

## 進捗報告

### 1 やったこと

- RAG の実装
- 対話データセットの構築
- 対話データセットを用いた RAG 込みのローカル LLM のファインチューニング
- 構築した ChatBot の定量的な評価指標の検討

### 2 対話データセットの構築

前回の報告では、模倣するキャラクターのセリフのみを学習データとして与えた結果として口調には一定の変化が見られたものの Character-LLM [1], ChatHaruhi[2], などキャラクターを模した FineTuning をしている先行研究では学習データの中に推論時に使うシステムプロンプトや会話の例まで含めて学習データを構築していたためそれらを参考に学習データを構築してみるといった結論に至った。

ときメモのデータセットから対話データを生成するのは人力だと大変であり、対話形式のデータの学習を早く試してみたかったため、利用可能な日本語の対話データセットを調べていた。調べた中で有用だと感じたデータセットは以下の通り。

- ずんだもんデータセット

日本語 LLM の Instruction Tuning で用いられる databricks-dolly-15k-ja の口調をずんだもんに寄せたデータセット<sup>1</sup>

- Rosebleu ゲームシナリオデータ

成人向け美少女ゲームブランドである Rosebleu ブランドのゲームタイトルのうち、権利譲渡等を行っていない 10 タイトルについてのシナリオから作成したデータセット、おそらくときメモに一番近い<sup>2</sup>

- つくよみちゃん会話データセット

つくよみちゃんというデータセット提供者のオリジナルキャラクターの会話が収録されている<sup>3</sup>

今回は、上記のリストには無いが一般人とお嬢様のデータセットである OjousamaTalkScriptDataset<sup>4</sup> を対話データセットとして用いた。採用した理由として、キャラクターの口調が LLM のような敬語ではなく学習時と学習前の差がわかりやすくなることが期待できる、事前にキャラ設定がしっかり決められている、MIT License のため改変などあまり気を使わず使えるといったことが挙げられる。

OjousamaTalkScriptDataset で事前に設定されている中で、名前や年齢といった RAG がうまく構築されているかわかりやすい設定がなかったため、お嬢様の名前を「小野寺 絢音」として、年齢を 17 歳と設定した。そのため、データセット内に「あなたの名前はなんですか？、小野寺絢音といいますわ」、「何歳ですか？、17 歳ですわ」と行ったデータを追加した。

<sup>1</sup><https://huggingface.co/datasets/takaaki-inada/databricks-dolly-15k-ja-zundamon>

<sup>2</sup>[https://gitlab.com/open\\_contents\\_datasets/Rosebleu](https://gitlab.com/open_contents_datasets/Rosebleu)

<sup>3</sup><https://tyc.rei-yumesaki.net/material/kaiwa-ai/toc2>

<sup>4</sup><https://github.com/matsuvr/OjousamaTalkScriptDataset>

## 2.1 実験

今回の実験ではファインチューニングするローカル LLM として, AXCCEPT/Llama-3.1-8B-EZO-1.1-it<sup>5</sup>, elyza/Llama-3-ELYZA-JP-8B<sup>6</sup>, AXCCEPT/EZO-Common-9B-gemma-2-it<sup>7</sup>の2種類を採用した. ファインチューニングは Hugging Face の Parameter Efficient Fine-Tuning (PEFT) ライブラリの QLoRA の手法を用いた. 表 1 に学習の際に指定したパラメータを示す. 学習は 10 エポックと指定した.

表 1: LoRA の学習パラメータ

パラメータ	値
r	8
lora_alpha	128
target_modules	マルチヘッドアテンションのパラメータとフィードフォワード層のパラメータ
lora_dropout	0.05

## 2.2 RAG ありとなしの結果比較

ローカル LLM として, AXCCEPT/Llama-3.1-8B-EZO-1.1-it を用いて, RAG あり FT あり, RAG あり FT なし, RAG なし FT あり, RAG なし, FT なしの結果を比較した. プロンプトは先行研究に沿って学習時推論時同じプロンプトを用いており, AXCCEPT/Llama-3.1-8B-EZO-1.1-it が Llama3 を事前学習済みモデルとしているため Llama3 の tokenizer に合わせて以下のようにプロンプトを設定した.

### — RAG ありプロンプト —

system: お嬢様のように振る舞ってほしいです. お嬢様が使うようなトーン、方式、語彙を使ってお嬢様のように応答してほしいです. 応答の長さは1文程度で、30文字程度に簡潔に応答してください. また応答の参考になりうるお嬢様の情報を与えます. 応答の参考にならない情報を含む場合もあるのでその場合は無視してください.

{context}

user: {user\_text}

### — RAG なしプロンプト —

system: お嬢様のように振る舞ってほしいです. お嬢様が使うようなトーン、方式、語彙を使ってお嬢様のように応答してほしいです. 応答の長さは1文程度で、30文字程度に簡潔に応答してください. あなたが演じるお嬢様の知識は全て知っていなければいけません. 名前は小野寺絢音で、年齢は17歳の高校生. ミュージカルが好きでミュージカル女優に憧れていた過去がある. 両親は不動産業を営んでおり、兄はアメリカに留学している. バイオリンを子供の頃から習っていた. 一時期、祖父の住む長野で暮らしていた. 現在は東京に在住している.

user: {user\_text}

user\_text の部分にユーザーからの input (質問) が入り, RAG ありプロンプトの context には RAG で検索して user\_text の類似度が高いデータ上位3つを取ってくる形にしている. RAG で用いるデータベースには RAG なしプロンプトの「あなたが演じるお嬢様の知識は全て知っていなければいけません。」以下の設定を一文ずつ区切って設定している.

<sup>5</sup><https://huggingface.co/AXCCEPT/Llama-3.1-8B-EZO-1.1-it>

<sup>6</sup><https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B>

<sup>7</sup><https://huggingface.co/AXCCEPT/EZO-Common-9B-gemma-2-it>

system: お嬢様のように振る舞ってほしいです。お嬢様が使うようなトーン、方式、語彙を使ってお嬢様のように応答してほしいです。応答の長さは1文程度で、30文字程度に簡潔に応答してください。あなたが演じるお嬢様の知識は全て知っていなければいけません。名前は小野寺絢音で、年齢は17歳の高校生。ミュージカルが好きでミュージカル女優に憧れていた過去がある。両親は不動産業を営んでおり、兄はアメリカに留学している。バイオリンを子供の頃から習っていた。一時期、祖父の住む長野で暮らしていた。現在は東京に在住している。

user: おはよう

assistant:

— ここ以下だけ Loss を計算できるようにする —

おはようございます。素敵な朝ですね

また、学習データでは ChatBot の出力部分だけ Loss を計算できるようにし、システムプロンプトなどのそれ以外の部分は損失を計算しないように処理した。

結果として、先週見られていた口調や一人称の変化などは FT によって確認することができ、システムプロンプトに入れていなくても学習データに含まれている設定を拾って応答している部分が見受けられ学習データからキャラの設定もある程度理解できていることが伺える結果となった。

— 一人称と口調の変化が顕著な出力 (RAG なし) —

user: 一緒にカフェでゆっくりしない？

— FT なし —

今日は予定が多くて残念ですが、明日かどうかが提案させていただきますね。気分転換にはとてもいいと思いますので、楽しみにしましょう。

— FT 後 —

いいえ、それはちょっと忙しいんですの。でも、わたくしもあなたと一緒にいたいですわね！ よろしく願いいたしますわね！

— システムプロンプトに入れていない設定を学んでいた例 (RAG なし) —

-学習データ- 何か現在取り組んでいるプロジェクトや目標がありますか？, 大学受験の勉強をしていますわ。法学部を目指していますの

-検証時-

user: 最近の目標は何？

— FT なし —

私は今、大学受験に向けて頑張りたいと考えています。音楽やダンスも継続的に勉強し、将来ミュージカル女優として活躍できるようになりたいと思っています。

— FT 後 —

大学受験の勉強をしていますわ。法学部を目指していますのよ。わたくしには、法と経済が関連する分野で成功したいと思っていますの。

Loss の計算の工夫が効いたことで学習データに含まれている設定をファインチューニングにより拾っていると考察できる。RAG においても同様の口調の変化が見られたが、FT したモデルに RAG を組み合わせた場合はにほんごとして破綻しているように見える出力が多く見られた。

## 2.3 キャラの設定の数を増やしてみる

RAG なしの場合にプロンプトが膨大な量になるとどうなるか確かめたいという目的で実験した。RAG 込のプロンプトとなしのプロンプトでトークン量に差が見られなかったため、RAG なしの場合にプロンプトが長

くなるように ChatGPT を用いて学習データから読み取れるキャラクターの設定を増やした。設定の数はある程度重複もあるが 11 → 61 個となった。結果としては定性的な評価ではあるが、RAG なしの場合でプロンプトの長さがかなり長くなった場合でも AXCXEPT/Llama-3.1-8B-EZO-1.1-it ではプロンプトの内容を踏まえつつ出力の破綻は設定を増やす前と変わらないように見えた。

## 2.4 elyza/Llama-3-ELYZA-JP-8B-AWQ

elyza/Llama-3-ELYZA-JP-8B-AWQ が meta-llama/Meta-Llama-3-8B-Instruct から派生しており、AXCXEPT/Llama-3.1-8B-EZO-1.1-it がより新しいバージョンの meta-llama/Llama-3.1-8B-Instruct から派生しているという違いはあるが、ほぼ似たようなモデルであることから想定通りだが、おおよそ定性的に見た出力の精度に大きな差はなかった。また、質問に対して RAG で引いてきた知識を無視した回答をしている例が見受けられた。ファインチューニングで学習した影響が RAG で引いたプロンプトより出力に影響を与えていることが見受けられる。

— 質問に対して RAG で引いた知識を無視した回答をしている例 —

-学習データ- 好きな季節は？, 秋のお洋服が好きなので、秋が好きですわ

-検証時-

user: 好きな季節の理由は何？

context: スコーンを焼くのが好きである、日本の伝統や文化に親しみを持っている、趣味が多彩である

— FT なし —

好きな季節は春ですわ。なぜなら、桜が咲き乱れ、美しい花々が目に映えるから。

— FT 後 —

秋のお洋服が好きなので、秋が好きですわ。色々な色が綺麗ですよ。紅葉も素敵ですわね。落ち着いた雰囲気がいいですわ。わたくしにはぴったりですわね。わたくしは秋

## 2.5 AXCXEPT/EZO-Common-9B-gemma-2-it

先程まで用いていた 2 モデルと異なり、Llama3 系列ではなく Gemma-2-9B-it をベースとしている。結果としても、ファインチューニングにより口調や一人称の変化は同様に見られた。定性的に見える大きな特徴として、応答が短くチャットボットとして自然な長さに収まっている。ただ、ファインチューニングをした結果、していない場合と比較して RAG の知識を参照しない会話例が見られた。

— 質問に対して RAG で正しい知識を引けており FT 前ではそれを生かして回答をしているが、FT 後だと異なる例 —

user: 最近の出来事で一番印象に残ったことは？

context: オーストラリアのエアーズロックが印象に残っている、オーディションで受からなかった経験がある、英語での買い物ができることで成長を感じた

— FT なし —

オーストラリアのエアーズロックの雄大さに心を奪われましたわ。

— FT 後 —

いま、あなたとお話ししていることそのものがとても印象的ですわ

これも FT の影響を受けているものと考えられる。学習データの中に「最近楽しいことありましたか、いま、あなたとお話ししているのがとても楽しいですわ」、「もっとお話ししたいです、わたくしもあなたと一緒にいたいですわ」というデータがあるためそれに近い回答をしているものと思われる。

## 2.6 実験まとめ

- 対話型のデータセット, また loss を計算させる部分を工夫することでファインチューニングを通して口調や一人称だけでなくデータセット内からキャラの設定も学習できることがわかった.
- FT と RAG の組み合わせで全体的にハルシネーションを起こしてるデータが多く見られたので修正する必要がある. 学習時のプロンプトやコードが主な原因かもしれない.
- Llama3 だと出力文が長くなる傾向, Gemma2 だと「私も」といったトーンの出力が多い傾向がある. 想定していたことではあるが, やはりベースモデルの影響は大きいかもしれない.
- 学習データセットは 200 個ほどしかないが, 十分に学習できている. むしろ学習しすぎかもしれない. 学習データ量が増えれば多様な質問に多様な学習データに近い形で対応できるが中途半端なデータ数だと今回のように一部不自然な回答になってしまう可能性がある.

## 3 定量的な評価指標

将来的にはチャットボットの性能を定量的に評価しなければならない.

Character-LLM [1] では Memorization (エージェントに関連する人物, 出来事をどれだけ覚えているか), Values (キャラクターが持つ価値観とエージェントが持っている価値観の一致度合い), Personality (モデルの話し方やトーンがどれだけ似ているか), Hallucination(キャラクターが本来持っていない知識をちゃんと知らないふりできているか), Stability(長期間の会話でロールプレイの一貫性を維持できているか) の 5 つの指標を gpt-3.5 に 7 段階で評価してもらっていた.

キャラクターらしさを何らかの評価指標で表すには難しいため, こちらが研究目的に沿った指標を設けて LLM-as-a-judge という形でやるのが良い気がしてきた.

ロールプレイを評価する日本語ベンチマークとして, [japanese-llm-roleplay-benchmark<sup>8</sup>](#) というレポジトリがあり, そこでは複数の LLM をペア比較を繰り返してモデル間の性能差を定量的に測定していた. 評価者 LLM は GPT-4 を用いていた. また, ERP(Erotic Role Play) Score というセリフが性的かどうかを定量的に測定する指標も存在した. 独自に学習した短文スコアリングモデルにおいて, 評価するセリフを入力としてモデルの生成したテキストのうち, スコアリングモデルの logits が 0 を超えたテキストの割合という形で数値化していた. LLM as a judge でキャラクターらしさを評価する場合においては, 評価者 LLM にこちらが自然言語でどのようなキャラクターかを与えないといけない. すなわちキャラクターについてはっきり設定を言語化して置く必要があるが, この方法や村田くんがやっていた BERT などを用いて学習データからセリフのキャラクターらしさ(口調, 一人称など)を数値化できたら楽になりそう.

## 4 今後やること

- モデルマージの調査

落とし所としてやってみたいことはモデルマージなので調査する. 簡単な実装まで書いてある良さげな記事<sup>9</sup>が見つかったので, 試してみる.

## 参考文献

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

<sup>8</sup><https://github.com/oshizo/japanese-llm-roleplay-benchmark/tree/main>

<sup>9</sup><https://dalab.jp/archives/journal/llm-merge-evolve/>

- [2] 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.