Artistic ChatBot PoC

Bartosz Grabek, Filip Kucia, Szymon Trochimiak

Warsaw University of Technology (WUT)

December 11, 2024

Overview

- Develop a voice-to-voice chatbot for the Faculty of Media Art.
- Teach the chatbot to interact creatively and understand artistic attitudes.
- Answer questions in Polish about the faculty's history, professors, and future.
- Challenges in artistic domains: subjectivity, abstract concepts, and creativity.
- Gap: Few AI solutions cater to artistic and cultural contexts.

Research Questions

The answers to the following research questions are to be investigated throughout the project:

- **RQ1:** How should an LLM be fine-tuned to represent a specialized artistic domain authentically?
- **RQ2:** What methods can be used to ensure effective retrieval of accurate, context-aware information from the knowledge base during real-time question-answer (QA) interactions?
- **RQ3:** How to allow the chatbot to make predictions and conjectures about the future based on historical data and deliver the responses in a creative way?

Hypotheses

We also plan to test the following hypotheses:

- **H1:** Fine-tuning is less efficient than RAG for improving LLMs QA capabilities if the data available is scarce relative to the size of the model.
- **H2:** A Retrieval-Augmented Generation (RAG) pipeline can dynamically and accurately retrieve curated information for nuanced questions in the artistic domain.
- **H3:** Generating creative responses with future conjectures is possible without providing explicit training data with predictions about the future.

Methodology

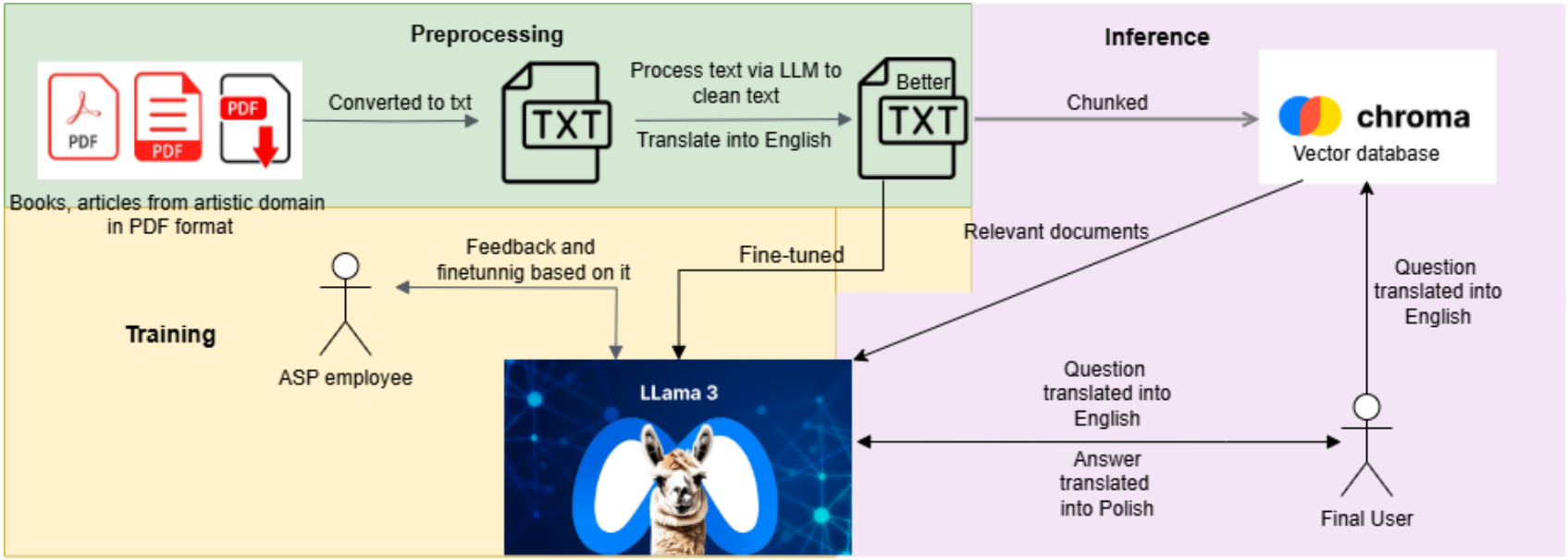
- **Fine-tuning**: Use LoRA (Low-Rank Adaptation) for efficient parameter optimization.
- **Information Retrieval**: Integrate Retrieval-Augmented Generation (RAG) for QA.
- **Dataset**: 165 PDFs curated from books, articles, and archives.
- **Tools**: LLama 3.2 model, OpenAI GPT models, third-party ASR/TTS services.

Data Preparation and Fine-Tuning

Pipeline Overview:

- **Input Data:** Books, articles, and resources from the artistic domain in PDF format.
- **Text Conversion:** PDFs are converted into plain text files for further processing.
- **Text Cleaning:** The raw text is cleaned using a Large Language Model (LLM) to improve quality and consistency.
- **Feedback Loop:** Feedback from ASP employees is collected, refining the model through fine-tuning based on the artistic domain-specific text.

Architecture Overview

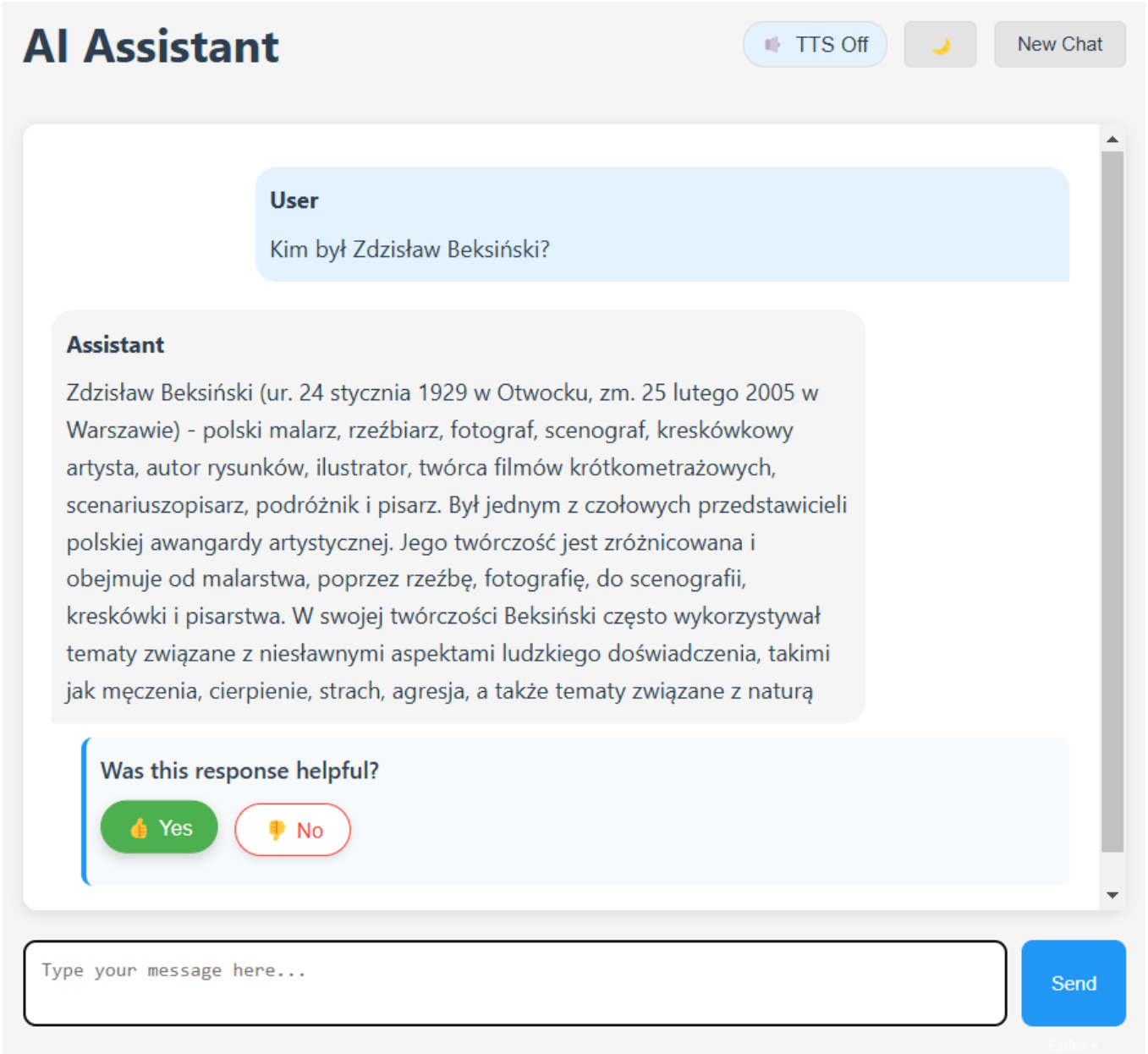


Inference and User Interaction

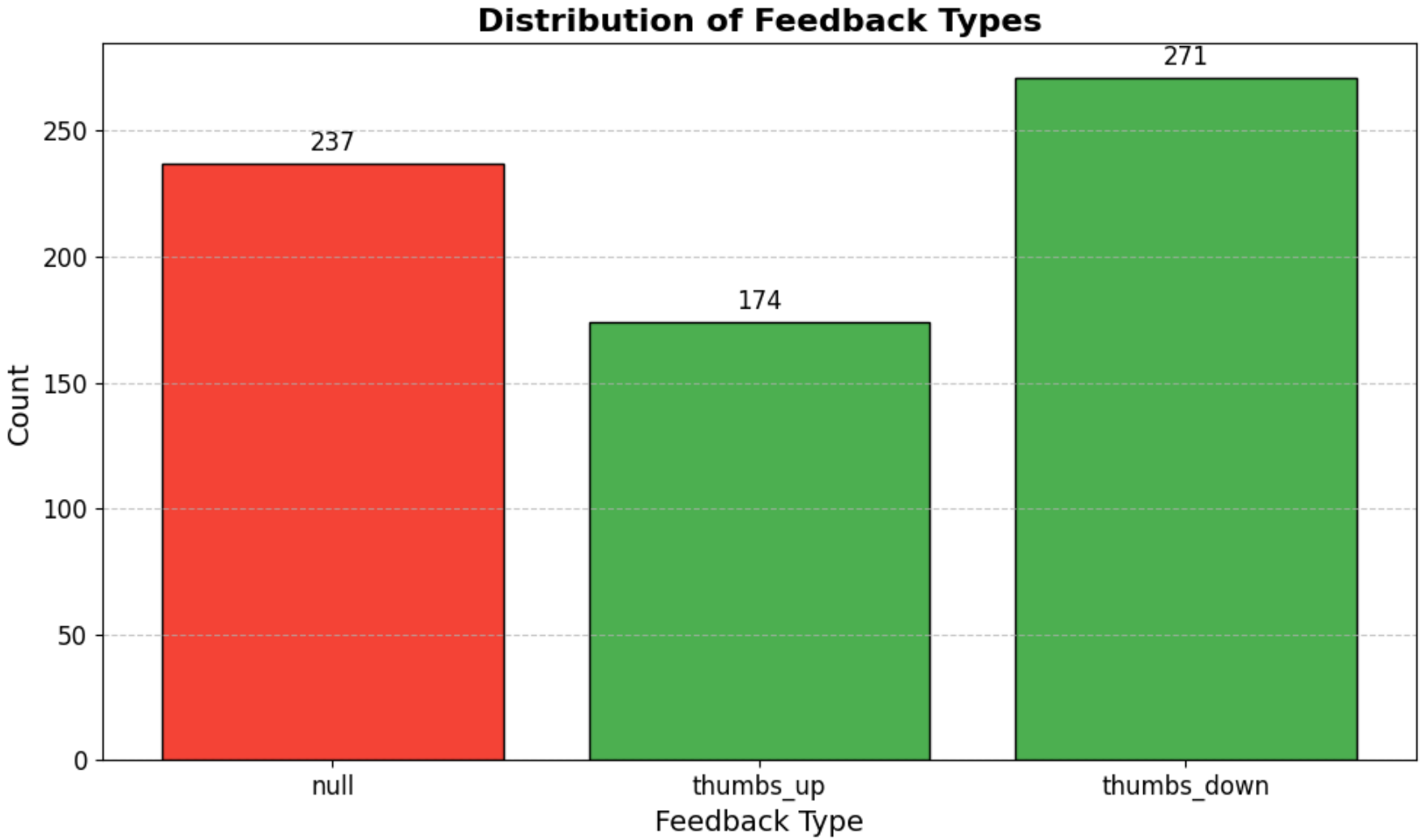
Pipeline Overview:

- **Input Question:** User's question is translated into English, because the input is always in polish.
- **Chroma Vector Database:** The question is matched with relevant chunks of documents stored in a vector database.
- **Response Generation:**
  - The LLM processes the relevant chunks of documents and generates a response in English.
  - The response is translated back into Polish for the user.

Application for collecting QA data & feedback



First feedback from ASP employees



Word Cloud out of input questions

Challenges and Risks

Challenges:

- Limited dataset size for fine-tuning.
- Risk of overfitting and repetitive responses.
- Tight deadlines and limited computational resources.

Mitigation Strategies:

- Feedback-driven iterative fine-tuning.

Conclusion and Next Steps

- The project bridges AI and the arts, fostering creativity in NLP.
- Deployment scheduled for the university exhibition in mid-January 2025.
- Future work: Analyze exhibition feedback and improve model performance.

Thank you for your attention! Questions?

References

- Franceschelli, G., & Musolesi, M. (2024). Creativity in Large Language Models
- Grattafiori, A., et al. (2024). The LLama 3 Herd of Models
- Hu, E. J., et al. (2021). Low-Rank Adaptation for LLMs
- More references in the project report