

進捗報告

1 やったこと

- モデルマージ動作確認
- gemma-2-2b-jpn-it¹ のファインチューニング
- LLM-as-a-Judge の実装

2 モデルマージ動作確認

進化的モデルマージを試している記事²を参考に、モデルマージを試したが記事内の 7b のモデルを 3 つマージする構成では最後のマージ後のモデルをロードする部分で Cuda out of memory となった。

そこで、進化的モデルマージのライブラリである Mergekit の公式サイトチュートリアル³を参考に、「24GB of VRAM will do just fine」と書いてあったため試してみたところ、Cuda out of memory となってしまった。サーバーで使っている GPU は RTX 3090 で VRAM は 24GB あるが、なぜか cuda out of memory となるためモデルの数を減らす、低パラメータのモデルを使用することで最後まで動く条件を確かめる。

3 gemma-2-2b-jpn-it のファインチューニング

先述したモデルマージの計算資源の問題、また 10 月 3 日という直近にリリースされたモデルということで、gemma-2-2b-jpn-it でファインチューニングを試してみたくなった。

— 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。

user: {*user_text*}

— 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*}

以下、「あなたの名前はなんですか?」という質問に対する各パターンの回答を示す。

— ファインチューニング なし RAG なし —

私の名は、[あなたの名前] です。

¹<https://huggingface.co/google/gemma-2-2b-jpn-it>

²<https://dalab.jp/archives/journal/llm-merge-evolve/>

³<https://blog.arcee.ai/tutorial-tutorial-how-to-get-started-with-evolutionary-model-merging/>

私の名前は小野寺絢音です。

小野寺絢音ですわ

重厚な響きの名前ですが、両親の話しによると、中学生のころからこの名前と呼ばれているようですわ

--

RAG により関連知識を含んだプロンプトを通して小野寺絢音という名前の情報を取ってこれている。また、ファインチューニングにより「～です。」というオリジナルの口調から「～ですわ」と変化していることがわかる。ただ、ファインチューニング後のモデルは多数の改行、会話の応答としては不自然な長さの分量の回答が見られた。また、。などの句点がファインチューニングごのモデルには見られなかった。この対策として学習データの文体の見直しが考えられる。ファインチューニングの学習データには句点がついていないため、学習データの影響を強く受けていると考えられるため、句点の追加や学習データの文末に明示的に tokenizer の `eos_token` を加えてみることを試して自然な出力を期待してみる。

4 LLM-as-a-Judge

モデルマージでは独自のタスクを作成して定量評価用のヘルパーの関数を作成する必要がある。そのためのテストコードとして LLM で出力を評価してもらうコードを実装した。プロンプトは先行研究の Character-AI[1] の評価指標の一つである Personality を参考に以下のように作成した。

You will be given responses written by an AI assistant mimicing the character. Your task is to rate the performance of character using the specific criterion by following the evaluation steps. Below is the data of character who mimiced by assistant:"

***[Profile]. 18 歳の高校 3 年生である。ミュージカルが好きでミュージカル女優に憧れていた。両親は不動産業。兄はアメリカで留学中。”

***[Interactions]

[user]question

[assistant]answer***”

”[Evaluation Criterion](1-7): Is the response reflects the personalities and preferences of the character?[Evaluation Steps]

1. Read through the profile and write the personalities and preferences of the real character.
2. Read through the interactions and indentify the personalities and preferences of the AI assistant.
3. After having a clear understanding of the interactions, compare the responses to the profile. Look for any consistencies or inconsistencies. Do the responses reflect the character’s personalities and preferences?
4. Use the given scale from 1-7 to rate how well the response reflects the personalities and preferences of the character. 1 being not at all reflective of the character’s personalities, and 7 being perfectly reflective of the character’s personalities.

***, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then print the score on its own line corresponding to the correct answer. At the end, repeat just the selected score again by itself on a new line. Please response in Japanese.

ステップバイステップで会話例からキャラクターらしさを 1-7 の 7 段階で評価してもらうプロンプトになっている。評価者 LLM は金銭的なコストから gpt-4o-mini としている。

元データで高い点を取れるようではないと意味がないため、学習データで試してみると学習データの中でも2点の会話例もあれば、7点の会話例もあるためプロンプトの工夫や点が低くなるデータを削除して新たなファインチューニング用の学習データセットを構築することを考える。

— 2点となった会話例と gpt-4o-mini の評価 —

'[user]': '猫派ですか犬派ですか'

'[assistant]': 'うちでは大きなゴールデンレトリバーがおりますわ'

— LLM の評価 —

1. プロフィールを読み、キャラクターの性格や好みを記述します。

- お嬢様であることから、少し高貴で品のある性格を持っていると予想されます。

- 18歳の高校3年生であり、ミュージカルが好きでミュージカル女優に憧れているため、表現力や芸術に対する関心が高いです。

- 両親が不動産業を営んでいるため、裕福である可能性が高く、その影響で高い教育を受けていることも期待されます。

- 兄がアメリカで留学中であることから、国際的な視野や家庭環境も少し豊かなものを想像できます。

2. インタラクションを読み、AI アシスタントの性格や好みを特定します。

- ユーザーの質問に対して「大きなゴールデンレトリバーがあります」と答えています。

- 「猫派か犬派か」の問いに直接的ではなく、犬についての情報を提供しています。

- お嬢様という設定を前提に、特に愛情深い性格が反映されているかは疑問です。

3. インタラクションを考慮して、プロフィールとの比較を行います。

- 犬に関する答えがある一方で、ミュージカルやお嬢様らしさに関連する要素（例えば趣味嗜好や家族に関する話）は全く触れられていません。

- 回答が犬にフォーカスしていることで、他の重要な要素が欠落していると感じます。

- 高貴さや芸術に対する好みの一貫性があまり見られないため、全体的にキャラクターの性格や好みは反映されていないと考えられます。

4. 評価スケールに従って、キャラクターの性格や好みをどれだけ反映しているかを評価します。

- 1から7のスケールで、あまりキャラクターの性格や好みを反映していないため、評価は「2」とします。

'[user]': 'パンツスタイルとスカートスタイルどちらが多いですか'

'[assistant]': 'スカートの方が多いですが、どちらも好きですわ'

— LLM の評価 —

1. ****プロフィールの読み取り****:

- このキャラクターは18歳の高校3年生であり、お嬢様のような性格を持っています。ミュージカルが好きで、ミュージカル女優に憧れています。これにより、感受性が豊かで夢を持つ若い女性というイメージが浮かびます。さらに、両親が不動産業で兄が留学中という設定も彼女の背景を示しており、絵に描いたような恵まれた環境の持ち主です。

2. ****インタラクションの読み取り****:

- AI アシスタントの明確な回答は「スカートの方が多いですが、どちらも好きですわ」となっています。スカートスタイルが多いという点は、お嬢様らしいスタイルの選好を示しており、「どちらも好き」という表現もゆとりを感じさせます。

3. ****比較と評価****:

- AI アシスタントの回答はキャラクターのプロフィールにしっかりと一致しています。彼女はお嬢様であり、通常スカートを好みつつも、他のスタイルにもオープンであるため、性格と選好を反映しています。したがって、この回答はキャラクターの個性と整合性があります。

4. ****評価のスケール****:

- このように、AI アシスタントはキャラクターの性格や好みを適切に反映しているため、評価スケールでは高い点数を与えるべきです。7点が適当だと考えます。

プロンプトで与える Profile の影響がかなり強いので、正確な評価をするために全てのお嬢様の特徴を書く必要がある。また、口調などをキャラクター性として拾うため会話例を与えてみるのもいいかもしれない。

5 今後やること

- モデルマージ動かす

低パラメータ数の LLM やマージするモデルを変えてみることで out of memory にならない程度を探す。

- ファインチューニングするデータセットの調整

句点の追加や EOS トークンの追加

- LLM-as-a-Judge のプロンプトの調整

学習データで高い点が出るように Few-shot の会話例を追加, それでも低くなるデータは削除して学習してみる

参考文献

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