アニメコンテンツ資料を活用した
チャットボットの開発

ChatGPT を始めとした生成系 AI はその正確性や、他人の著作物を許可なく利用している可能性が指摘されており、利用の際には注意が必要である。また、個人情報や機密情報を入力すると、そのデータが学習に利用されてしまうという問題も発生した。そのため、本作品では無償で公開されているデータを用いて AI のトレーニングを行うことで、正確性と著作物の問題を解消しようと試みた。また、トレーニングすることによって、自分が知りたい情報のために最適化された AI を作成することができると考えた。

本作品では、「メディア芸術データベース・ラボ」というサイトから取得したデータを jsonl 形式に変換するための movie.py と regular.py、そのプログラムを実行することによって作成される LD_AnimeMovie.jsonl と LD_AnimeRegular.jsonl、それらのデータをベクトル化して ChromaDB に保存するための list1.py（本作品解説では、LD_AnimeMovie.jsonl のデータのみ利用する）、そして、ChromaDB に保存されたデータを利用してチャットをするための chatbot2.py という 6 つのプログラム及びデータから構成されている。また、全てのプログラムを VScode で作成している。

今回作成したチャットボットは、アニメコンテンツ資料を活用したことによってある程度回答の精度が上昇したと言える。しかし、gpt-3.5-turbo に渡すデータは text-embedding-ada-002 の性能に依存しているため、このモデルの性能次第で回答の精度が大きく変化してしまうことがわかった。また、gpt-3.5-turbo の一度に扱えるトークン数に上限があることで、いくつか前の会話を踏まえた回答の生成が難しいこともわかった。

本作品は、現時点では改善点が多く実用的とは言えないが、ChatGPT は性能が日々向上しており、今後リリースされる性能の高いモデルを使用することによって十分実用的になりうると考えられる。しかし、正確性の問題については未だ改善すべき点が多く、このような問題を解決するためには、より膨大な量の最新かつ正確なデータを学習させることや、間違っている部分を AI が自分で修正できるようになる必要があるだろう。また、ユーザーが AI の限界を認識し、回答を盲目的に信じるのではなく、必要に応じて他の情報源を参照することが重要である。