Chatbot

Projekt z przedmiotu ZZSN w semestrze 24L

Maciej Kaczkowski Piotr Kitłowski *

Czerwiec 2024

Spis treści

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

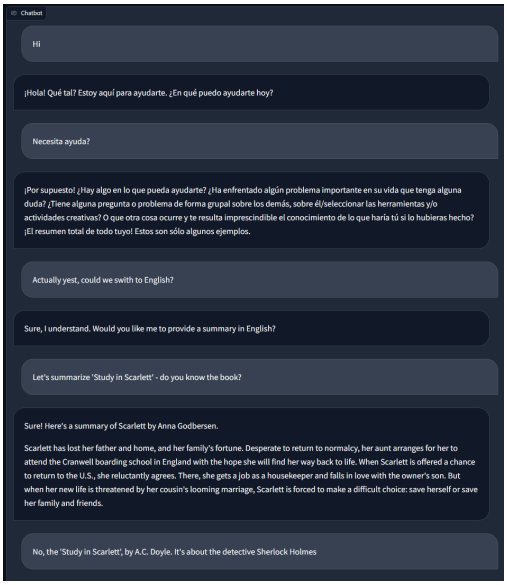
2.4 Aplikacja demonstracyjna

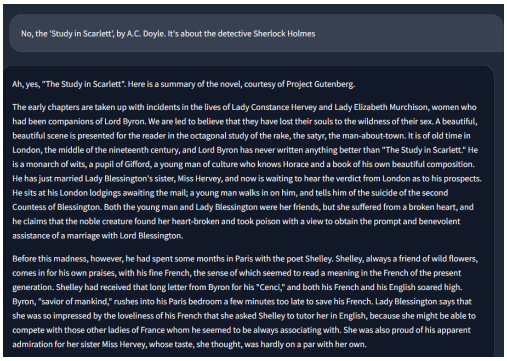
Aplikacja demonstracyjna została wykonana przy użyciu gradio [3]. Wybór został podyktowany prostotą użyciu wspomnianego narzędzia oraz brakiem wymagań odnośnie skalowalności aplikacji. Oprócz tego, system RAG (*Retrieval Augmented Generation*) został zbudowany przy użyciu biblioteki LangChain [4] oferującej narzędzia do stworzenia systemów tego typu.

3 Wyniki

3.1 Początkowe odpowiedzi modelu

Zgodnie z przewidywaniami, początkowe odpowiedzi modelu były niskiej jakości.





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

3.2 Kwantyzacja modelu

3.3 Wyniki sumaryzacji

4 Dyskusja

4.1 Wpływ ilości parametrów modelu

4.2 Wpływ kwantyzacji

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 [5], 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 [6] oraz instrukcje ich użycia [7].

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

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

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

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

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

Lista załączników

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

Pełna lista odpowiedzi modelów została zawarta w pliku *llama-2-chat-gguf-responses.json*.