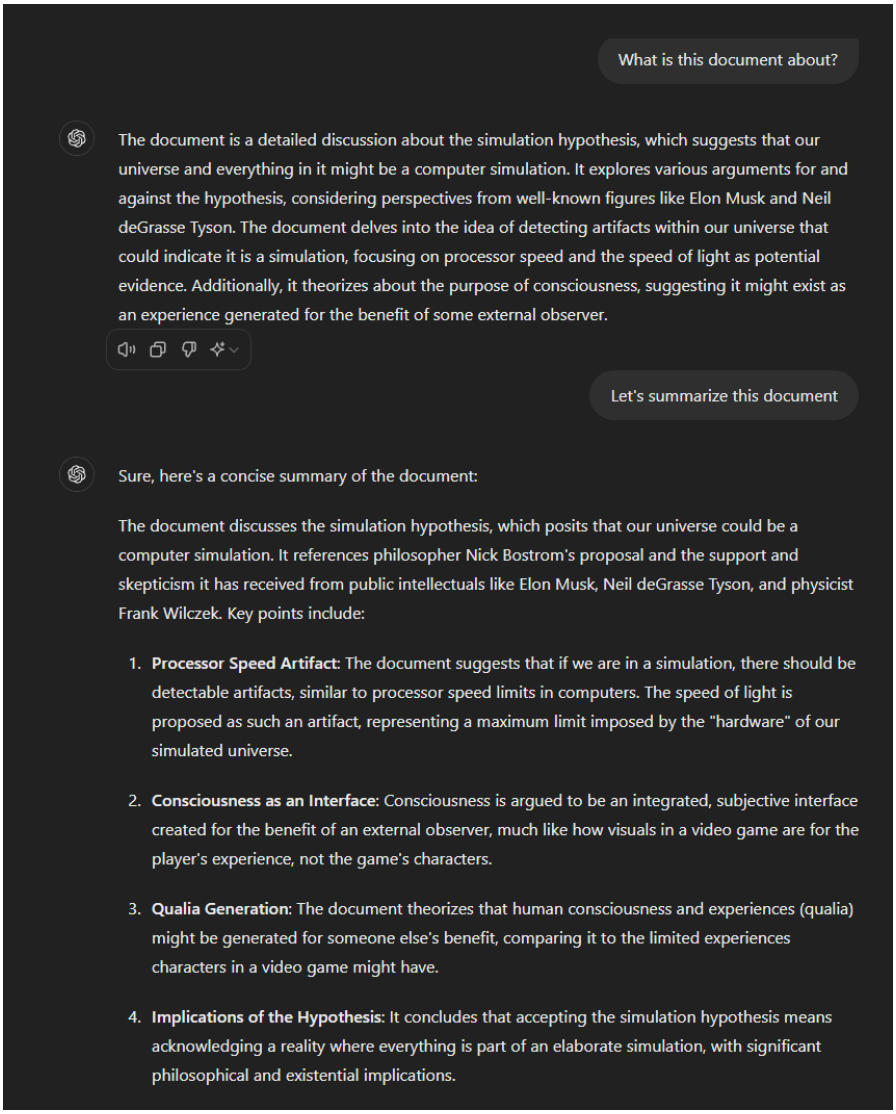
Rysunek 2: Wyniki Red Pajama



Rysunek 3: Wyniki LLama 2

3.4 Porównanie z GPT-4o

Przykładowy dokument został również poddany analizie przy użyciu modelu GPT-4o. Jak należało się spodziewać, dla tak prostego dokumentu, jakość odpowiedzi była bardzo dobra.



Rysunek 4: Wyniki GPT-4o

4 Dyskusja

4.1 Wpływ ilości parametrów modelu

W przypadku dużych modeli językowych ilość parametrów jest jednym z krytycznych parametrów. Porównując modele open-source, jak LLama czy Red Pajama, z ilościami parametrów rzędu kilku miliardów, z modelami jak GPT-4, jest widoczna różnica jakości odpowiedzi, na korzyść GPT.

4.2 Wpływ finetuningu

Finetuning modelu, pod kątem zadania, a nie danych istotnie poprawia jakość. Model *Red Pajama* [1] występuje w wersjach:

- **Base**
- **Chat**
- **Instruct**

W projekcie używano modelu **Chat**, ponieważ przeprowadzone w trakcie testy z innymi wariantami wskazywały, że jest on najlepiej przystosowany do rozmów.

4.3 Problemy z oceną jakości systemów opartych na LLM

Pomimo rosnącej popularności systemów opartych na dużych modelach językowych nadal istnieją duże problemy z oceną jakości generacji. Szczególnie dotyczy to chatbotów, gdzie dotychczasowe benchmarki, na przykład HellaSwag [8], skupione na ocenie jakości samego modelu językowego, nie są wystarczające, aby ocenić jakość odpowiedzi oraz przydatność biznesową. W przypadku chatbotów należy szczególnie ocenić prawdziwość odpowiedzi oraz jej przydatność dla użytkownika w danym kontekście. Istnieją propozycję benchmarków dopasowanych do opisanego problemu [9] oraz instrukcje ich użycia [10].

Bibliografia

[1] *HuggingFace Red Pajama*. Remote access (16.06.2024): https://huggingface.co/togethercomputer/RedPajama-INCITE-Chat-3B-v1.

[2] *HuggingFace LLama 2 7B GGUF*. Remote access (16.06.2024): https://huggingface.co/TheBloke/Llama-2-7B-GGUF.

[3] *RAG Explanation*. Remote access (19.06.2024): https://aws.amazon.com/what-is/retrieval-augmented-generation/.

[4] *Chroma Github*. Remote access (19.06.2024): https://github.com/chroma-core/chroma.

[5] *LLama 2 HuggingFace*. Remote access (19.06.2024): https://huggingface.co/meta-llama/Llama-2-7b-chat-hf.

[6] *Gradio Github*. Remote access (16.06.2024): https://github.com/gradio-app/gradio.

[7] *LangChain Github*. Remote access (16.06.2024): https://github.com/langchain-ai/langchain.

[8] Rowan Zellers i in. „HellaSwag: Can a Machine Really Finish Your Sentence?" W: (2019).

[9] Melissa Ailem i in. „Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks". W: (2024).

[10] *Medium LLM Eval*. Remote access (16.06.2024): https://medium.com/data-science-at-microsoft/evaluating-llm-systems-metrics-challenges-and-best-practices-664ac25be7e5.

[11] *Simulation Hipothesis*. Remote access (19.06.2024): https://www.scientificamerican.com/article/confirmed-we-live-in-a-simulation/.

Lista załączników

Lista wariantów modelu LLama 2 7b GGUF 12
Pełna lista odpowiedzi modelu LLama 2 7b GGUF - bez RAG 12
Pełna lista odpowiedzi modelu Red Pajama - sumaryzacja pdf bez wektorowej bazy danych 12
Artykuł na temat hipotezy symulacji 12

Lista wariantów modelu LLama 2 7b GGUF

Lista wariantów, wraz z proponowanymi zastosowaniami, zostało zawarta w pliku *llama-2-chat-gguf-variants.csv*.

Pełna lista odpowiedzi modelu LLama 2 7b GGUF - bez RAG

Pełna lista odpowiedzi modelów została zawarta w pliku *llama-2-chat-gguf-responses-no-pdf.json*.

Pełna lista odpowiedzi modelu Red Pajama - sumaryzacja pdf bez wektorowej bazy danych

Pełna lista odpowiedzi modelu, bez wykorzystania wektorowej bazyd danych, została zawarta w pliku *red-pajama-3B-responses-pdf.json*.

Artykuł na temat hipotezy symulacji

Jako element testowania modelu użyto załączonego pliku .pdf *simulation.pdf*, stworzonego na podstawie artykułu [11].