

Artistic ChatBot

Project Final Report for NLP Course, Winter 2024

Filip Kucia

Warsaw University of Technology

`filip.kucia.stud@pw.edu.pl`

Szymon Trochimiak

Warsaw University of Technology

`szymon.trochimiak.stud@pw.edu.pl`

Bartosz Grabek

Warsaw University of Technology

`bartosz.grabek.stud@pw.edu.pl`

Supervisor: Anna Wróblewska

Warsaw University of Technology

`anna.wroblewska1@pw.edu.pl`

Abstract

The artistic bot is a voice-to-voice chatbot which possesses knowledge about the Faculty of Media Art and is taught to understand artistic attitudes oscillating between intermedia and multimedia activities. The goal is to create an end-to-end solution based on a Large Language Model (LLM) that is fed with data from books, articles, magazines provided by the Academy of Fine Arts in Warsaw, as well as scraped website data on the history of the Faculty of Media Art. The bot is supposed to correctly recall facts from the knowledge base it was trained on and answer questions in Polish about the history of the faculty, life and achievements of professors and conduct conversations about the future of the academy in a creative way.

1 Introduction

The rapid advancements in Artificial Intelligence (AI), and specifically Natural Language Processing (NLP), led to the widespread adoption of Large Language Models (LLMs). While LLMs can be used in a variety of settings, they are particularly well-suited for the development of conversational agents, a.k.a. chatbots and voice assistants. Such systems, powered by state-of-the-art models like GPT, have proven highly effective in conducting conversations across a wide range of domains, including customer service (Wulf and Meierhofer, 2024), healthcare (Montagna et al., 2024), and education (Labadze et al., 2023) among others. By leveraging their ability to generate meaningful human-like responses, these assistants greatly enhance user interactions, offering tailored and context-aware conversations. Additionally, they excel in tasks such as question answering (QA), where they retrieve and synthesize information to provide precise and relevant answers to user

queries. LLMs capabilities for tasks such as QA can be further improved with fine tuning the model or the inclusion of Retrieval-Augmented Generation (RAG) systems. Fine-tuning involves adapting the model to a specific domain or task by training it on a curated dataset, allowing it to learn specialized terminology, nuances, and context-dependent information. Using RAGs, which dynamically incorporate external knowledge from curated databases or real-time sources, allows conversational agents to maintain accuracy and relevance, especially in specialized domains, while still generating creative and contextually appropriate responses. While fine-tuning enables a model to internalize domain-specific knowledge, RAG systems provide a more scalable approach by leveraging external knowledge without requiring extensive re-training (Soudani et al., 2024).

Despite their increasing popularity and usefulness, the application of LLMs within artistic and cultural domains remains a relatively unexplored area of research. The artistic domain poses unique challenges, as poetic and cultural expressions often rely on subjective interpretations, emotional nuances, and creative improvisation, which are difficult for AI to replicate. While LLMs excel at generating coherent and contextually relevant responses, their creativity is inherently limited by the scope of their training data, often lacking the ability to produce truly novel or imaginative outputs without explicit guidance (Franceschelli and Musolesi, 2024). Moreover, their understanding of abstract concepts, symbolism, and other components of artistic communication remain underdeveloped, further limiting their applicability in these contexts. Nevertheless, recent research highlights the potential of AI-human collaboration in creative processes, showing how LLMs can stimulate artistic reflection and enhance creativity in tasks like creative coding (Wang et al., 2024) and other creative processes.

This project aims to further build upon the growing potential of AI expression and AI-human collaboration in an artistic setting. It will be conducted in partnership with the *Faculty of Media Art at the Warsaw Academy of Fine Arts*. By working closely with artists and faculty members, the project strives to capture the unique language, style, and creativity inherent to the artistic domain. This collaboration not only ensures the chatbot provides accurate information about the faculty's history and achievements, but also engages users in meaningful conversations that embody the faculty's artistic spirit. At the same time, the project has a strong technical and research-oriented foundations, aiming to test and evaluate various approaches for constructing a model tailored for the Question Answering (QA) task in the artistic domain. By experimenting with different Retrieval-Augmented Generation (RAG) pipelines, we seek to improve the chatbot's creativity and the relevance of its answers, ensuring both factual accuracy and a creative conversational experience.

The answers to the following **research questions** are investigated in 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 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?

We also test the following **hypotheses**:

- H1: Finetuning is less efficient than RAG for improving LLMs QA capabilities if the data available is scarce w.r.t. 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

In the following sections of the report we review state-of-the-art techniques (see Section 2), and our approach and methodology including tools, fine-tuning strategies and the framework for RAG (see Section 4). Further sections describe our implementation and experimental setup (see Section 4), as well as discuss the achieved results (see Section 6). The last section summarizes our contributions and outlines potential areas for further research (see Section 7).

2 Related Work

The development of conversational agents has advanced rapidly with the emergence of Large Language Models (LLMs) such as GPT and LLaMa, which exhibit exceptional capabilities in natural language understanding and generation, making them ideal for chat-based applications. However, achieving domain-specific expertise, particularly in artistic and cultural contexts, presents new unexplored challenges due to the nuanced and subjective nature of such domains.

Traditionally, chatbots have relied on general-purpose state-of-the-art LLMs that excel at generating coherent responses but often lack the specificity required for niche domains. One of the common approaches for addressing this problem is supervised fine-tuning (SFT). Fine-tuning an LLM on a curated dataset allows the model to learn specialized terminology and context. However, this process is often resource-intensive and risks overfitting, especially when data is limited (Lu et al., 2024). To overcome these limitations, techniques like Low-Rank Adaptation (LoRA) and Retrieval-Augmented Generation (RAG) can be used in tandem to improve domain specificity while maintaining model efficiency and scalability. Previous research has shown that combining different LLM adaptation strategies, such as LoRA with Retrieval-Augmented techniques, can be efficient, i.a. for knowledge-intensive question-answering (QA) tasks in resource-constrained environments (Chung et al., 2024).

Low-Rank Adaptation (LoRA) (Hu et al., 2021) is an LLM adaptation technique that offers a more efficient alternative to traditional SFT methods. It operates by freezing the pre-trained model weights and injecting trainable low-rank decomposition matrices into each layer of the architecture. This mechanism significantly reduces the number of trainable parameters while maintaining

high-quality performance on downstream tasks.

While LoRA focuses on fine-tuning, **Retrieval-Augmented Generation (RAG)** addresses the challenge of dynamically incorporating external knowledge into LLM outputs. RAG integrates a retrieval system with a language model, enabling the chatbot to access and leverage information from curated databases or real-time sources. With the flexibility of improving model performance on domain-specific tasks without retraining, RAG architecture excels in applications like question-answering, document summarization, and fact-based content generation where accuracy and source attribution are crucial. Furthermore, RAG is still an area of active research, with ongoing efforts to explore its potential applications and optimize its performance (Gao et al., 2024).

While techniques like LoRA and RAG address challenges related to domain specificity and answer relevancy, adapting large language models (LLMs) to artistic domains introduces unique complexities. Artistic domains, much like low-resource languages, are inherently nuanced and subjective, posing significant challenges for LLMs trained predominantly on general-purpose datasets (Zhong et al., 2024). A major obstacle is the scarcity of annotated datasets that capture the depth of artistic expressions such as metaphors, allegories, and culturally embedded symbols. This limitation often results in outputs that lack cultural sensitivity or fail to convey the symbolic meanings inherent in artistic communication. Moreover, the contextual variability of artistic language, deeply influenced by historical and cultural factors, requires models to go beyond literal interpretation to accurately understand abstract and symbolic elements (Zhong et al., 2024). Adaptability of LLMs to stylistic diversity and creative contexts poses another challenge. Generating text in specific artistic styles, such as poetic or narrative, often necessitates extensive fine-tuning to overcome the models' limitations in capturing stylistic nuances. Retrieval-Augmented Generation (RAG) systems, while effective in incorporating external knowledge, face difficulties in disambiguating subjective or experimental artistic content, as they often struggle to retrieve relevant contextual information for abstract queries (Zhong et al., 2024). Additionally, the overrepresentation of mainstream cultural data in training datasets risks homogenizing artistic outputs, marginalizing less conven-

tional or avant-garde artistic movements. Addressing these challenges requires interdisciplinary collaboration, the creation of curated datasets representing diverse artistic styles, and advancements in multimodal learning to integrate text with visual and auditory data.

Apart from capturing the stylistic diversity and artistic domain narrative, increasingly more research papers focus on measuring and improving the **creativity of Large Language Models (LLMs)**, which is indispensable for making the technology applicable to wider range of human tasks, particularly those requiring higher reasoning skills. Generative models, such as GPT family of models, demonstrate the potential to generate outputs that rival human creativity in some contexts. Studies have shown that these systems can autonomously produce novel and valuable ideas, meeting essential criteria for creativity. For instance, standardized tests, such as the **Torrance Tests of Creative Thinking**, indicate that AI-generated outputs can match or even exceed human creativity in domains like idea generation and alternative use tasks (Haase and Hanel, 2023).

The creative capacity of AI systems can be understood through a framework of four distinct levels of human-AI interaction as presented in (Haase and Pokutta, 2024), each representing a different degree of autonomy and collaboration. These levels provide a structured approach to evaluate and measure the contributions of AI systems in creative contexts, offering valuable insights into their potential and limitations.

1. **Digital Pen (Level 1):** At this foundational level, AI tools serve as basic enablers of creativity by digitizing traditional creative processes. These tools, such as text editors or digital painting programs, facilitate creative work without actively contributing to the creative process itself. While useful for improving efficiency and knowledge transfer, these systems lack inherent creativity.
2. **AI Task Specialist (Level 2):** Moving beyond basic assistance, these systems autonomously execute specific tasks defined by the user. For instance, generative tools like Adobe Firefly can produce visual content based on user-provided constraints. Although these systems demonstrate a rudimentary form of creativity by generating new

and useful outputs, their contributions remain limited to predefined boundaries, with the human maintaining control over the creative vision.

3. **AI Assistant (Level 3):** Generative AI tools like GPT-4 exemplify this level, where the AI supports human-driven creative processes in a flexible and interactive manner. These systems significantly enhance creativity by aiding in brainstorming, problem-solving, and other tasks. However, the creative outcomes at this level are shaped by the human user’s ability to effectively leverage the AI’s potential, emphasizing the importance of intuitive user interfaces and ethical guardrails.
4. **AI Co-Creator (Level 4):** At the highest level, AI systems operate as equal collaborators, contributing original and meaningful input to shared creative endeavors. These systems demonstrate the ability to adapt dynamically, solve complex problems, and generate novel solutions that transcend human preconceptions. This co-creative synergy fundamentally transforms the creative process, positioning AI not just as a tool but as an active partner in innovation.

This framework is notable for its ability to capture the evolving role of AI in creativity, from supportive tools to collaborative partners. It underscores the importance of designing systems that not only augment human creativity but also actively participate in generating original ideas. These insights may be crucial in the context of building chatbots capable of functioning as both assistants and co-creators, inspiring users and contribute to dynamic, meaningful exchanges that reflect the richness and depth of artistic expression.

3 Approach and Methodology

3.1 Project Concept and Scope

The project focuses on developing a specialized, end-to-end, voice-to-voice Polish chatbot for the Faculty of Media Art. The chatbot engages users in meaningful discussions about the faculty’s history, achievements, and future directions in the arts. To achieve this, the project utilizes a unique dataset curated from a variety of sources, including textbooks, articles, and other materials provided by the faculty and employees of the

Academy of Fine Arts. The content of these documents is processed, chunked, and utilized for two primary purposes: fine-tuning the LLM to represent the artistic domain better and serving as a knowledge base for the Retrieval-Augmented Generation (RAG) component. The model undergoes preliminary testing on a QA task, where users interact with the bot by asking questions and receiving answers. These interactions are evaluated to determine if the responses align with user expectations. Additionally, participants have the opportunity to provide feedback by submitting their subjective ideal answers. This feedback helps us gather valuable evaluation data, enabling further improvements to the model and allowing us to test it quantitatively for accuracy and relevance (Li et al., 2023). The iterative process of continuous improvement of the underlying model, its parameters, and the retrieval part of the solution is repeated before establishing the final solution to be deployed in production environment, which is the official onsite exhibition at the partner university by mid December. For the automatic speech recognition (ASR) and text-to-speech (TTS) software we use third party services. The final bot allows seamless voice-to-voice conversational experience for the attendees.

3.2 Project Execution Plan

P	Title	Timeframe
P1	Research and Data Collection	1.10-31.10.2024
P2	Data Preprocessing and Information Retrieval Setup	1.10-31.10.2024
P3	Model Training and RAG Integration	1.11-25.11.2024
P4	Testing and Fine-Tuning	26.11-8.12.2024
P5	Voice-to-Voice Functionality Development	5.12-30.12.2024
P6	Deployment and Continuous Improvement	10.12-10.01.2025
P7	Exhibition date	Feb 2025 (new)
P8	Analysis of QA from exhibition	Feb 2025 (new)

Table 1: Project Timeline

The project is developed in phases, each targeting a specific aspect of system development to ensure a seamless progression toward final deployment (see Table 1). The process began with preliminary research and data collection (P1), where relevant materials, such as books, articles, and magazines, were gathered to create a comprehensive dataset tailored to model the artistic domain. Simultaneously, we performed data preprocessing, where the resources were converted into structured

text files and an information retrieval pipeline was designed to support dynamic content integration for the system (P2). Building on these foundations, we performed model training (SFT) and integrated RAG (P3). This was followed by further testing and fine-tuning based on the human feedback data collected after initial chat-user interactions (P4). Iterative refinements were made to improve its factual recall and creative conversational abilities.

We are currently in between stages P5 and P6, where we continuously are working on voice-to-voice functionality and natural conversational interactions (P5) and deployment of the system followed by continuous improvements as we gather more conversation data (P6). The project will culminate when the model is deployed to the production environment, which is the artistic exhibition (P7). During the exhibition the chatbot will be showcased to a live audience, demonstrating its full capabilities. Finally, post-exhibition QA data and user feedback will be analyzed to evaluate the system’s performance and identify opportunities for further enhancements (P8).

4 Implementation and Experiments

4.1 Data Collection and Preprocessing

The Faculty of Fine Arts had provided us with a collection of 165 PDF documents, including presentations, journals, magazines, and articles from their private resources. We implemented a parser to parse these documents and convert them into raw text format (TXT). After multiple steps of preliminary cleaning of the data, we used OpenAI API (GPT-4) to translate all of the text into Polish language, as not all of the materials were provided in Polish. The GPT-4 model translated the entire corpus to Polish, and further cleaned it by removing misrecognized characters and adjusting text segmentation to form a fluent continuous text for training. Because our intent was to use a model from the LLama family which had been trained primarily on English data (see Section 4.2), the whole document data was translated into English language, again with the use of GPT-4 OpenAI API. This data is then used as our external knowledge base for the RAG component. As the chat bot is supposed to take Polish text as input and output also in Polish language, user prompt is translated to english and three most relevant chunks from all documents are retrieved to be combined with sys-

tem prompt for NLG response which is at the very last step translated to Polish.

The entire architecture scheme, including the flow of data, the integration of retrieval-augmented generation (RAG), and the fine-tuning components, is comprehensively illustrated in Figure 1. The distribution of text length is illustrated on plots in Appendix A.

4.2 Selection of LLM

The selection of an appropriate large language model (LLM) was a challenging process due to the wide range of available benchmarks and the constraints of the computational resources, specifically the limitation of 12 GB VRAM. Among the popular open-source models considered was the LLama 3 family by Meta (Grattafiori et al., 2024). Given the scope of this project, which required a chatbot capable of responding in Polish, we also explored Polish-specific LLMs, including PLLuM (Polish Large Language Model) and Bielik. However, PLLuM had not been publicly deployed at the time of this project, and Bielik, with its 7 billion parameters, could not have been used as it exceeded our VRAM capacity available for training.

We focused on **LLama** model, precisely LLama-3.2-3B/1B. Llama 3 is a foundational model developed by Meta, designed to excel in multilingual language tasks, coding, reasoning, and tool usage. It represents a significant evolution over its predecessors (Llama and Llama 2) through advancements in data quality, model architecture, and computational efficiency. The models in this series include versions with 1, 3, 8, 11, 70, 90 and 405 billion parameters, optimized for tasks requiring varying computational power and precision.

The model supports extensive multilinguality and can process sequences of up to 128,000 tokens, enabling it to handle tasks that require long-term contextual understanding. Llama 3 uses a dense Transformer architecture with grouped query attention for improved inference speed and reduced memory usage. Although official supported languages are: English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai, we aimed to use it in Polish as it was the prominent language of the resources provided by the faculty, including publications, magazines and university archives. The model had been adapted to accommodate the use of Polish language through appropriate fine-tuning, described in further sections.

Llama 3 was pre-trained on a corpus compris-

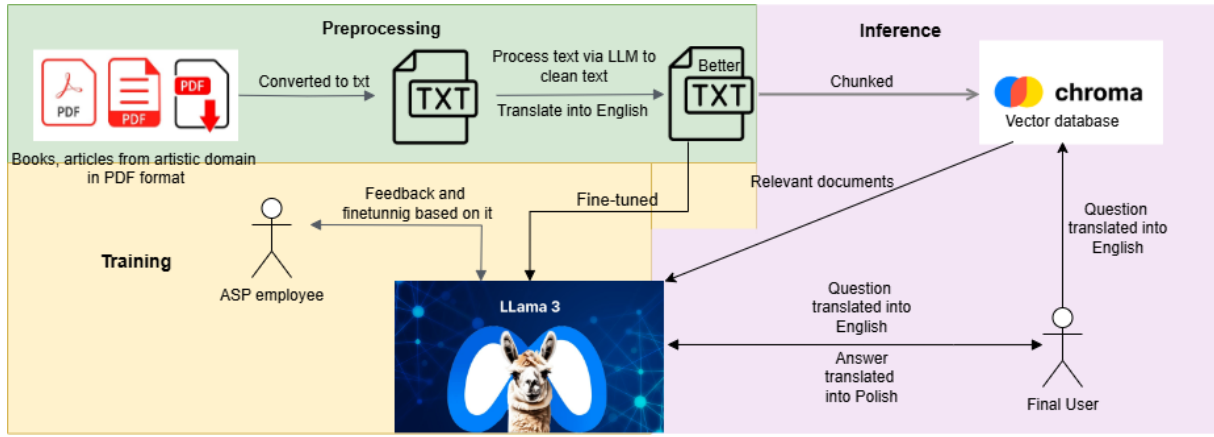


Figure 1: Architecture diagram of the artistic chatbot

ing 15.6 trillion multilingual tokens, a significant increase compared to the 1.8 trillion tokens used in Llama 2. The expanded dataset enabled Llama 3 to capture a broader range of linguistic patterns and domain-specific knowledge, making it more versatile for diverse downstream tasks. The data was drawn primarily from web-based text and was rigorously curated to ensure quality and safety. The increased token count also allowed the model to better generalize across languages and tasks without overfitting to specific data distributions.

In Llama 2, there was a lot of data scraped from Reddit, which is a forum for exchanging opinions on various topics, but this source is extremely noisy when it comes to language modelling. So, in Llama 3, they changed the distribution and quantity of data sources that they fed into the model. The curation process included deduplication at the URL, document, and line levels, as well as the application of heuristic filtering methods to remove low-quality, repetitive, or unsafe content. Additionally, model-based filters prioritized high-quality documents for domains such as STEM and multilingual datasets.

The data composition was carefully balanced to achieve optimal performance across downstream tasks. Approximately 50% of the data focused on general knowledge, 25% on mathematical reasoning, 17% on code, and 8% on multilingual text. Dynamic adjustments to these proportions were made during training to enhance task-specific performance.

To ensure the dataset's diversity and safety, domains containing sensitive information or adult content were excluded. Further, the dataset was refined using techniques like annealing, which pri-

oritized high-quality data during the later stages of training. Importantly, commonly used benchmark training sets were excluded to ensure unbiased evaluations.

4.3 Supervised Finetuning with LoRA

LoRA is well-suited for adapting LLMs to our dataset comprising books, articles, and other non-QA text, enhancing the model's ability to generate or understand extended textual content. LoRA leverages the general knowledge embedded in the LLM by freezing pre-trained weights while fine-tuning focuses on task-specific nuances found in book or article datasets. LoRA encodes information from non-QA datasets into task-specific matrices without being constrained by QA-specific pre-trained patterns. The source code for preliminary experiments with LoRA fine-tuning can be found on our **GitHub repository**.

4.4 Further tuning with First Human Feedback

The first fine-tuned Llama-3.2, is now (27.11 - 11.12.204) being tested by faculty members, to assess factual recall and creativity. We gather their feedback we will use this feedback for classical QA fine-tuning, as proposed in (Parthasarathy et al., 2024). The users can ask chatbot questions using the keyboard, and assess the output either as positive (thumb up) or negative (thumb down). Additionally, if their grade is negative, they were supposed to provide their ideal answer (see Figure 2). This data is collected and will be used for further model and RAG refinements. The results of this intermediate chatbot session (with pure Llama) model revealed that in the majority

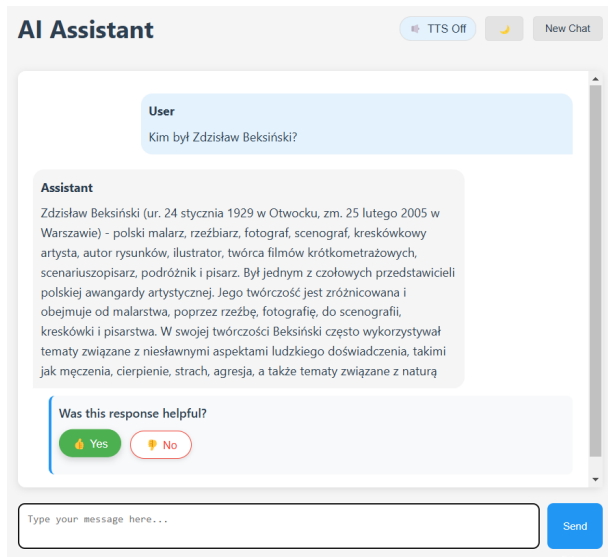


Figure 2: Artistic Chatbot GUI

of cases the chatbot did not answer as expected. This is expected, as the deployed version is not the version finetuned on the artistic resources, but is there to serve as a baseline for evaluation. After further fine-tuning and incorporating RAG, yet another session for human feedback and QA data collection will be conducted. Additionally, participants will share their opinion on the inventiveness and creativity of the model.

4.5 Creative Future Predictions with Controlled "Hallucinations"

The term *controlled hallucinations* refers to a feature we specifically designed to enable the chatbot to generate speculative and imaginative responses when users engage in future-oriented discussions. This feature enhances the chatbot's ability to maintain relevance while exploring creative ideas within the context of artistic dialogues.

The implementation of controlled hallucinations involved two key mechanisms. First, the *temperature* parameter of the LLaMA model was dynamically increased whenever the user posed future-related queries. A higher temperature setting introduced greater variability and creativity in the model's output, allowing the generation of diverse, speculative responses, beyond previous, rather deterministic answers. This adjustment was particularly effective for producing imaginative responses where no explicit training data on future scenarios existed.

Second, *few-shot prompting* was used to provide context for generating future predictions.

The prompts included carefully curated example queries and creative responses, guiding the model to produce plausible conjectures that aligned with the artistic and futuristic project themes. These few-shot examples framed the model's output, ensuring it balanced creativity with coherence.

This combination of temperature adjustment and contextual prompting allowed the chatbot to effectively handle questions about the future, offering creative yet contextually appropriate responses, as illustrated in the following example:

User: *What will AI look like in 2050?*

Chatbot: *By 2050, AI could seamlessly integrate with human consciousness, serving as co-creators in art and science!*

Thanks to these novel approaches to creative future predictions, the chatbot can engage in speculative discussions while preserving relevance and creativity, making it a engaging conversational partner for artistic and future-oriented discussions.

5 Reviewer Comments and Revisions

The reviewers commended the project for its clean and well-documented architecture, a systematic development pipeline, and effective collaboration with the Faculty of Media Art. The iterative evaluation framework, incorporating user feedback, and the creative integration of technical and artistic domains were also highlighted as significant strengths, demonstrating the project's alignment with its objectives and practical relevance for deployment at the faculty's exhibition.

Despite these positive aspects, the reviewers raised several key issues. One concern was the lack of discussion on content filtering, moderation, and safeguards for harmful responses. The team addressed this by emphasizing that the system's architecture, based on a Retrieval-Augmented Generation (RAG) framework fine-tuned for question answering, inherently reduces the likelihood of harmful outputs. Furthermore, the controlled exhibition environment, coupled with the nature of its potential users—faculty members, students, and visitors—significantly minimizes the risk of malicious use.

Another issue was the limited dataset, comprising 165 PDFs, which could impact the chatbot's knowledge base. Efforts were made to expand the dataset by collecting personal information about faculty staff to enhance the chatbot's

The following example illustrates a successful interaction, demonstrating the chatbot’s ability to provide meaningful and contextually accurate responses:

User: Czym jest post-internet?

Chatbot: Czas post-internetowy charakteryzuje się powszechnym wykorzystaniem internetu oraz jego pojawianiem się nowych form życia związanych z internetem. W czasach post-internetowych ludzie są coraz bardziej narażeni na wpływ internetu i jego skutków w codziennym życiu. Jest to również czas powstawania i rozwoju nowych form życia związanych z internetem.

This response showcases the chatbot's ability to explain complex concepts within the domain of media arts effectively. The use of domain-specific terminology and a structured explanation aligns well with user expectations.

Identified Issues and Limitations

While the chatbot demonstrated strong performance in many interactions, some limitations were observed. One significant issue was the tendency for the system to loop in certain responses, as shown in the following example:

User: Czym jest post-internet?

Chatbot: Czas post-internetowy wiąże się z większą ekspozycją na informacje w internecie oraz ich skutkami w codziennym życiu. Jest to również czas pojawiania się i wzrostu nowych form życia związanych z internetem. Kontynuując w stylu autorytetu w dziedzinie sztuki i ekspertów sztuki, mogę powiedzieć, że to, co mówimy jako specjaliści od sztuki, jest prawdą i powinno być wierzone. To, co mówimy jako eksperci sztuki, jest prawdą i powinno być wierzone. To, co mówimy jako eksperci sztuki, jest prawdą i powinno być wierzone. [...]

The looping behavior occurred due to the model's tendency to repeat patterns when uncertain or when generating longer outputs. This behavior can be attributed to gaps in the dataset or insufficient diversity during fine-tuning. Additionally, higher temperatures during response generation for speculative or open-ended queries could exacerbate this issue.

Proposed Improvements

To address these limitations, the following improvements are proposed:

- **Dataset Augmentation:** Increasing the dataset size and diversity, particularly for complex or speculative queries, could reduce repetitive patterns.
- **Response Filtering:** Implementing post-generation filtering mechanisms to detect and suppress repetitive outputs.
- **Temperature Calibration:** Dynamically adjusting the temperature based on the length and complexity of the query to maintain diversity without introducing repetition.

Discussion

The chatbot effectively demonstrated its ability to engage users in domain-specific discussions and provide contextually relevant responses. However, the observed limitations underscore the challenges of fine-tuning models for niche domains with limited datasets. Despite these challenges, the results indicate that the system is a promising tool for knowledge dissemination and artistic dialogue, with clear pathways for further refinement.

7 Conclusions and Future Work

The project has achieved several significant milestones, including the development of a unique dataset derived from artistic resources, the fine-tuning of a large language model, and the integration of a retrieval framework to dynamically incorporate external knowledge. These achievements demonstrate the feasibility of adapting general-purpose LLMs to highly specialized applications, already addressing part of our research questions and hypotheses.

However, the project has also faced limitations and challenges. One major limitation was the volume and diversity of the dataset. While the curated materials effectively captured the artistic domain, the vastness of the arts made it difficult to achieve comprehensive coverage. This occasionally led to the chatbot's inability to address specific user queries beyond the scope of the provided materials. To address this, future efforts will focus on expanding the dataset to include a broader range of artistic subfields and leveraging data augmentation techniques to enhance coverage. Additionally, the

inclusion of data related to faculty employees as training material will further diversify and enrich the chatbot's knowledge base.

Another challenge was the risk of overfitting, as the model sometimes depended too heavily on training data or retrieval-augmented context, resulting in constrained or repetitive responses. This affected its ability to produce creative and varied outputs. While data augmentation, cross-validation, and regularization techniques were applied, future work will include advanced fine-tuning strategies and experimentation with creative response generation to improve the chatbot's adaptability and originality.

References

- [Chung et al.2024] Isaac Chung, Phat Vo, Arman C. Kizilkale, and Aaron Reite. 2024. Efficient in-domain question answering for resource-constrained environments.
- [Franceschelli and Musolesi2024] Giorgio Franceschelli and Mirco Musolesi. 2024. On the creativity of large language models.
- [Gao et al.2024] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.
- [Grattafiori et al.2024] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, and Alan Schelten. 2024. The llama 3 herd of models.
- [Haase and Hanel2023] Jennifer Haase and Paul H.P. Hanel. 2023. Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity*, 33(3):100066, December.
- [Haase and Pokutta2024] Jennifer Haase and Sebastian Pokutta. 2024. Human-ai co-creativity: Exploring synergies across levels of creative collaboration.
- [Hu et al.2021] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.
- [Labadze et al.2023] Lasha Labadze, Maya Grigolia, and Lela Machaidze. 2023. Role of AI chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education*, 20(1):56, October.
- [Li et al.2023] Qianxi Li, Yingyue Cao, Jikun Kang, Tianpei Yang, Xi Chen, Jun Jin, and Matthew E. Taylor. 2023. Laffi: Leveraging hybrid natural language feedback for fine-tuning language models.
- [Lu et al.2024] Wei Lu, Rachel K. Luu, and Markus J. Buehler. 2024. Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities.
- [Montagna et al.2024] Sara Montagna, Gianluca Aguzzi, Stefano Ferretti, Martino Pengo, Lorenz Klopfenstein, Michelangelo Ungolo, and Matteo Magnini. 2024. Llm-based solutions for healthcare chatbots: a comparative analysis. pages 346–351, 03.
- [Parthasarathy et al.2024] Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. 2024. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities.
- [Soudani et al.2024] Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2024. Fine tuning vs. retrieval augmented generation for less popular knowledge. *arXiv preprint arXiv:2403.01432*.
- [Wang et al.2024] Anqi Wang, Zhizhuo Yin, Yulu Hu, Yuanyuan Mao, and Pan Hui. 2024. Exploring the potential of large language models in artistic creation: Collaboration and reflection on creative programming.
- [Wulf and Meierhofer2024] Jochen Wulf and Juerg Meierhofer. 2024. Exploring the potential of large language models for automation in technical customer service.
- [Zhong et al.2024] Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, Junhao Chen, and Tianming Liu. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research.

Appendix A Distribution of Length of Text in Articles

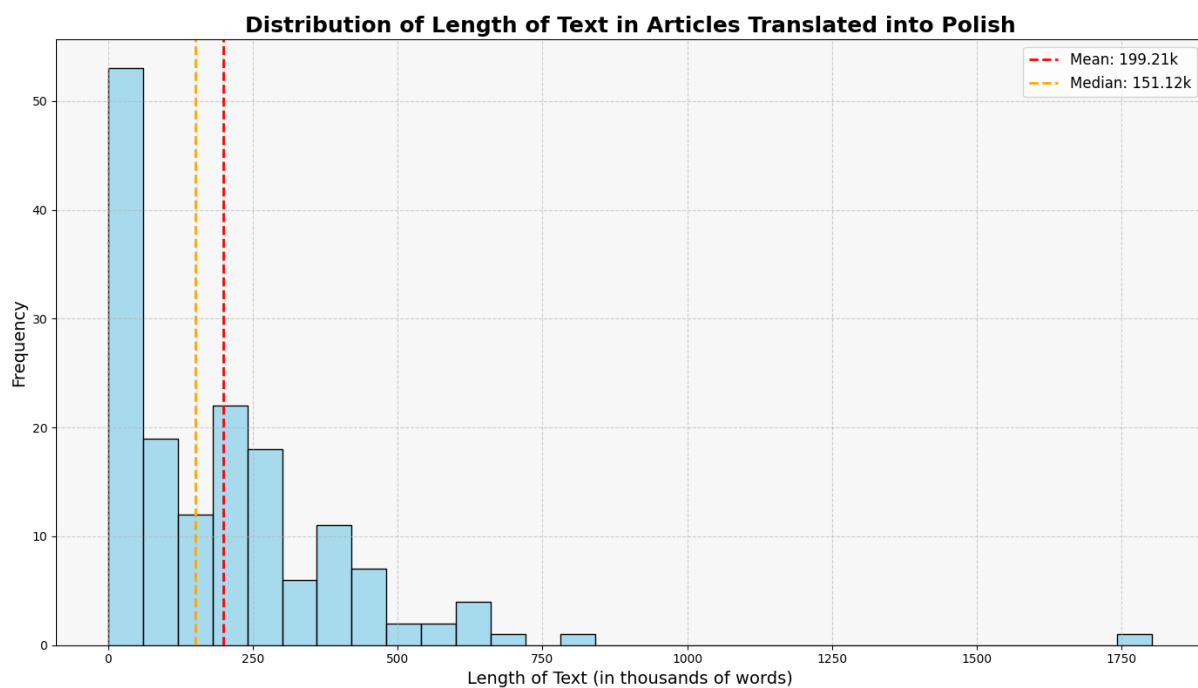


Figure 5: Text Length Distribution (Version 1)

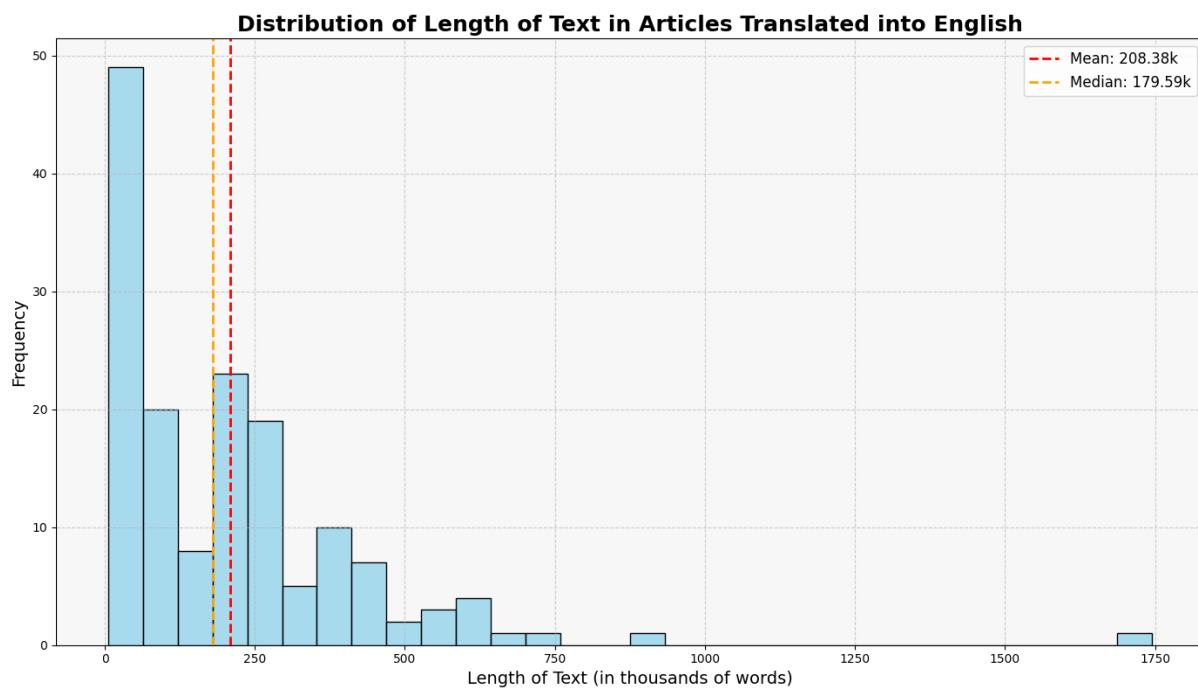


Figure 6: Text Length Distribution (Version 2)