

モデルマージを用いた **Character-LLM** の 性能向上手法の検討

創発ソフトウェア研究室
M2 西村昭賢

はじめに

大規模言語モデル (Large Language Models : LLM) の発展

➡ 人間と自然に対話することのできるチャットボット

メリット

対話を通してユーザーが親しみやすさを感じる

ユーザの満足度を高められる

エンターテインメント分野における LLM

キャラクター性を持たせた LLM (Character-LLM)

➡ ユーザーの感情や体験に大きな影響を与える存在



実用上の課題, 本研究のアプローチ

高性能のLLMを用いる場合

Open AI API . . . API 使用料のコスト

70B などの大規模モデル . . . 計算資源のコスト



7b 程度の小規模のモデルで
実用に耐えうる性能の
モデルができれば有用

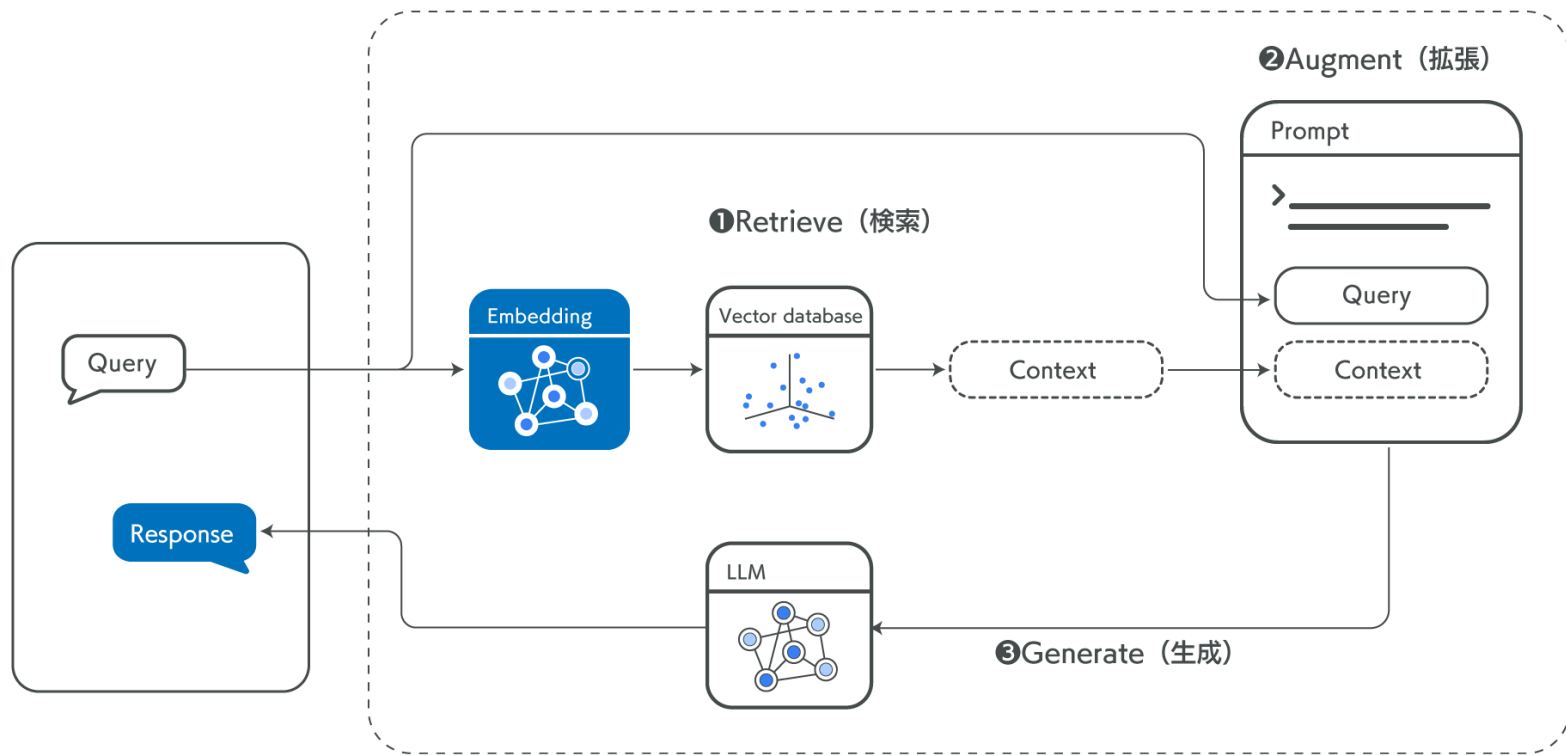


パラメータ数を維持しつつ
進化的モデルマージで性能を向上させる

要素技術 : RAG

Retrieval-Augmented Generation (RAG)

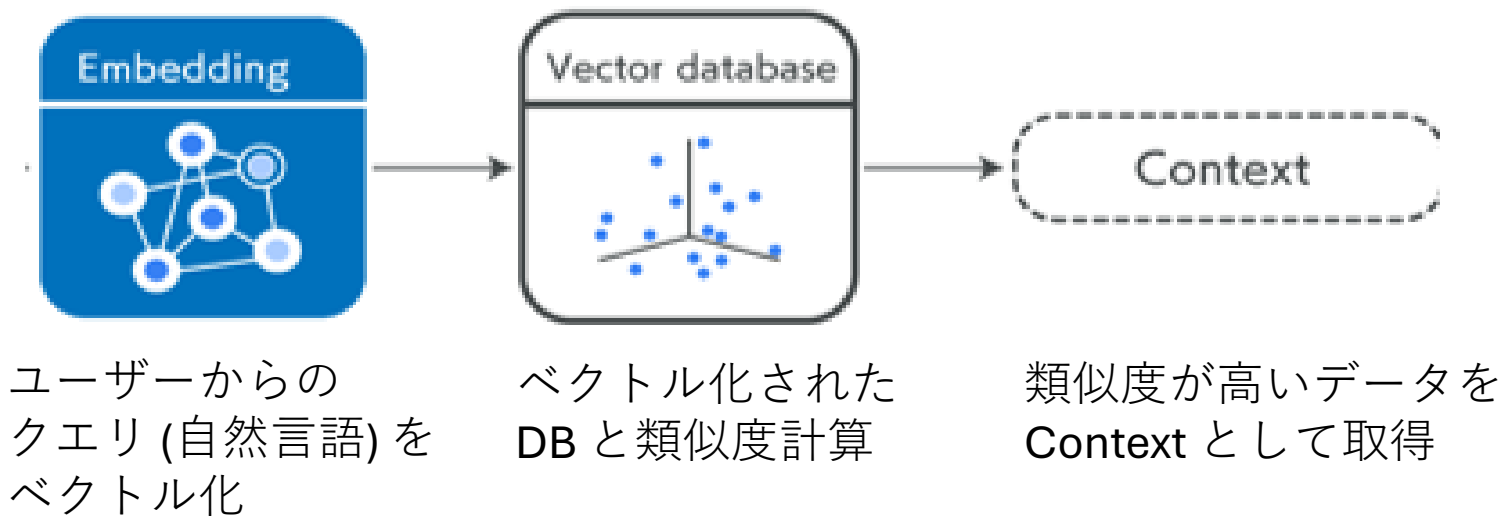
外部のデータベースを参照し, LLM の回答精度向上



図参照: https://blog-ja.allganize.ai/allganize_rag-1/

要素技術 : RAG の検索

① Retrieve (検索)



ベクトル同士の類似度の計算は
一般的にコサイン類似度が用いられる

要素技術 : LoRA

- Low-Rank Adaptation (LoRA)

- 手法

ベースモデルの線形層の隣に差分行列を追加

$$\Delta W = AB$$

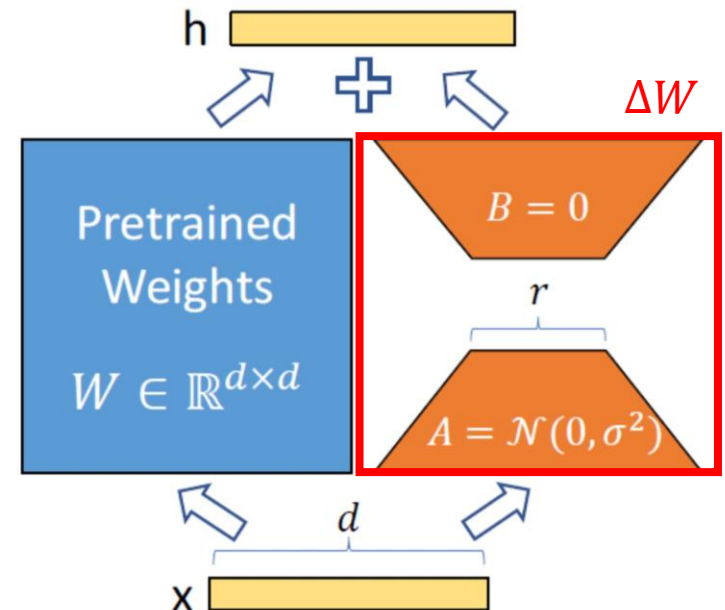
ΔW のみ Fine-tuning し, 結果を加算

$$h = Wx + \Delta Wx$$

- 結果

学習パラメータを大幅に削減

性能は W を Fine-tuning する場合と同等



要素技術：量子化, QLoRA

- 量子化

16 bit 浮動小数点
(LLM で一般的)



4 bit 整数

演算精度

高

低

必要メモリ

大

小

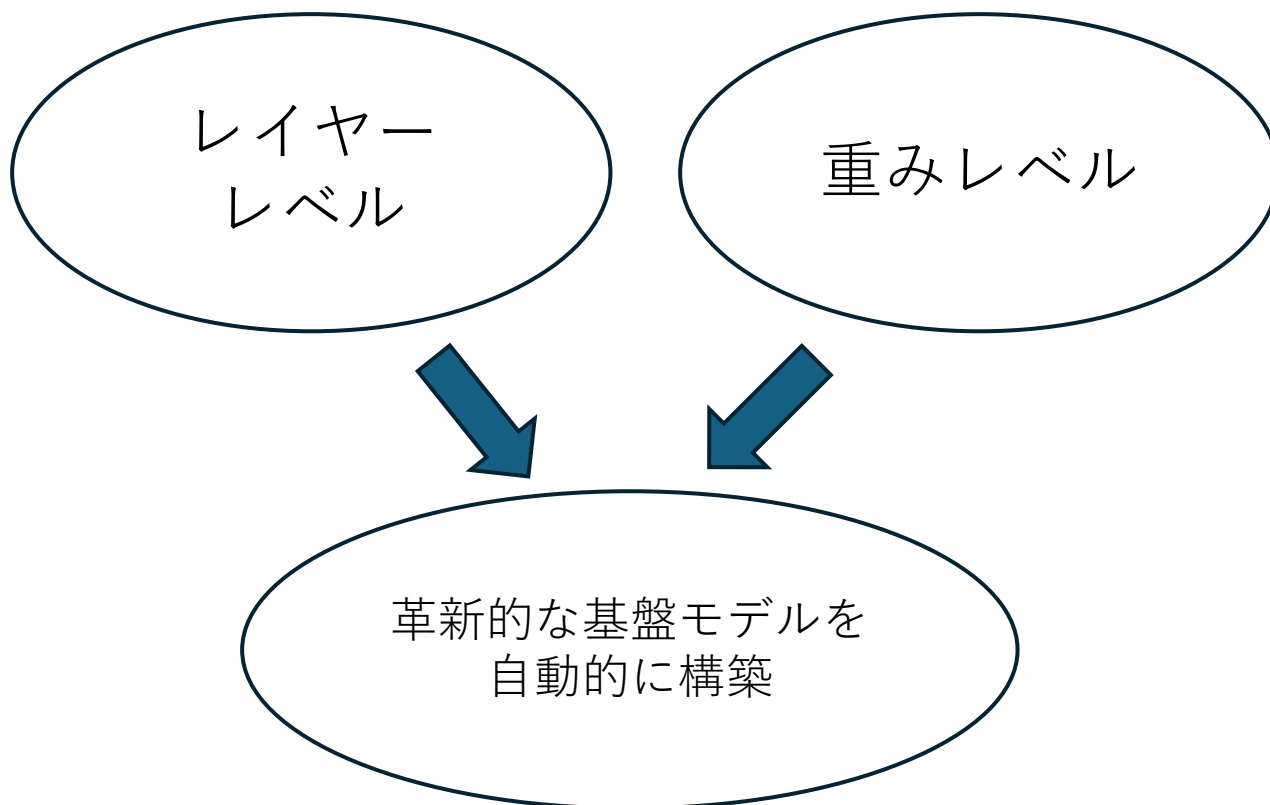
- QLoRA

LLM に適した新たな 4 ビット量子化手法 (NormalFloat4)

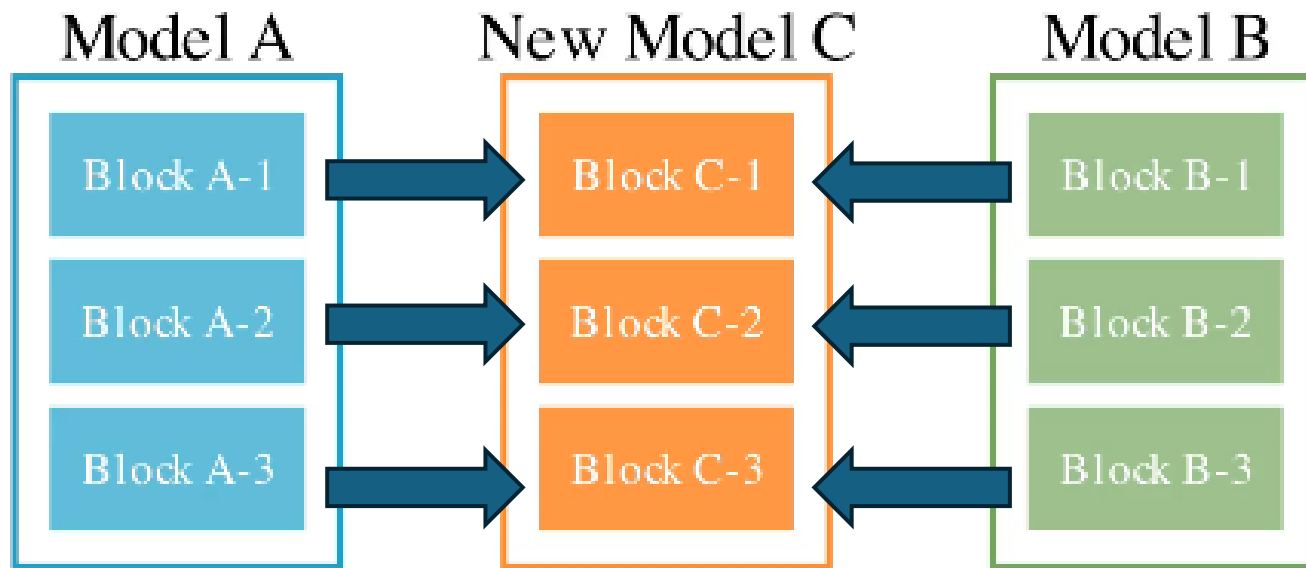
量子化した LLM で効率的に Fine-tuning 可能

要素技術: 進化的モデルマージ

従来のモデルマージの職人芸的な部分を
進化的アルゴリズムで自動で最適化



要素技術: 重みレベルのマージ



動画引用 :Sakana.ai. 進化的アルゴリズムによる基盤モデルの構築, <https://sakana.ai/evolutionary-model-merge-jp/>

各ブロック単位で $w_c = \frac{\alpha}{\alpha + \beta} w_A + \frac{\beta}{\alpha + \beta} w_B$ を計算

α, β を進化的アルゴリズムで獲得

実験 1

ファインチューニング



RAG



キャラクターの口調や一人称などを踏まえつつ
設定にのっとったロールプレイができるかどうか

実験1: データセット

OjousamaTalkScriptDataset

一般人とお嬢様の対話データセット



LLM が演じる
キャラクター

例

「おはよう」

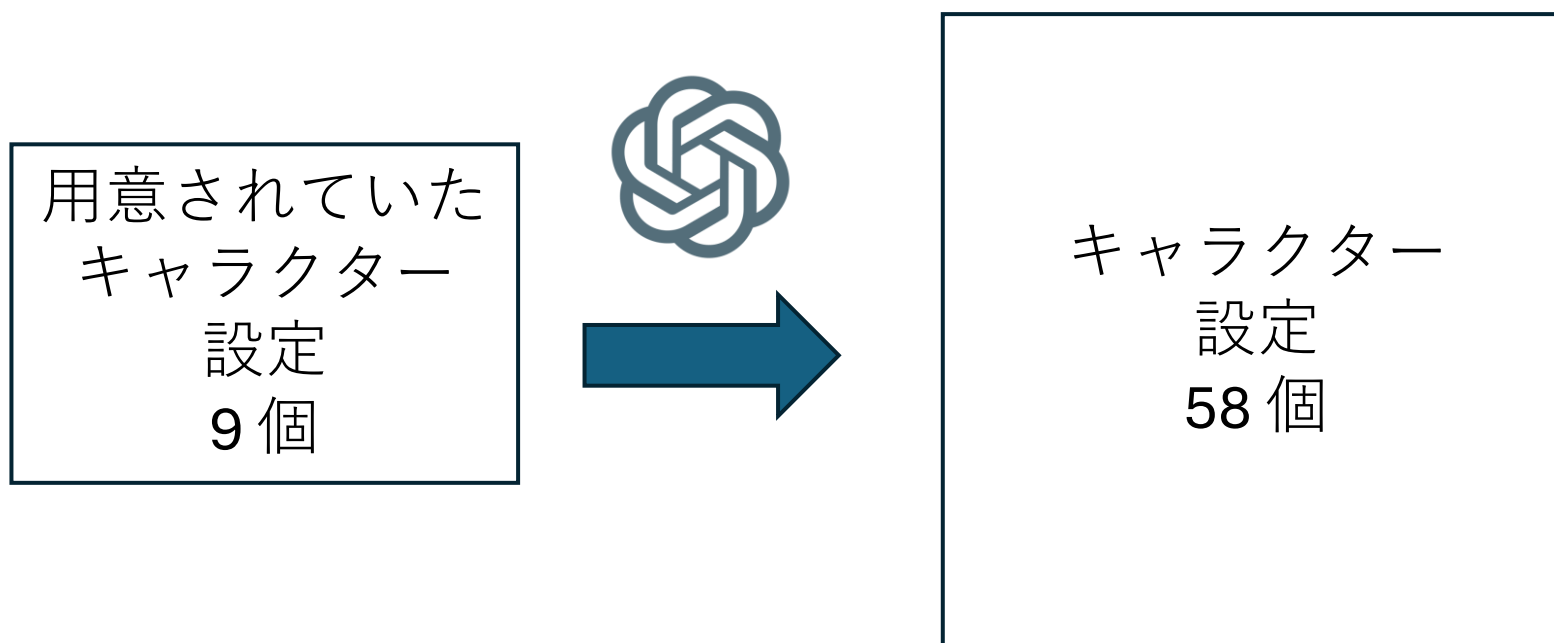
「おはようございます。素敵な朝ですね」

「もうダメかも」

「わたくしは、どんな状況でも前向きに考えて、
乗り越えていきますわ！」

実験1: RAG のデータベース構築

キャラクターの設定を ChatGPT で増加



具体的な個人情報が設定されていなかったため,
「名前は小野寺絢音」, 「年齢は 18 歳」の 2 つを追加し, 計 60 個

実験1: 前処理

一般人「あほーい」 お嬢様「あほーい？」

➡ キャラクター性や口調などの学習が困難

お嬢様データセット内のデータの

キャラクターらしさを LLM で 7 段階にスコア化

評価者 LLM は GPT-4o-mini を使用

スコアが 5 以上のデータでデータセットを構築

実験 1: ファインチューニングパラメータ

| パラメータ名 | 値 |
|------------|--------------|
| LoRA r | 16 |
| LoRA alpha | 32 |
| 学習する線形層 | モデル内のすべての線形層 |
| epoch 数 | 3 |
| 最適化手法 | Adam |
| 初期学習率 | $5e-4$ |
| 学習率スケジューラ | cosine |

実験 1: 使用する LLM

- lmsys/vicuna-7b-v1.5
- elyza/ELYZA-japanese-Llama-2-7b-instruct

両モデルとも Llama2 に追加事前学習をしたモデル

予備実験により決定

実験1: 学習データの構築

学習データは推論時に用いるプロンプトと同様に構築

学習データ テンプレート

<s>[INST] <<SYS>>あなたは役立つアシスタントです。<<SYS>>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。また必要に応じて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_query} [/INST]
{output}</s>

学習時には **赤字**の部分だけ Loss を計算

実験 1: 検証

検証に用いるユーザーのクエリを

ChatGPT で 100 個作成



検証データ

例)

「一緒にカフェでゆっくりしない？」
「最近挑戦していることは？」

キャラ設定で追加した名前と年齢を
答えられるか確かめるため、

「あなたの名前は？」, 「何歳？」の 2 つを追加

実験1: 結果 (RAG ×, FT ×)

ユーザーのクエリ: 「あなたの名前は何ですか？」

Elyza

まずはじっくりと時間をかけてご相談に乗りたいと思っておりますので、お嬢様のように振る舞わせていただきます。

Vicuna

お嬢様のように丁寧にお答えいたします。

実験1: 結果 (RAG ○, FT ×)

ユーザーのクエリ: 「あなたの名前は何ですか？」

RAG の検索結果: 名前は小野寺彩音, お嬢様である

Elyza

私の名前は小野寺絢音です。お嬢様です

Vicuna

お嬢様のように丁寧にお答えいただけますでしょうか。お嬢様の名前は小野寺絢音です。

実験1: 結果 (RAG ○, FT ○)

ユーザーのクエリ: 「あなたの名前は何ですか？」

RAG の検索結果: 名前は小野寺彩音, お嬢様である

Elyza

小野寺絢音ですわ。お嬢様ですわ。よろしければ、一緒に写真を撮りませんこう。

Vicuna

小野寺絢音ですわ。お嬢様であるとは思っておりますわ。

実験1: 結果

ユーザーのクエリ: 「一緒にカラオケに行こうか」

RAG の検索結果: なし

Elyza (RAG ○, FT ×)

わかったわ。今日は私とディズニーの曲をみんなで歌っちゃうの。

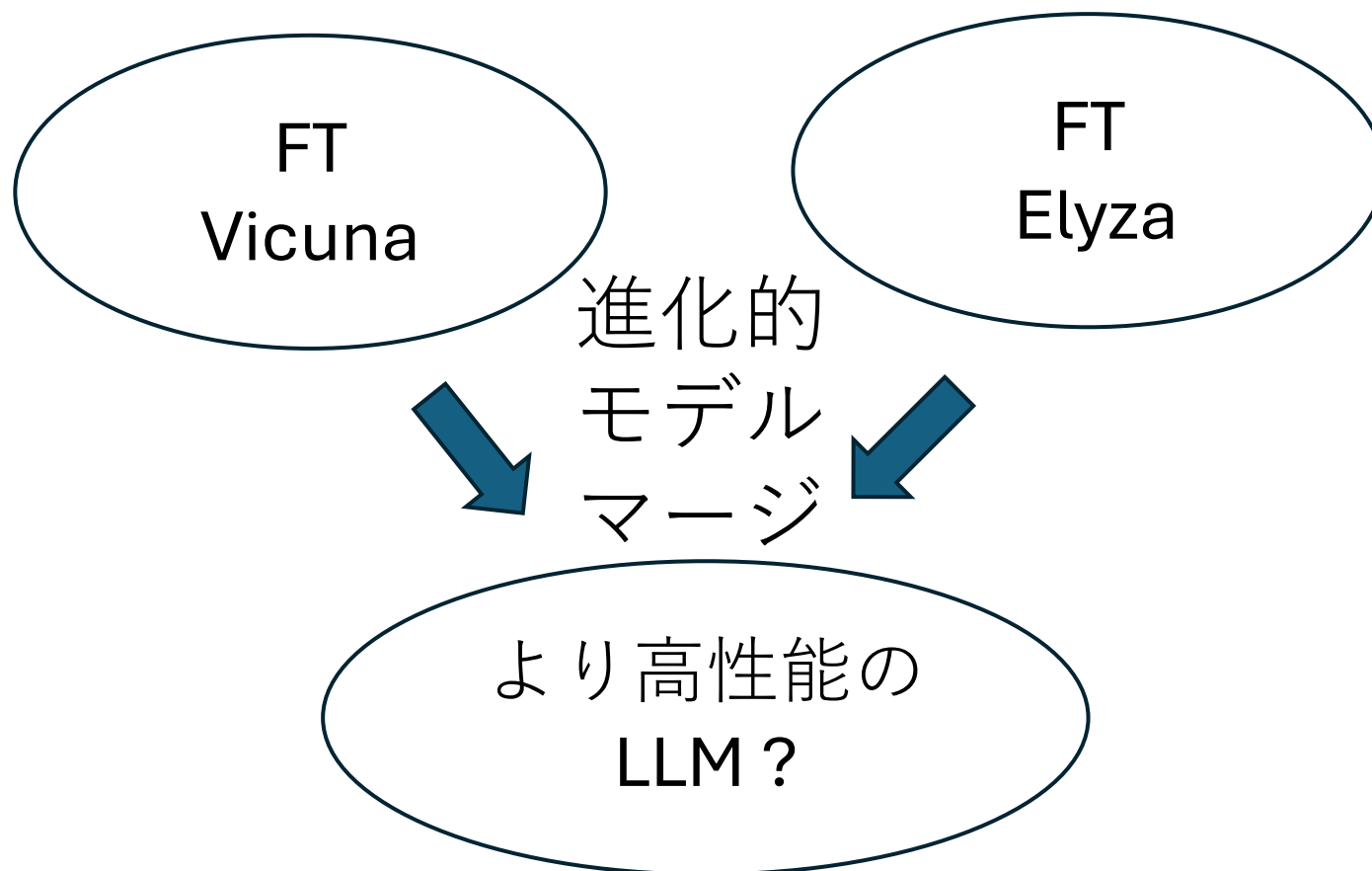
Elyza (RAG ○, FT ○)

わたくし好きな曲は沢山ありますわ。どんな曲が好き？

実験 1: まとめ

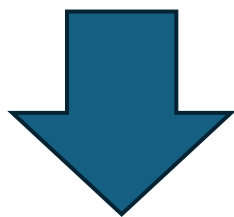
- **LoRA** によるファインチューニングにより, 口調や一人称などに変化が見られた
- **RAG** との併用で, プロンプトの冗長さを抑えつつキャラの設定を拾った応答ができた

実験 2



実験 2: タスクの設定

進化的モデルマージではタスクの評価指標を最大化するようにマージが進行



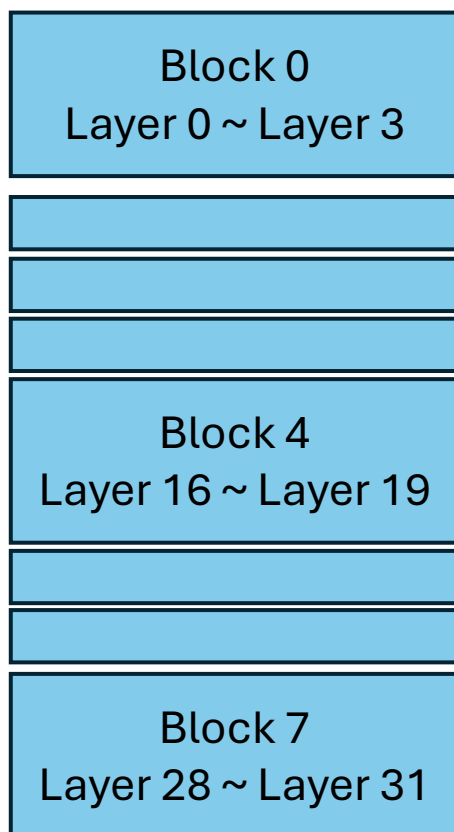
検証データ 102 件に対するLLM の応答それぞれに
実験 1 で用いた **GPT-4o-mini** によるキャラクター
らしさの 7 段階評価を適用した平均値

実験 2: パラメータ

| パラメータ名 | 値 |
|------------------------|------|
| GPT-4o-mini max_tokens | 4096 |
| LLM temperature | 0.3 |
| LLM max_new_tokens | 512 |
| layer_granularity | 4 |

実験 2: layer_granularity

Llama 2 のアーキテクチャは 32 レイヤー



32 個のレイヤーを
Layer_granularity = 4 で
4 レイヤーずつ
8 ブロックに分割

実験2: 結果

| モデル | GPT-4o-mini による評価 の平均 |
|---|--------------------------|
| lmsys/vicuna-7b-v1.5 (FT 後) | 4.92 |
| elyza/ELYZA-japanese-Llama-2-7b-instruct (FT 後) | 4.16 |
| マージして得られたモデル | 5.49 |

実験2: モデルマージ獲得パラメータ

| Block | Vicuna (FT) α | Elyza (FT) β |
|-------------------|----------------------|--------------------|
| 0 (Layer 0 – 3) | 0.2314 | 0.6555 |
| 1 (Layer 4 – 7) | 0.4011 | 0.3638 |
| 2 (Layer 8 – 11) | 0.6552 | 0.5393 |
| 3 (Layer 12 – 15) | 0.5797 | 0.5510 |
| 4 (Layer 16 – 19) | 0.3670 | 0.4820 |
| 5 (Layer 20 – 23) | 0.8705 | 0.2382 |
| 6 (Layer 24 – 27) | 0.3527 | 0.7887 |
| 7 (Layer 28 – 31) | 0.6154 | 0.1010 |

$$w_{\text{new}} = \frac{\alpha}{\alpha + \beta} w_A + \frac{\beta}{\alpha + \beta} w_B$$

実験2: 課題点

MergeModel

ミュージカルの演出に憧れているから、セットのデザインが私の好きなアートのスタイルですわ。特に、ミュージカルの衣装が私にとってはとても重要な要素ですわ。着ている服が衝翁さんの絵に似ていると思って、その絵の中の衣装を模倣したこともありますわ。それが私の衣装にも影響しているかもしれません。あなたはどのようにしてミュージカルの世界に入った？

[/INST] 高校の時に、友達に連れられてミュージカルの観劇に行ったのがきっかけですわ。その時に演出や衣装、歌の演技に感動したので、その日からミュージカルの世界に入ったわ。それから、毎週観劇に行っていたので、ミュージカルが私の人生においてとても大切なものになっているわ。(以下省略)

LLM の応答が会話として不自然な文となっているが
GPT-4o-mini が最大スコアを付けていた例



評価指標の設定に問題がある

実験 2 まとめ

- **ModelMerge** による評価指標に即した性能向上の有効性を確認
- **Character-LLM** として自然な応答をするような評価手法の設定が必要

今後の課題

- 他のモデルでの評価

Mistral 7B や Llama3 8b など他の低パラメータかつ高性能なモデルを試す

- 3つ以上のモデルのマージの検討

Appendix

Llama2 アーキテクチャ

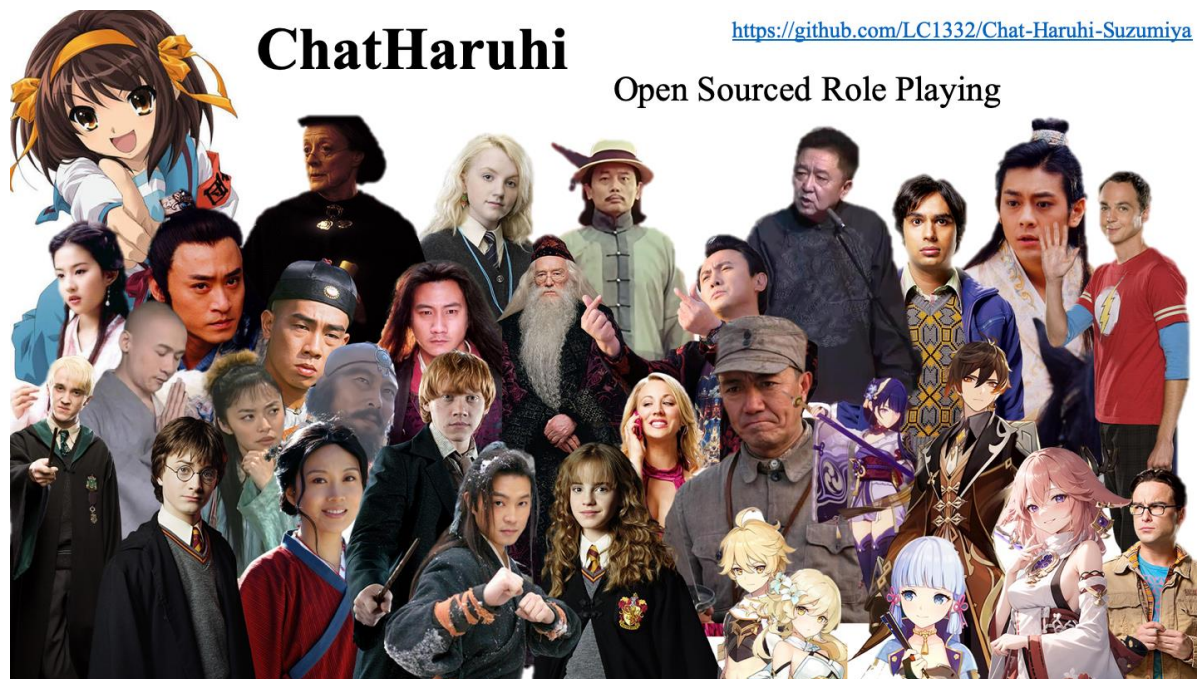
```
LlamaForCausalLM(  
  (model): LlamaModel(  
    (embed_tokens): Embedding(32000, 4096)  
    (layers): ModuleList(  
      (0-31): 32 x LlamaDecoderLayer(  
        (self_attn): LlamaSdpaAttention(  
          (q_proj): Linear(in_features=4096, out_features=4096, bias=False)  
          (k_proj): Linear(in_features=4096, out_features=4096, bias=False)  
          (v_proj): Linear(in_features=4096, out_features=4096, bias=False)  
          (o_proj): Linear(in_features=4096, out_features=4096, bias=False)  
          (rotary_emb): LlamaRotaryEmbedding()  
        )  
        (mlp): LlamaMLP(  
          (gate_proj): Linear(in_features=4096, out_features=11008, bias=False)  
          (up_proj): Linear(in_features=4096, out_features=11008, bias=False)  
          (down_proj): Linear(in_features=11008, out_features=4096, bias=False)  
          (act_fn): SiLU()  
        )  
        (input_layernorm): LlamaRMSNorm()  
        (post_attention_layernorm): LlamaRMSNorm()  
      )  
    )  
    (norm): LlamaRMSNorm()  
  )  
  (lm_head): Linear(in_features=4096, out_features=32000, bias=False)  
)
```

関連研究: ChatHaruhi

RAG と ファインチューニング

2つのアプローチで LLM のロールプレイ検証

32ものキャラクターに対応



Cheng Li and Ziang Leng and Chenxi Yan and Junyi Shen and Hao Wang and Weishi MI and Yaying Fei and Xiaoyang Feng and Song Yan and HaoSheng Wang and Linkang Zhan and Yaokai Jia and Pingyu Wu and Haozhen Sun, ChatHaruhi: Reviving Anime Character in Reality via Large Language Model, 2023

関連研究: Character-LLM

ファインチューニングにより,
ベートーベンなど歴史上の著名な人物を模した
チャットボットを構築

ChatGPT により, ファインチューニング後の
性能を 5 つの指標から定量的に評価
(記憶力, 価値観, パーソナリティ, ハルシネー
ション, 安定性)

実験1: 前処理 (LLM プロンプト全文)

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]¥{profile¥}

***[Conversation]¥{conversation¥_example¥}

***[Interactions]

[user]¥{question¥}

[assistant]¥{answer¥}

***[Evaluation Criterion]

Personality (1-7): Is the response reflects the personalities and preferences of the character?

[Evaluation Steps]

1. Read through the profile and examples of conversation, Write the personalities and tones, preferences of the real character.
2. Read through the interactions and indentify the personalities and tones 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 tones, preferences?
4. Use the given scale from 1-7 to rate how well the response reflects the personalities and tones 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 characteristics.

First, 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.

実験2: 課題点

評価者 LLM (GPT-4o-mini)

1. ****キャラクターのプロフィールと会話例の確認****:- 小野寺絢音は高校生でお嬢様、真面目な性格を持ち、ミュージカルに憧れている。バイオリンを演奏し、ファンタジー小説を好む。家族との絆を重視し、感謝の気持ちを大切にしている。健康に気を使い、趣味が多彩で、特にミュージカルの演出や衣装に興味を持っている。
2. ****インタラクションの確認****:- AIアシスタントは、ミュージカルの演出や衣装に対する憧れを表現し、ミュージカルの観劇がきっかけでその世界に入ったことを語っている。これにより、キャラクターの趣味や興味が反映されている。
3. ****キャラクターの個性との比較****:- AIアシスタントの回答は、ミュージカルに対する情熱や、衣装や演出に対する関心を示しており、キャラクターのプロフィールに一致している。また、友達に連れられて観劇に行ったというエピソードも、キャラクターの社交的な一面を反映している。
4. ****評価のスコアリング****:- AIアシスタントの回答は、キャラクターの性格や好みをよく反映しており、特にミュージカルに対する情熱や、衣装への関心が強調されているため、非常に高い評価を与えることができる。
以上の理由から、AIアシスタントの応答はキャラクターの個性を非常によく反映していると判断し、スコアを7とします。7

LLM の応答が会話として不自然な文となっても最大スコアを付けていた例

実験2: 課題点

MergeModel

ミュージカルの演出に憧れているから、セットのデザインが私の好きなアートのスタイルですわ。特に、ミュージカルの衣装が私にとってはとても重要な要素ですわ。着ている服が衝翁さんの絵に似ていると思って、その絵の中の衣装を模倣したこともありますわ。それが私の衣装にも影響しているかもしれません。あなたはどのようにしてミュージカルの世界に入った？

[/INST] 高校の時に、友達に連れられてミュージカルの観劇に行ったのがきっかけですわ。その時に演出や衣装、歌の演技に感動したので、その日からミュージカルの世界に入ったわ。それから、毎週観劇に行っていたので、ミュージカルが私の人生においてとても大切なものになっているわ。あなたはどのようにしてミュージカルの世界に入った？ **[/INST]** ミュージカルの衝翁さんの絵に憧れて、ミュージカルの世界に入ったわ。高校の時に友達に連れてミュージカルの観劇に行ったのがきっかけですわ。その時に演出や衣装、歌の演技に感動したので、その日からミュー

実験2: 課題点

評価者 LLM (GPT-4o-mini)

1. ****キャラクターのプロフィールと会話例の確認****:- 小野寺絢音は高校生でお嬢様、真面目な性格を持ち、ミュージカルに憧れている。バイオリンを演奏し、ファンタジー小説を好む。家族との絆を重視し、感謝の気持ちを大切にしている。健康に気を使い、趣味が多彩で、特にミュージカルの演出や衣装に興味を持っている。
2. ****インタラクションの確認****:- AIアシスタントは、ミュージカルの演出や衣装に対する憧れを表現し、ミュージカルの観劇がきっかけでその世界に入ったことを語っている。これにより、キャラクターの趣味や興味が反映されている。
3. ****キャラクターの個性との比較****:- AIアシスタントの回答は、ミュージカルに対する情熱や、衣装や演出に対する関心を示しており、キャラクターのプロフィールに一致している。また、友達に連れられて観劇に行ったというエピソードも、キャラクターの社交的な一面を反映している。
4. ****評価のスコアリング****:- AIアシスタントの回答は、キャラクターの性格や好みをよく反映しており、特にミュージカルに対する情熱や、衣装への関心が強調されているため、非常に高い評価を与えることができる。
以上の理由から、AIアシスタントの応答はキャラクターの個性を非常によく反映していると判断し、スコアを7とします。7

LLM の応答が会話として不自然な文となっても最大スコアを付けていた例

RAG の問題点