進捗報告

1　研究方針

先週の金曜日に森先生と相談した通り, ときメモのデータセットを用いて特定のキャラクターらしさを学習した LLM を作成, その LLM からより低いパラメータ数の LLM への蒸留を目指していく形で, 研究に取り組んでいく予定です. もう逃げない

2　シナリオデータからキャラクターを分類するタスク

村田君の JSAI 論文や卒論, 田中さんの研究会資料などを参考にとりあえず BERT でランク付けという形で分類タスクに取り組んでいます.

3　ローカル LLM のファインチューニング

特定のキャラクターらしさを表現する LLM を構築する方法は主に 2 種類存在し, 1 つ目はキャラクターの対話データから Retrieval Augmented Generation (RAG) を用いてユーザーからのプロンプトに関連するデータを与える方式, 2 つ目は対話データを用いて LLM をファインチューニングする方法がある.

キャラクターらしい LLM を構築した研究例として ChatHaruhi [1] がある. 図 1 に示すようにこの研究では, 収集した対話データと GPT-4, GPT-3.5 により生成したデータを用いて RAG 用のデータベースを構築し, LLM にキャラクターのペルソナを与えて RAG によりキャラクターらしい応答を生成している.

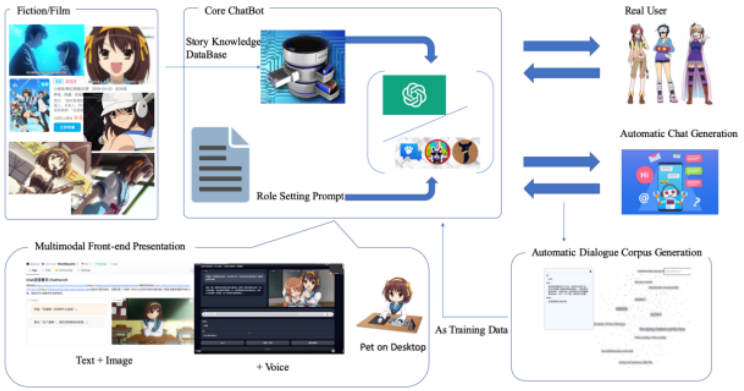

図 1: ChatHaruhi のアーキテクチャ

また, ファインチューニングの手法を主とした研究 [2] も存在しベートーベンやクレオパトラなど歴史上の人物をモデルとしてキャラクター LLM を構築していた. 学習の際には, GPT-3.5 でキャラクターのデータを生成し, Llama 7b のモデルをファインチューニングしていた. どちらも Github 上にコードが公開されていたため, 実際に動かしてみて参考にしようと思っています.

参考文献

[1] Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. Chathaviving anime large language model, 2023.

[2] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing, 2023.