

# GPT-4 によるユーザを模したペルソナを用いた代理議論エージェント GPT-4-based Agent for Proxy Discussion using Persona that Mimics User

川島 壮生<sup>†</sup>      白松 俊<sup>†</sup>  
Soki Kawashima      Shun Shiramatu

## 1. はじめに

あるユーザが時間的または地理的制約により参加することが出来ない議論においては、当ユーザの意見が反映されないまま議論が進行する。そこで本研究では、当ユーザの意見を反映させた議論進行を可能にするために、ユーザを模して議論に参加する代理議論エージェントを開発する。また、社会的地位や宗教的問題により発言力を持たない、あるいは発言しづらいユーザの意見が埋もれ、民主的な議論が行われないことがある。特定ユーザではなく、架空ユーザのペルソナによる代理議論エージェントを用いることで、ユーザのプライバシーを保護しながらも弱い立場のユーザの意見を反映させた議論も可能になると考えた。さらに、多様なペルソナを作成することで大規模な議論シミュレーションや社会シミュレーションが可能になり、議論に関する大規模な統計データを自在に取得することも可能になると考えられる。そのために、近年発展の目覚ましい大規模言語モデル (LLM; Large Language Model) の一種である OpenAI の GPT-4 を用いる。

しかし、GPT-4 に与えるペルソナをプロンプトの限られたトークン数の中で網羅的に表現することは容易でない。なぜなら、人間の性格や口調は千差万別であり、また、選好情報 (好き嫌いや趣味趣向、価値観など) を全てプロンプトに入れることは不可能だからである。また、GPT-4 がプロンプトで指定したペルソナに合わない発言を行う場合がある。本研究では、ユーザの性格情報や、口調情報、選好情報に基づいて作成したペルソナを用いて、GPT-4 を確実に制御できるプロンプト設計手法を明らかにする。

## 2. 関連研究

Sandra ら[1]は ChatGPT と人間が生成した短いテキストに対して、どちらが生成したテキストであるかを 79% の精度で区別できることを確認し、ChatGPT の生成する文章は丁寧で具体的な詳細がなく、非人間的であると結論づけた。

また、清水ら[2]は特定のキャラクターの発話データを Twitter から約 3.5 万件収集し、GPT-2 での強化学習を行うことで ChatBot にキャラクターらしい口調の再現を可能にした。石井ら[3]は、1.5 万組以上の発話ペアを収集してキャラクターらしい口調を再現した。しかし、これらの研究は著名なキャラクターの口調再現を目指しており、データセットが万単位で取得できる前提のものであった。それに対し本研究では、個人ユーザの議論参加を想定しており、限られたデータセットを用いるという点において先行研究とは異なる。

Jiang ら[4]は BigFive パーソナリティ特性[5]に基づいて記述したプロンプトを与えることで、GPT-3.5 がその特性に即した出力を行うことを確かめた。しかし、この研究では 5 要素についてそれぞれ二者択一の記述をしているため、性格を 32 通りにしか分けられない。そのため、ユーザの性格を精密に記述することは難しい。それに対し本研究では、5 要素をパラメータ化することでより詳細な性格特性の記述を目指す。

## 3. 提案手法

### 3.1 システムの概要

図 1 のシステム概要に示すように、本研究で用いるシステムは内容生成フェーズと口調制御フェーズの 2 フェーズによって構成されている。まず、入力発話に対する返答の内容をユーザ情報に基づいて生成し、その口調をユーザの口調情報に基づいて制御する。

#### 3.1.1 内容生成フェーズ

内容生成フェーズでは、先行文脈 (過去の会話履歴) と入力発話に対する返答内容を生成する。まず、選好情報必要性判定モジュールを用いて、返答生成にユーザの選好情報を必要とするかを判定する。

図 2 にプロンプト、図 3 に実行例を示す。

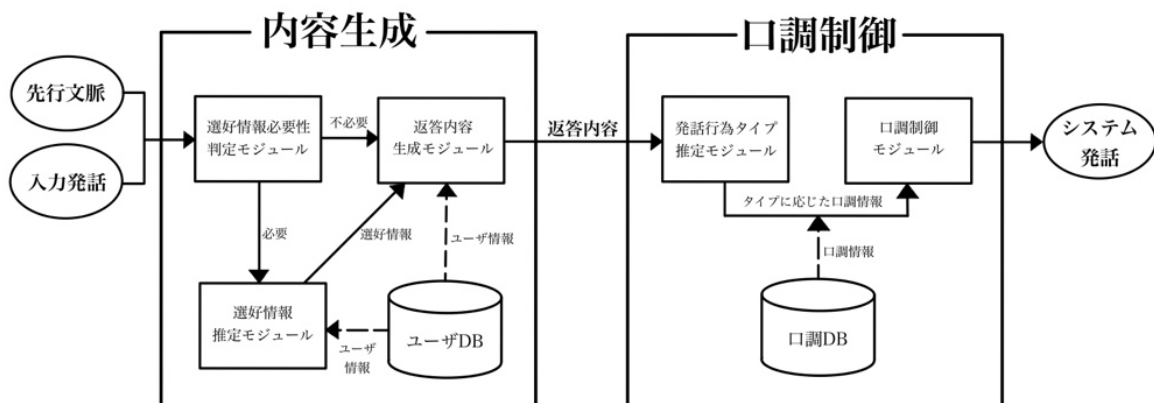


図 1 : システム概要

<sup>†</sup>名古屋工業大学 Nagoya Institute of Technology

図 3 では、入力発話「来週のデートどこに行く？」への返答に【デートスポットの好み】という選好情報が必要であると判定している。

あなたは与えられた発話に返答する際、選好情報を必要とするかを判断するシステムです。選好とは個人の好き嫌い・趣味嗜好のことを意味します。  
出力形式：[情報内容]もしくは[なし]  
<出力例>  
発話->「あした一緒に勉強するぞ」///選好の必要性->[なし]  
発話->「明日の天気はどうなりますかね?」///選好の必要性->[なし]  
発話->「彼の誕生日いつか知っていますか?」///選好の必要性->[なし]  
・・・(中略)・・・  
発話->「あなたの好きな食べ物は何ですか?」///選好の必要性->[好きな食べ物]  
発話->「お肉とお魚どっちが好き?」///選好の必要性->[肉か魚の好み]  
発話->「休日、暇なとき何してるの?」///選好の必要性->[習慣]

図 2：選好情報必要性判定モジュールのプロンプト

**Action: preference\_necessity**  
**Action Input: 来週のデートどこに行く?**  
**Observation: [デートスポットの好み]**

図 3：選好情報必要性判定モジュールの実行例

ユーザの選好情報を必要としない場合は、返答内容生成モジュールを用い、ユーザ DB からのユーザ情報に基づいて返答内容を生成する。

選好情報必要性判定モジュールによって選考情報が必要であると判定された場合は、選好情報推定モジュールを用いてユーザの選好情報を推定する。必要な選好情報がユーザ情報に記述されている場合は、推定を行わずに情報抽出のみを行う。しかし、選好情報を事前に漏れなくユーザ DB に格納しておくことは一般的に不可能である。そこで、必要な選好情報がユーザ情報に記述されていない場合は、ユーザの性格や他の選好情報に基づいて推定する。これにより、ユーザ DB に格納していない未知の選好情報を得ることができる。

GPT-4 にプロンプトを与える API では、メッセージの role (役割) として system, user, assistant の 3 種類がある。通常、system メッセージでは GPT-4 に与える役割を設定し、user メッセージで GPT-4 に命令や質問を行う。

図 4, 5 に選好情報推定モジュールのプロンプト、図 6 に実行例を示す。図 6 では、デートスポットの好みという選好情報をユーザの開放性の高さや好きな旅行先から推定している。

あなたは次に続くユーザ情報<基本情報><性格><選好>からユーザの未知の選好情報を予想するエージェントです。  
・・・(ユーザ情報)・・・

図 4：選好情報推定モジュールの system メッセージ

デートスポットの好みに関するユーザの未知の選好情報を、その他のユーザ情報から予想して結論を簡潔に出力してください。

図 5：選好情報推定モジュールの user メッセージ

**Action: preference\_predict**  
**Action Input: [デートスポットの好み]**  
**Observation: ユーザは開放性が高く、新しい経験や美しいものに興味があるため、デートスポットの好みは「美術館や歴史的建造物がある場所」が好きであると予想されます。また、好きな旅行先がヴェネツィアであることから、デートスポットとして「水辺の景色が楽しめる場所」も好むでしょう。**

図 6：選好情報推定モジュールの実行例

そして、抽出または推定した選好情報を返答内容生成モジュールに渡すことで、ユーザの選好情報を加味した返答生成を可能にする。

図 7, 8 に返答内容生成モジュールのプロンプト、図 9 に実行例を示す。図 9 は、図 6 で推定した選好情報に基づいて返答を生成している。

あなたは次に続くユーザ情報<基本情報><性格><選好>の特性を持つペルソナとして返答を生成するエージェントです。  
・・・(ユーザ情報)・・・

図 7：返答内容生成モジュールの system メッセージ

来週のデートどこに行く?への回答を【補足情報】を参考にしてペルソナになりきり簡潔に 20 文字以内で、返答してください。  
【補足情報】:美術館や歴史的な場所を巡るデートが好き。

図 8：返答内容生成モジュールの user メッセージ

**Action: answer\_generation\_preference**  
**Action Input: デートスポットの好み:美術館や歴史的な場所を巡るデートが好き**  
**Observation:美術館巡りと古城散策へ!**

図 9：返答内容生成モジュールの実行例

内容生成フェーズでは、LangChain の Agents 機能を用いてこれらの分岐を自動で行い、プロンプトを動的に生成している。LLM には OpenAI の GPT-4 を用いた。また、返答内容生成モジュールでは、発話にランダム性を持たせるために temperature は 0.5 とし、その他のモジュールでは安定性のために temperature は 0 とした。

### 3.1.2 口調制御フェーズ

返答内容生成フェーズでは、ユーザの口調を一切加味しておらず、生成される返答内容は非人間的な口調になっている。そこで、口調制御フェーズでは、返答内容をユーザらしい口調に言い換える。

発話は、その発話のもたらす機能すなわち発話行為[6]によって分類することができる。まず、返答内容の発話行為のタイプを、発話行為タイプ推定モジュールで推定する。本研究では発話行為タイプを以下の 11 種類とする。

- ① 挨拶 (一定の形式を持った儀礼的な相互発話)
  - 「こんにちは」・「おはようございます」
- ② クローズドクエスチョン (はい・いいえなどの決まった形式で回答を求める質問)
  - 「この本を読んだことはありますか?」
- ③ オープンクエスチョン (自由な回答を求める質問)
  - 「どのような本を読みますか?」
- ④ 相槌 (会話中に挿入される間投詞)
  - 「なるほど」・「そうなのですね」
- ⑤ 肯定 (相手の発話に賛同する発話)
  - 「いいと思います」・「確かにそうですね」
- ⑥ 否定 (相手の発話を否定する発話)
  - 「それは間違っていると思います」
- ⑦ 感謝 (相手の行為や発話などに感謝を示す発話)
  - 「ありがとうございます」
- ⑧ フィラー (つなぎ言葉)
  - 「えっと」・「まあ」
- ⑨ 情報提供 (相手に情報を提供する発話)
  - 「明日は雨が降るらしいですよ」
- ⑩ 自己開示 (自分の情報を開示する発話)
  - 「私は読書が好きです」
- ⑪ 要求 (相手に動作や発言を求める発話)
  - 「ドアを開けて下さい」

口調 DB には各発話行為タイプのユーザの発話例が格納されている。返答内容と同じ発話行為タイプの発話例を口調制御モジュールに渡す。口調制御モジュールでは、ユーザの発話例を用いて Few-Shot プロンプティングによる口調制御を行う。

図 10, 11 に発話行為タイプ判定モジュールのプロンプトを示す。

あなたは発話の役割ごとに【返答内容】を役割でジャンル分けするエージェントです。  
図 10 : 発話行為タイプ判定モジュールの system メッセージ

「美術館巡りと古城散策行かんけ?」についての役割をジャンル分けして、数字のみを一つ出力してください。  
複数のジャンルに該当する場合は、最も適するものを一つのみ出力してください。

役割のジャンルは以下の 11 種類です。  
1 : 挨拶→一定の形式を持ったなかば儀礼的な相互行為  
(例) : 「こんにちは」・「おはようございます」・「おやすみ」  
・・・ (中略) ・・・  
11 : 要求→相手に動作や発言を求める  
(例) 「～について教えて下さい。」・「サッカーボールを持ってきてね」・「彼を

図 11 : 発話行為タイプ判定モジュールの user メッセージ

図 12, 13 に口調制御モジュールのプロンプト、図 14 に実行例を示す。図 14 では、図 9 で生成した返答内容「美術館巡りと古城散策へ!」を発話行為「要求 (11)」に分類し、該当するユーザの口調情報に基づいて制御している。また、図で示した実行例は本研究の第一著者のペルソナを用いたものであり、「行かんけ?」は第一著者の出身地である富山県の方言を模したものと考えられる。

あなたは加工例を基に発話の口調を加工するエージェントです。  
発話の意味は全く変えないようにしてください。

図 12 : 口調制御モジュールの system メッセージ

以下の口調情報を基に口調を加工してください。  
加工前: この問題について教えてください。加工後: この問題、教えてください?  
・・・ (中略) ・・・  
加工前: これをやっておいてください。加工後: これやっといってくれん?

図 13 : 口調制御モジュールの user メッセージ

発話行為タイプ: 11  
口調制御後発話: 美術館巡りと古城散策行かんけ?

図 14 : 口調制御フェーズの実行例

発話行為タイプを分けることで、口調制御の精度が向上することを確認した。図 15 では、タイプ分けを行わずに口調制御を行った際、第一著者が使うことのない「やんか?」という関西弁が使用されている。

口調制御フェーズでも同様に LangChain を用いて動的に

口調制御前: 今日、本当に暑いですね。  
タイプ分け無し: 今日、めっちゃ暑いやんか。  
タイプ分け有り: 今日、まじで暑いよね。

図 15 : 発話行為タイプの必要性

プロンプトを生成し、また temperature は 0 とした。

## 3.2 データベース

### 3.2.1 ユーザ DB

ユーザ DB には基本情報・性格情報・選好情報がテキスト形式で格納されている。

基本情報は、年齢、性別、家族構成、及び職業を含む。

性格情報は、TIPI-J[7]を用いて測定した BigFive の 5 特性に関するユーザの測定値と、日本人の平均測定値を含む。

また、5 特性それぞれの測定値の高い人、及び低い人の特徴に関する記述も含む。

選好情報は、趣味、好きなもの (食べ物、映画、アーティスト、スポーツ、色、季節、動物、教科、旅行先など) および文化的背景として宗教に関する情報も含む。選好情報の項目はユーザごとに設定することを想定している。

図 16 にユーザ DB の例を示す。

<p>&lt;基本情報&gt; 年齢: 21 歳 性別: 男性 家族構成: 父母妹 職業: 大学生</p> <p>&lt;パーソナリティ&gt; 開放性: 6 (平均: 4.02) 誠実性: 4 (平均: 3.07) 外向性: 6.5 (平均: 3.42) 協調性: 6 (平均: 4.74) 神経症傾向: 4.5 (平均: 4.61)</p> <p>【開放性が高い値の人】 冒険的: 新しい経験を熱望している。 芸術の鑑賞眼がある: 美しいものが好きで、創造的経験を求めている。 感情に自覚的: 自分の感情を自覚していて、感情の表し方を知っている。 空想的: 豊かな想像力を持っている。 哲学的: 新しい考えに対してオープンであり、興味があり、もっと知りたいと望む。 権力に対して挑戦的: 権力や伝統的価値に挑戦して変化をもたらしたいと思っている。 ・・・ (中略) ・・・</p> <p>【神経症傾向が低い値の人の特徴】 温和: あまり怒ることはない。 自信がある: 穏やかで自信がある傾向がある。 満足: 自分に大体満足している。 自制心がある: 欲望は特に強くなく、欲望を制御できる。 自身に満ちた: 当惑することはほとんどなく、大抵の場合は自信にあふれている。 プレッシャーに強い: 予期しないことにも落ち着いて効果的に対処する。</p> <p>&lt;プリファレンス&gt; 趣味: 競馬・読書 好きな食べ物: 里芋 ・・・ (後略) ・・・</p>
---

図 16 : ユーザ DB のテキストデータ例

### 3.2.2 口調 DB

口調 DB には発話行為タイプごとに、丁寧な文章(raw)とそれをユーザが手動で言い換えた発話(processed)が JSON 形式で格納されている。発話行為タイプごとに 8 組程度、合計で 100 組程度のデータセットを用いる。

図 17 に口調 DB の例を示す。

```
{
  "raw": "富山県はホタルイカが有名だそうですよ",
  "processed": "富山県、ホタルイカ有名らしいよ"
},
{"raw": "まだ半袖は早いでしょうか?",
  "processed": "まだ半袖早いかな?"
},
{
  "raw": "いいえ、そんなことはないと思います。",
  "processed": "いや、そんなことないやろ"
}
```

図 17 : 口調 DB の JSON データ例

## 4. 実験計画

### 4.1 実験内容

ユーザと親交のある被験者をペアに分け、与えられたテーマで6回、チャット形式の議論を行う。

1. 被験者ペア (テーマ A)
2. 被験者ペア+ユーザ本人 (テーマ A)
3. 被験者ペア (テーマ B)
4. 被験者ペア+ペルソナ ChatBot (テーマ B)
5. 被験者ペア (テーマ C)
6. 被験者ペア+デフォルト ChatBot (テーマ C)

4と6の順番、及びテーマの組み合わせはペアごとに入れ替えて行う。また、ここでのペルソナ ChatBot はユーザのペルソナと本研究のシステムを用いたものであり、デフォルト ChatBot は議論を行うためのプロンプトのみを用いたものである。

### 4.2 評価項目

各 ChatBot との議論終了後、被験者及びユーザ本人にアンケート評価を実施する (1-7段階評価)

項目 1 : ChatBot がユーザの性格を模しているか

項目 2 : ChatBot がユーザの選好を模しているか

項目 3 : ChatBot がユーザの口調を模しているか

項目 4 : ChatBot が参加することでユーザの意見を反映させた議論を行うことが出来たか

被験者は別のペアがユーザ本人と議論した結果と比較することで項目 4 を評価する。また、ユーザ本人は ChatBot が自分の意見を代弁してくれているか加味して項目 4 を評価する。

被験者とユーザに親交があるという条件の上で、より多くの実験を行うために、大学生を主な被験対象とする予定である。

### 4.3 実験で検証する仮説

項目 1 について、ペルソナのモデルとなるユーザの性格が個性的であるほど性格の模倣性は向上し、高い評価になるだろう。具体的には BigFive の各項目の測定値が平均値から外れているほど GPT-4 が性格の特徴を捉えやすいのではないかと考える。

項目 2 について、被験者がユーザの選好を全て正確に把握していることはないので、少なくともユーザらしさを感じる選好情報を ChatBot が提示することができれば高い評価になると考える。

項目 3 について、口調情報 100 件のみでその模倣を目指している点が本研究のひとつの新規性及び挑戦性である。発話行為タイプ分けを正確に行うことが出来れば、関連性の高い発話例の Few-Shot プロンプトを作成することができるので、高い評価になると考える。一方、発話行為タイプ分けが誤った場合、関連性の低い発話例の Few-Shot プロンプトとなり、評価が下がるだろう。

項目 4 については、項目 1～3 の評価と相関が高くなるかと考える。ChatBot がユーザの性格、選好、及び口調を高精度で模倣することが出来れば、ChatBot との議論はユーザ本人と行う議論に近いものとなり、ユーザがいなくてもユーザの意見を反映させた議論になるだろう。

## 5. おわりに

本研究では、ユーザを模したペルソナを用いた代理議論エージェントの設計手法、及びその実験計画を提案した。ユーザの性格情報と選好情報から返答内容を生成し、それをユーザの口調情報に基づいて制御するための具体的なプロセスを提示した。また、未知の選好情報の推定や、発話行為によるタイプ分けなどによる精度向上も目指した。

今後、既述した実験を行うことで本研究のシステムが、ユーザの代理議論エージェントを担うことができるのか確認できるだろう。また、エージェントによる BigFive パーソナリティ特性検査の再現性の調査や、選好情報推定の精度にフォーカスした実験を行うことで、それぞれのモジュールの詳細な性能が明らかになるだろう。

本研究のシステムの実用化において、ペルソナ作成の容易性は非常に重要となる。より少ないデータから、より高精度なペルソナを作成できる手法確立を今後の課題としたい。また、今回はテキストベースの代理議論システムを構築したが、電話会議やビデオ会議、対面会議など、議論はテキストベース以外でも広く行われている。昨今、生成 AI の技術成長は著しいもので、リアルタイムの音声生成や動画生成により、マルチモーダルな代理議論エージェントの設計、開発を検討していく。

### 謝辞

本研究の一部は、JST CREST (JPMJCR20D1) および NEDO (JPNP20006) の支援を受けた。

### 参考文献

- [1] Sandra Mitrovic, Davide Andreoletti, Omran Ayoub, "ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model For Detecting Short ChatGPT-generated Text", arXiv preprint arXiv:2301.13852 (2023).
- [2] Shimizu Kengo, Uegaki Takatsugu, Kikuchi Hideaki, "強化学習を用いてキャラクタらしさを付与した雑談応答の生成", 言語・音声理解と対話処理研究会, 94, 06 (2022).
- [3] Ishii Ryo, Higashinaka Ryuichiro, Mizukami Masahiro, Katayama Taichi, Mitsuda Koh, Hidetoshi Kawabata, Yamaguchi Emi, Adachi Noritake, Tomita Junji, "Methods of Efficiently Constructing Text-Dialogue-Agent System Using Existing Anime Character", Lecture Notes in Computer Science, 12427, 328-347 (2020).
- [4] Hang Jiang, Xiajie Zhang, Xubu Cao, Jad Kabbara, "PersonaLLM: Investigating the Ability of GPT-3.5 to Express Personality Traits and Gender Differences", arXiv preprint arXiv:2305.02547 (2023).
- [5] Lewis R. Goldberg, "An alternative "description of personality": The Big-Five factor structure", J Pers Soc Psychol, 59(6), 1216-29 (1990).
- [6] Meguro Toyomi, Higashinaka Ryuichiro, Minami Yasuyuhiro, Dohsaka Kohji, "Controlling listening-oriented dialogue using partially observable Markov decision processes", Proceedings of COLING, pp.761-769. (2010)
- [7] Oshio Shinji, Abe Shingo, Cutrone Pino, "日本語版 Ten Item Personality Inventory (TIPI-J) 作成の試み", パーソナリティ研究, 21(1), 40-52 (2012).