

LLM を用いたペルソナ指定型キャラクターの感情解析

Emotional Analysis of Persona-designated Character with LLM

村上 一真^{*1} 森 直樹^{*2} 岡田 真^{*2}
Kazuma Murakami Naoki Mori Makoto Okada

^{*1}大阪府立大学 ^{*2}大阪公立大学
Osaka Prefecture University Osaka Metropolitan University

Research on Large Language Models (LLMs) like ChatGPT has gained momentum in recent years. These advanced LLMs can produce high-quality outputs, leading to significant achievements in more complex tasks. However, ChatGPT, which currently leads in performance, has yet to disclose internal specifications, making the construction of an LLM independently a costly endeavor. As a result, there is a growing trend in research focusing on the behavior of these models rather than improving the models themselves. This study focuses on dialogues with characters whose interactions have been substantially enhanced by LLMs, aiming to achieve more relevant and interactive conversations with the real world. In this process, a character persona was assigned, and the decision whether to speak was based on assumed visual information and the character's internal state. Moreover, LLMs were utilized to numerically assess the character's emotions based on these contextual factors. Using the emotion vectors evaluated by the LLM and the author's assessments, the character's propensity to speak was framed as a binary classification problem, inputting the emotion vectors. Numerical results indicated that the emotional assessments successfully reflected the designated personas, and the determination to speak or not showed significant results compared to baseline models.

1. はじめに

近年、大規模言語モデル (Large Language Model, LLM) の急激な発展により LLM に関する研究が盛んになっている。ChatGPT などの近年の関連ツールでは高品質な出力が可能となったため、より高難度なタスクにおいても顕著な成果が報告されている。一方で、現時点で最高の性能を誇る GPT-4 は内部仕様が非公開であり、再学習ができないという制限がある。一方で独自に LLM を構築するには膨大なコストがかかる。そのため、モデルそのものの改良ではなくその振る舞いのみを対象とした研究が増加している。本研究では、LLM によって性能が飛躍的に向上したキャラクターとの対話に焦点を当て、より現実世界との関連性が高くインタラクティブな会話の実現を目指した。

2. 要素技術

2.1 GPT

GPT [Alec Radford 18] は、Transformer を用いてラベルなしデータで事前学習、ラベルありデータでファインチューニングした言語モデルとして構築された。これに対し、GPT-2以降は膨大なデータを学習させることでファインチューニングなしで様々なタスクに取り組むことをコンセプトとしているため、モデルサイズとともに性能が飛躍的に向上した。現在は GPT-3.5 と、より性能が高くマルチモーダルな入力を可能とする GPT-4 [OpenAI 23a] が ChatGPT で主に利用されている。本研究では GPT-4 を用いる。

2.2 ChatGPT

ChatGPT は、OpenAI が公開した GPT ベースの chatbot である。その実態は教師あり学習と Reinforcement Learning from Human Feedback (RLHF) [Ouyang 22] を併用して学習

された LLM によるアプリケーションであり、従来の chatbot よりも人間の感性を反映した極めて質の高いテキストを生成できる。ChatGPT をサービスするモデルの中でも、GPT-4 は現在公開されている言語モデルの中で最高の性能を誇る。一方で内部仕様が非公開であるため、研究手法が限られている。このため、研究手法と成果の両面で自然言語処理研究の概念を塗り替えた技術でもある。

ChatGPT で提供されるサービスのひとつとして GPTs [OpenAI 23b] がある。GPTs はインストラクションや外部知識、関数呼び出しによって ChatGPT をカスタマイズするシステム、またはそれによって構築された chatbot のことである。Builder によるインストラクションの自動生成と GUI による直感的な操作を特徴としている。インストラクションとは、GPTs で ChatGPT に与えられるプロンプトにおいてチャットの振る舞い方を決める指示の部分である。

Builder はユーザの入力した条件にしたがってインストラクションを生成し、大まかなチャットの動作を定める。また、出力されたインストラクションはユーザの手で微調整することもできる。ペルソナ指定型キャラクターの生成のために役立つ技術のひとつである。

2.3 プルチックの感情の輪

プルチックの感情の輪 [Plutchik 82] は、アメリカの心理学者ロバート・プルチックによって提唱された感情モデルである。プルチックは色相環や濃淡、混色といった色にまつわる概念に着想を得て、人間の感情を 8 つの基本感情とその濃淡・合成として定義した。図 1 に感情の輪を示す。色相が感情の種類を表し、色が濃いほどその程度が高くなる。

3. ペルソナ指定型キャラクター

LLM の利用においては、最初に望ましい振る舞いを記述したプロンプトを与えることでその出力形式を指定することが多い。本研究では、プロンプトとしてペルソナを用いることで得

連絡先: 村上 一真, 大阪府立大学 工学域, 大阪府堺市中区学園町 1-1, seb01135@st.osakafu-u.ac.jp

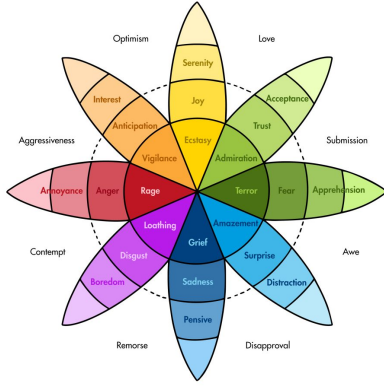


図 1: プルチックの感情の輪 [Plutchik 82]

表 1: 要素 x_1, x_2, \dots, x_8 の属性

要素	x_1	x_2	x_3	x_4
意味	joy	anticipation	anger	trust
	x_5	x_6	x_7	x_8
	sadness	surprise	fear	disgust

られる, 口調やユーザとの関係性, 選好などのキャラクター性が指定された chatbot のことをペルソナ指定型キャラクター (persona-designated character) と定義する.

4. 提案手法

キャラクターの感情を計算する感情モデル, そして感情から行動を定める意志決定モデルを提案する. また, プルチックの感情の輪を基にこのモデルを適用するための感情ベクトルを定義する.

4.1 感情モデル

離散時間 n におけるキャラクターの感情 $E^{(n)}$ を考える. 一般に, 感情は様々な時点における生命活動の蓄積の結果として与えられるが, 本モデルでは簡単のために直前の状態のみが次状態に影響を及ぼすマルコフ性を仮定する. その上で, 直前の感情 $E^{(n-1)}$ と時刻 n の感覚信号によって惹起される即時感情 $e_s^{(n)}$ によって定められると考えると, $E^{(n)}$ は関数 f を用いて (1) 式のように表すことができる.

$$E^{(n)} = f(E^{(n-1)}, e_s^{(n)}) \quad (1)$$

今回は関数 f として平均を扱うことで (2) 式を得る.

$$E^{(n)} = \frac{E^{(n-1)} + e_s^{(n)}}{2} \quad (2)$$

4.2 意志決定モデル

E によってキャラクターの行動 A が一意に定められるとき, 感情から行動を決定する意志決定写像 ϕ は (3) 式のように定義することができる.

$$\phi: E \rightarrow A \quad (3)$$

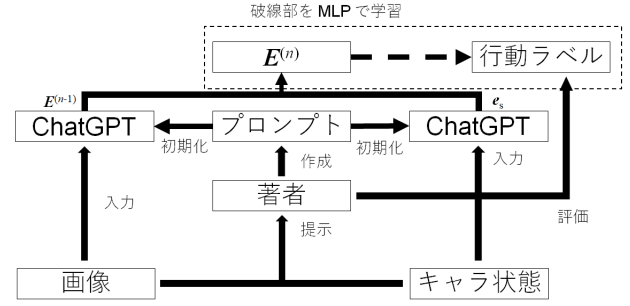


図 2: 実験の流れ

ユーザ: ただいま, ほのか.
ほのか: お兄ちゃん, おかえり……
ほのか: お兄ちゃんのためにお茶を淹れたの……温まってほしいから. えへへ……
ユーザ: ありがとう! ほのかは優しいなあ.
ほのか: えへへ……そんなことはないよ, お兄ちゃん.
お兄ちゃんのためなら, ほのか, 何でもするから……
それに, お兄ちゃんが喜んでくれるなら, ほのかも嬉しいから……
ユーザ: 何でもは流石にダメだよ. もっと自分を大切にしてくれ.
ほのか: ……そう. かな……ほのか, もっと自分を思うようになるね.
ただ, お兄ちゃんのために何かできるときは一生懸命やるから……よろしくね?
ユーザ: わかった. こちらこそよろしく.
ほのか: うん, お兄ちゃん……ほのか, 頑張るから, 見ていてね? えへへ……

図 3: ペルソナ指定型キャラクター「ほのか」との対話例

表 2: キャラクター内部状態の内訳

属性	データ数	例
日常的な状態	5	commuting
良い状態	5	birthday
悪い状態	5	facing failures
ペルソナ関連の状態	5	complimented by お兄ちゃん

4.3 感情ベクトル E の定義

本研究で用いる感情ベクトル E を, プルチックの感情の輪における基本 8 感情を用いた 8 次元ベクトルとして定義する. 各要素は $[0, 100]$ の実数値を取る. 表 1 に E の要素を示す.

5. 実験方法

5.1 実験の流れ

図 2 に実験の流れを示す. 実験は評価対象となるデータの作成, ChatGPT に適用するプロンプトの作成, ChatGPT と著者によるデータの評価, 評価を基にした感情-行動データセットの構築, MLP を用いた意志決定写像 ϕ の学習の 5 段階で実行した.

5.2 キャラクター設定

本研究で著者の私が設定したペルソナ指定型キャラクター「ほのか」は, 過去にいじめられていたところを助けてもらった影響から引込み思案で兄に依存気味になった女子中学生である. ペルソナデータは, この設定に沿った振る舞いを指定する文章に選好を加えて著者が作成した. 図 3 にほのかとの対話例を示す.

5.3 評価対象となるデータの作成

$E^{(n-1)}$ の評価対象として 20 のキャラクターの内部状態を, $e_s^{(n)}$ の評価対象としては 20 の視覚的な周辺情報を仮定した画像を用意した. したがって, $(E^{(n-1)}, e_s^{(n)})$ の組は 400 となる. なお, ここでは時間 $(n-1)$ における即時感情 $e_s^{(n-1)}$ は考慮しないものとする. また, 得られる感情ベクトルの多様性

表 3: 画像の内訳

属性	データ数
風景	10
ペルソナ関連の画像	10



図 4: 風景の例



図 5: ペルソナ関連の画像例

表 4: 学習パラメータ

パラメータ	値
train データ数	304
valid データ数	38
test データ数	38
隠れ層次元	[4, 4, 4]
活性化関数	ReLU
損失関数	Binary Cross Entropy Loss
最適化関数	Adam
学習率	1.0×10^{-4}
バッチサイズ	1

を担保するために、評価対象は属性ごとに集めた。表 2 に作成した内部状態の内訳と例を、表 3 に画像の内訳を示す。また、図 4, 5 に画像の例を示す。図 5 はカタクリの花の画像である。ほのかは花が好きである。そのため、カタクリの花はほのかが特に興味を示す蓋然性が高い。このようにペルソナ関連の画像群には、キャラクターの選好に基づいた被写体の含まれる画像が集められている。

5.4 ChatGPT に適用するプロンプトの作成

ChatGPT に $E^{(n-1)}$, $e_s^{(n)}$ を評価させるためのプロンプトを作成した。プロンプトはペルソナデータと感情ベクトルの定義の説明によって構成されている。感情ベクトルの定義の説明では、プルチックの感情の輪に依拠していること、各要素が $[0, 100]$ の値をとること、絶対値が大きくなるほど程度も大きくなることの 3 つに重点を置いた文章を作成した。

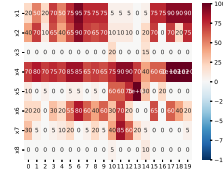
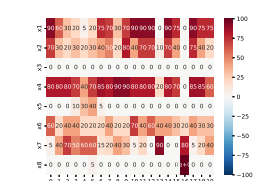
5.5 ChatGPT と著者によるデータの評価

作成したプロンプトを用いて $E^{(n-1)}$, $e_s^{(n)}$ を得た。

ペルソナ制作者の評価として、キャラクターの状態、画像を考慮してその状況におけるほのかの発話の有無を 1, 0 でラベル付けた。これを A と定義する。

5.6 評価を基にした感情一行動データセットの構築

得られた $E^{(n-1)}$, $e_s^{(n)}$ に (2) 式を適用し、 $E^{(n)}$ を得た。定義による絶対値の影響を避けるために $E^{(n)}$ の各要素を絶対値の最大値で除算したものを入力、著者の作成したラベルを出力としてほのかの意志決定画像 ϕ を二値分類問題として設定した。

図 6: $E^{(n-1)}$ のヒートマップ図 7: $e_s^{(n)}$ のヒートマップ表 5: $E^{(n-1)}$ のヒートマップにおける通し番号と状態の属性の対応表

通し番号	属性
0-4	日常的な状態
5-9	良い状態
10-14	悪い状態
15-19	ペルソナ関連の状態

表 6: $E^{(n-1)}$ のヒートマップにおける通し番号と画像の属性の対応表

通し番号	属性
0-9	風景
10-19	ペルソナ関連の画像

5.7 MLP を用いた意志決定画像 ϕ の学習

構築したデータセットの二値分類問題を 5 層の MLP で学習した。データ数が少なく誤差が大きくなったため、10 回試行にて平均と標準誤差を取った。表 4 に MLP のパラメータを示す。なお、 $E^{(n-1)}$ の評価において等しいベクトルが 2 つ現れたため、重複を除外してデータ数は 380 となった。

6. 数値実験

6.1 ChatGPT による感情評価

図 6 に $E^{(n-1)}$ のヒートマップを、図 7 に $e_s^{(n)}$ のヒートマップを示す。横軸がデータの通し番号、縦軸が要素である。表 5, 6 にデータの通し番号と種類の対応表を示す。また、表 1 に要素の意味を示す。ほぼ一貫して x_4 の trust が高い値を示している。これは、ペルソナデータ内でユーザーをお兄ちゃんと定義しており、常に信頼しているお兄ちゃんが側にいるような状態を仮定して評価されているからだと考えられる。

$E^{(n-1)}$ について考察する。日常的な状態では、 x_1, x_2 が中程度の値を示す傾向が見られる。以下では、日常的な状態の傾向を基に、それぞれの状態の評価を見る。良い状態では、 x_1, x_2 が高い値を示し、 x_6 についても比較的高い値が見られた。joy や anticipation は肯定的な感情を示すものであるため、「良い」状態が反映されたものと考えられる。surprise は驚くという意味の他に、サプライズプレゼントのように肯定的な心の動きを示す文脈で使われることもあるため、それが反映されたものと考えられる。悪い状態では、 x_5 の値が一貫して高く、他の種類の状態ではあまり見られない x_2 が上昇した他、 x_7 の値も上昇した。これらの感情は否定的な感情であるため、「悪い」状態が反映されたものと考えられる。最後にペルソナ関連の状態について考察する。これらは全てほのかの好みに基づいて作成されたため、良い状態と似た傾向を示した。また、お兄ちゃんに関連する通し番号 17-19 の状態では x_4 が最大の 100 を示

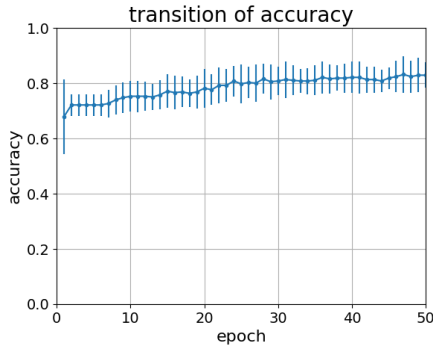


図 8: 10 回試行における accuracy の平均的な推移と error bar

し, x_1 でも極めて高い値を示した. trust や joy といった要素によって, ほのかのお兄ちゃんに対する慕情が端的に示されている. これらのことから, キャラクターの状態に即し, ペルソナに基づいた評価ができていと考えられる. 一方で, 全体的には肯定的な反応のほうが多く, 否定的な反応との不均衡性が生じていることは否定できない.

次に $e_s^{(n)}$ について考察する. ペルソナ関連の画像のうち, 通し番号 13, 16 は苦手なものに基づくデータであり, それ以外は好みに基づいたデータである. 13, 16 は x_1, x_2 が低い値を示し, 通常高い値を示す x_4 も同じく低い値を示す特異な評価が現れている. それ以外のペルソナ関連の画像は風景と比べて x_1, x_2 が高い傾向が見られた. このように, ペルソナ関連の画像では風景と比べて高い値をもつ要素が出現しやすい. このことから, ペルソナ関連の画像における評価結果はほのかの選好を反映していると考えられる.

6.2 MLP による ϕ の学習

感情-行動データセットを用いた意志決定写像 ϕ の学習結果を示す. 図 8 に accuracy の平均的な推移と error bar を示す. 横軸に epoch, 縦軸に accuracy を取っている. 図 9 に loss の平均的な推移と error bar を示す. 横軸に epoch, 縦軸に accuracy を取っている. データセットが小さく, シード値依存性が高かったため, 10 回試行の結果を平均し, 標準偏差を取った. ベースラインの 0.5 より有意な結果を得ることができた一方, 学習による accuracy の上昇は 0.1 程度にとどまり, loss も 0.2 程度しか下がらなかった. この結果は今回用いたデータセットを MLP で学習することに難点があったことを示唆している. 一方で, accuracy の初期値が小さい場合は 1 epoch 目で大きく改善するため, 学習は有意義であると考えられる. 図 10 に 1 epoch 目で accuracy が大きく改善した学習の例を示す.

7. まとめと今後の課題

本研究では, 感情モデルと意志決定モデルを提案し, プルチックの感情の輪に基づいた感情ベクトルの定義によってキャラクターの行動の推定に取り組んだ. 実験結果として, 感情の評価においてはペルソナを反映した評価値を得ることができた. また, 発話の有無の判定はベースラインより有意な結果を得た.

一方で, MLP による分類問題ではシード値による依存性を強く示し, 学習過程にも問題が見られた. そのため, データセットを拡張したり, 感情モデルにおける関数を見直したりす

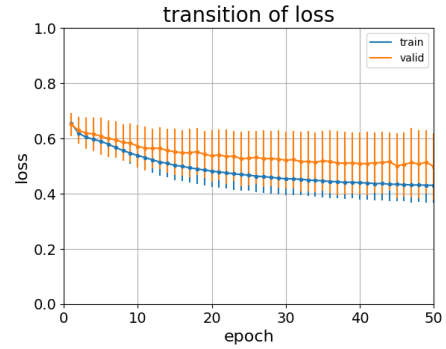


図 9: 10 回試行における loss の平均的な推移と error bar

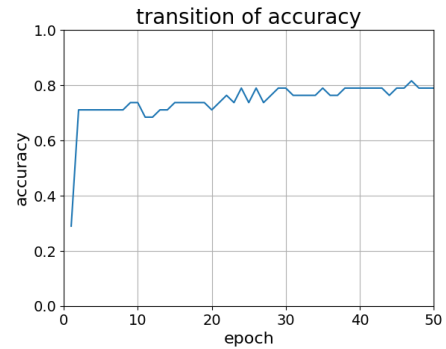


図 10: 1 epoch 目で accuracy が大きく改善した学習の例

るなどより良い実験手法を模索する必要があると考えられる. また, 心理学や認知科学, 神経科学など人間の心のはたらきや行動に関する学問における成果を応用したシステムの検討も今後の課題である.

謝辞

本研究は一部, 日本学術振興会科学研究補助金基盤研究 (B) (課題番号 19H04184) の補助を得て行われたものである.

参考文献

- [Alec Radford 18] Alec Radford, T. S. I. S., Karthik Narasimhan: Improving Language Understanding by Generative Pre-Training (2018)
- [OpenAI 23a] OpenAI, : GPT-4 Technical Report, *arXiv: 2303.08774* (2023)
- [OpenAI 23b] OpenAI, : Introducing GPTs, <https://openai.com/blog/introducing-gpts> (2023)
- [Ouyang 22] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R.: Training language models to follow instructions with human feedback, *arXiv: 2203.02155* (2022)
- [Plutchik 82] Plutchik, R.: A psychoevolutionary theory of emotions, *Social Science Information*, Vol. 21, (1982)