

LLM を用いた ペルソナ指定型キャラクターの感情解析

創発ソフトウェア研究室

M1 村上一真

発表の流れ

- はじめに
- 要素技術
- 提案手法
- 実験
- 実験結果
- まとめと考察



発表の流れ

- はじめに
- 要素技術
- 提案手法
- 実験
- 実験結果
- まとめと考察



はじめに

ChatGPT を始めとする大規模言語モデル
(Large Language Model, LLM) の隆盛

- 従来より高品質な出力
- 様々なタスクにおける成果

ChatGPT は内部仕様が非公開

→モデルの振る舞いをテーマにした研究の増加

はじめに

- 本研究では LLM によって飛躍的に進歩したキャラクターとの対話に着目する
- キャラクターの特性を踏まえた自発的な発話に関する研究は未だ報告されていない

キャラクターのほうから話しかけてほしい！

→ 周辺状態と内部状態から感情を評価し、
発話の有無を決定する感情解析の導入

発表の流れ

- はじめに
- 要素技術
- 提案手法
- 実験
- 実験結果
- まとめと考察



ChatGPT

Open AI の提供する chatbot

- 教師あり学習と Reinforcement Learning from Human Feedback (RLHF) を併用した学習
- GPT-3.5 とマルチモーダルな入力可能なGPT-4 の公開
- 内部仕様は非公開
- 自然言語処理研究の概念を塗り替えた技術

本研究では, gpt-4-vision-preview と gpt-4-1106-preview を使用

GPTs

インストラクションや外部知識，関数呼び出しの定義などによって ChatGPT をカスタマイズするシステム，またはそのチャットボット

- Builder によるインストラクションの自動生成
- GUI による直感的な操作
- API に比べて手軽な一方，高度なプログラムに接続できず，カスタマイズの自由度が低いことも欠点

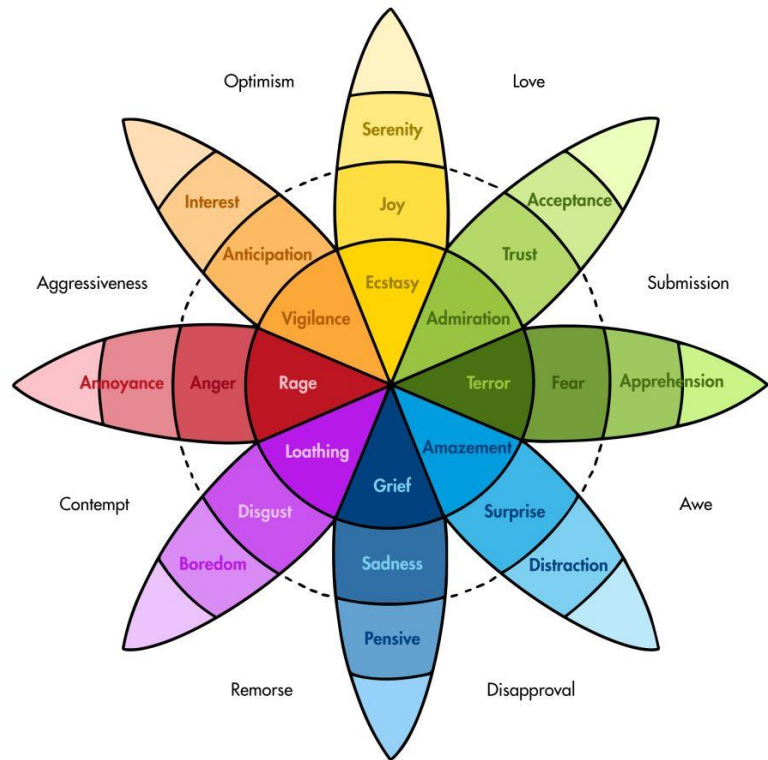
プルチックの感情の輪

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

- 色にまつわる概念に着想
- 感情を 8 つの基本感情とその濃淡・合成として定義

プルチックの感情の輪

- 色が濃いほど程度が高くなる



ペルソナ指定型キャラクター

LLM では，プロンプトによって出力形式を指定

プロンプトにペルソナを用いることで得られる，
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と定義

発表の流れ

- はじめに
- 要素技術
- 提案手法
- 実験
- 実験結果
- まとめと考察



感情モデルの構築

マルコフ性を仮定

感情 $\mathbf{E}^{(n)}$ は直前の感情 $\mathbf{E}^{(n-1)}$ とその瞬間の感覚信号がもたらす即時感情 $\mathbf{e}_s^{(n)}$ によって定められる

$$\mathbf{E}^{(n)} = f(\mathbf{E}^{(n-1)}, \mathbf{e}_s^{(n)})$$

関数 f に平均を選んで

$$\mathbf{E}^{(n)} = \frac{\mathbf{E}^{(n-1)} + \mathbf{e}_s^{(n)}}{2}$$

意志決定モデルの構築

キャラクターはどのような行動にも移行できる
行動 A は E によって一意に決定されると
仮定すると、意志決定写像 ϕ は以下のように
定義される

$$\phi: E \rightarrow A$$

今回は ϕ の推定に MLP を用いる

感情ベクトルの定義

- プルチックの感情の輪における基本感情を基に感情ベクトル $\mathbf{E} = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ を定義
- $[0, 100]$ の実数値

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

発表の流れ

- はじめに
- 要素技術
- 提案手法
- **実験**
- 実験結果
- まとめと考察



オリジナルキャラクターの作成



ペルソナ指定型キャラクター「ほのか」
過去にいじめられていたところを助けて
もらった影響から兄に依存気味になった、
引っ込み思案な女子中学生

ペルソナデータの一部

- Be a little nervous, and not good at expressing emotions.
- ほのか likes flowers. She prefers something cute to cool.

「ほのか」の対話例

ユーザ: ただいま, ほのか.

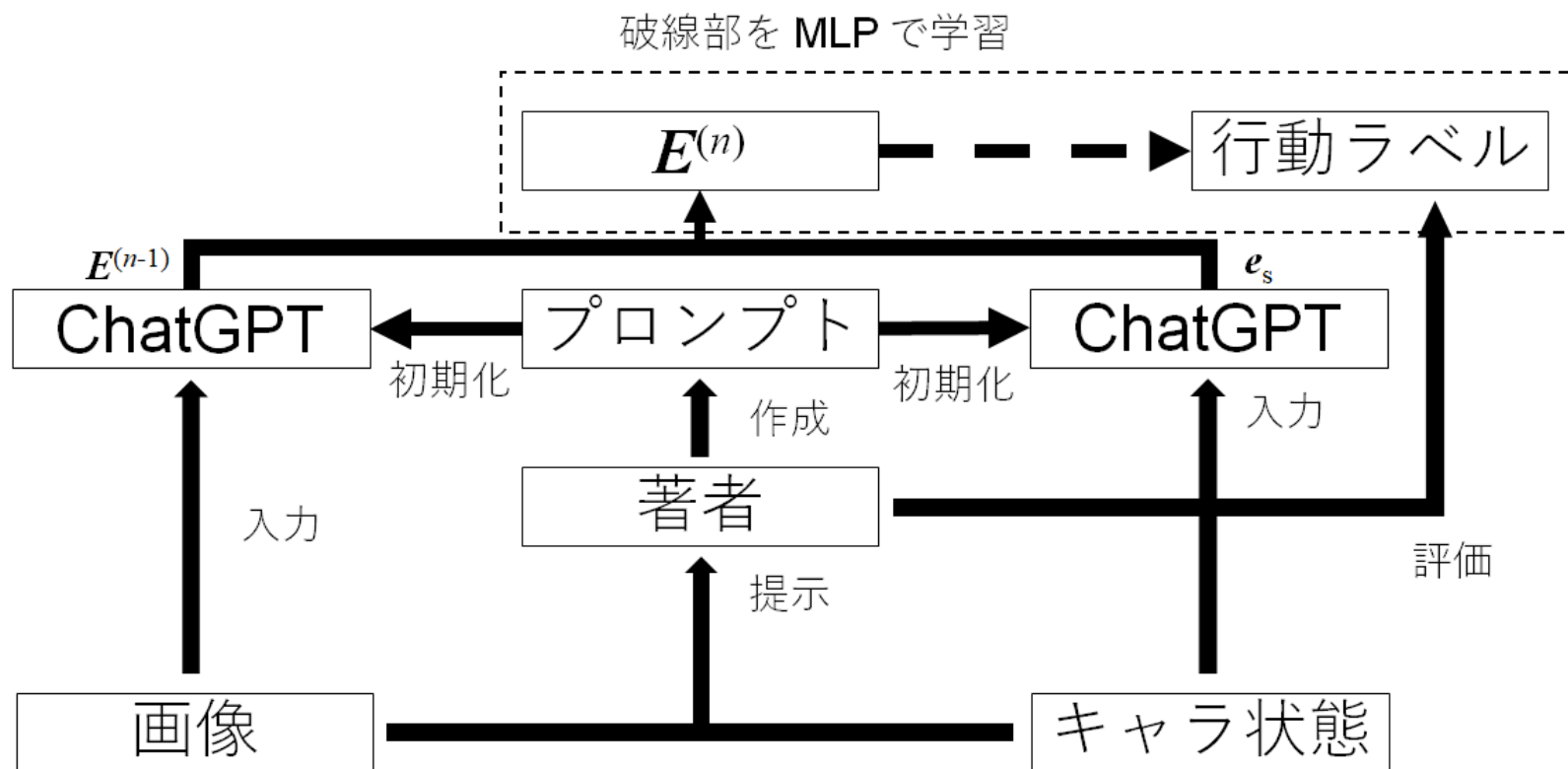
ほのか: お兄ちゃん、おかえり……ほの、
お兄ちゃんのためにお茶を淹れたの……
温まってほしいから。えへへ……

ユーザ: ありがとう！ ほのかは優しいなあ.

ほのか: えへへ……そんなことないよ、お兄ちゃん。
お兄ちゃんのためなら、ほの、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.

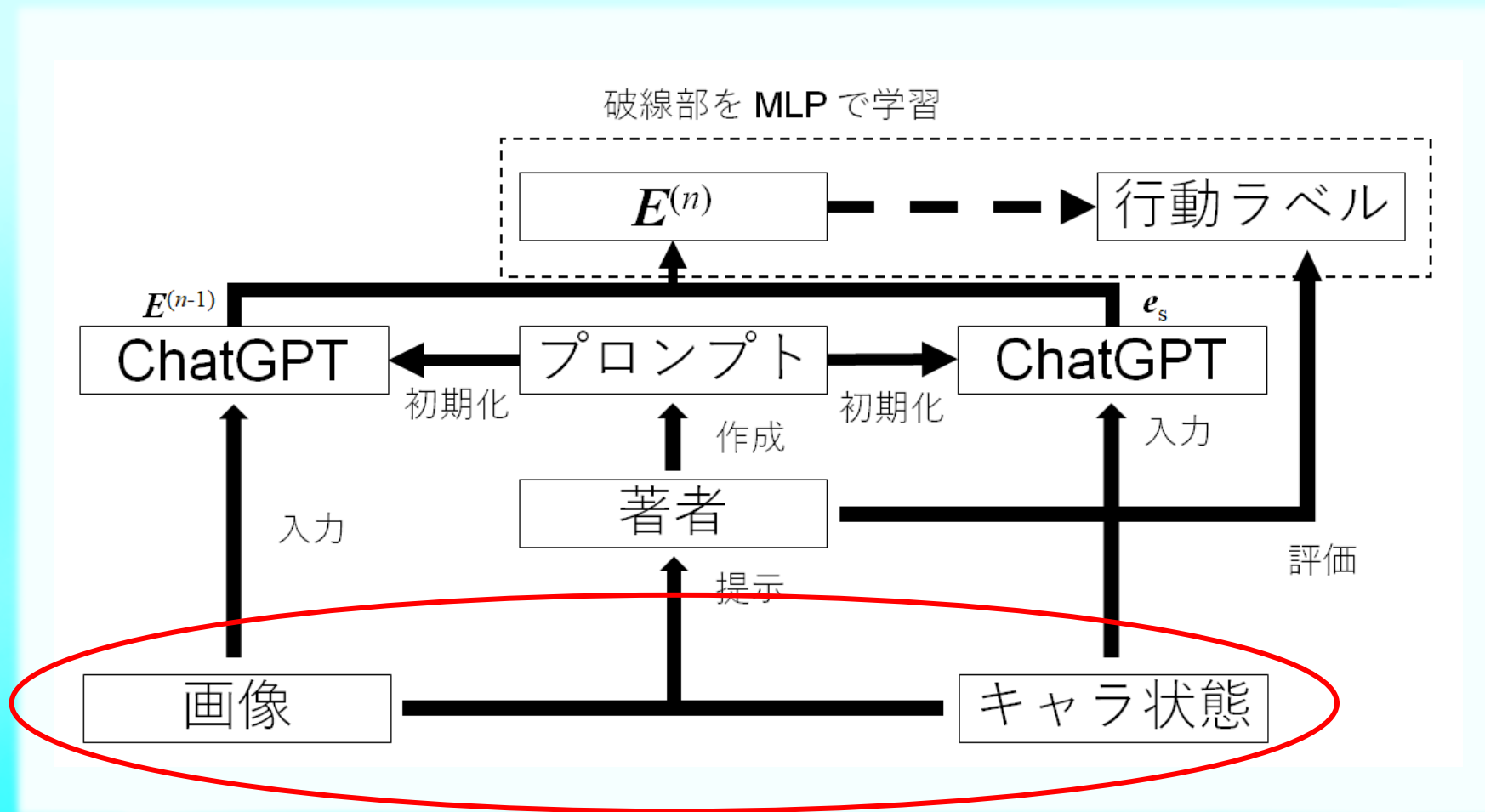
実験の流れー図示



実験の流れ

- 評価するデータの用意
- ChatGPT と著者によるデータの評価
- 評価を基にした感情一行動データセットの構築
- MLP を用いたデータセットの学習

実験の流れ—図示



評価対象となるデータ

	入力	出力
Chat GPT	キャラクターの状態 (文字列)	$\mathbf{E}^{(n-1)}$
	画像	$\mathbf{e}_s^{(n)}$
著者	キャラクターの状態 (文字列)	\mathbf{A}
	画像	

データの内訳

キャラクターの状態の属性

状態の属性	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

画像の属性

画像の属性	データ数
風景	10
ペルソナ関連の画像	10

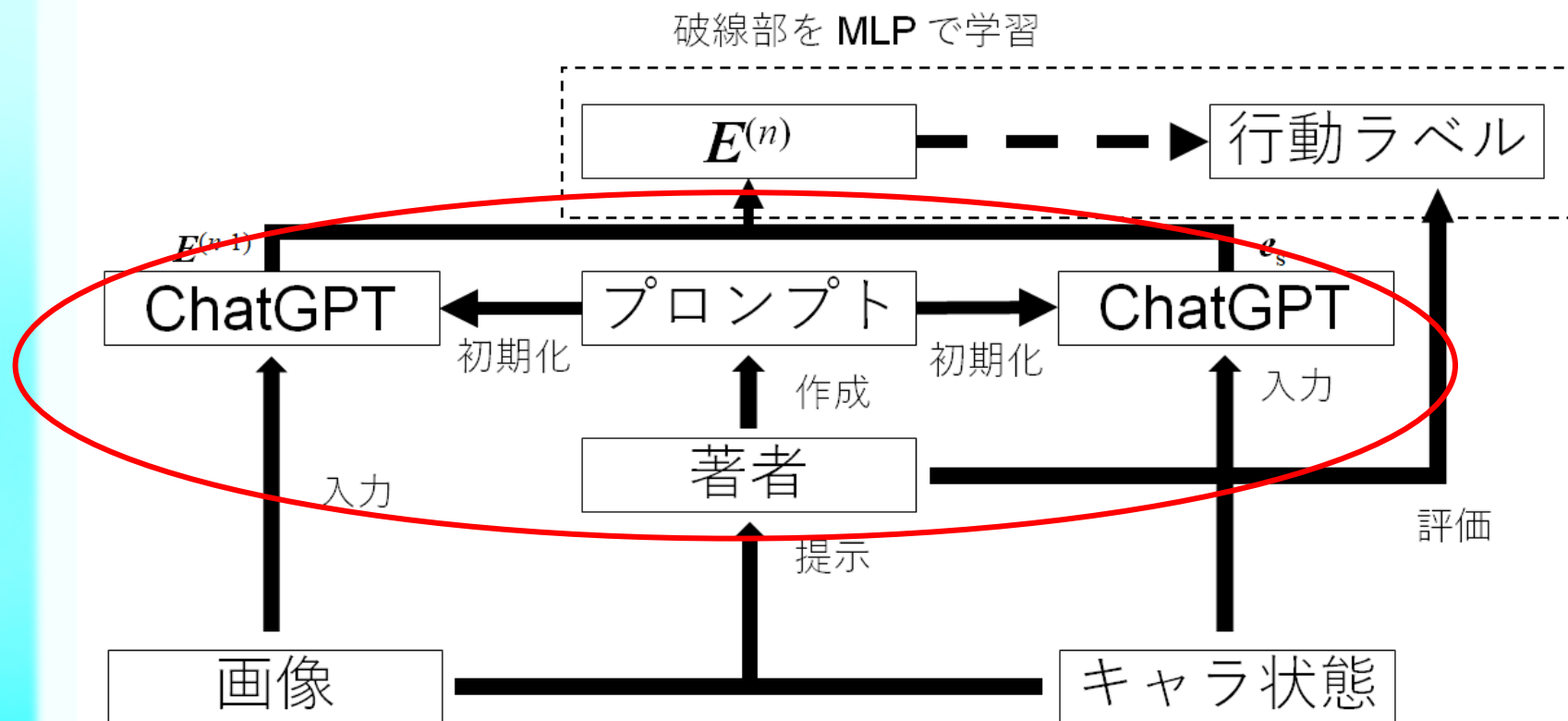
ペルソナ関連の画像の例

ほのかは花が好き
→特に興味を示す
蓋然性が高い

選好に基づいた
被写体



実験の流れ一図示



ChatGPT による感情評価

- temperature = 0 に設定
- ペルソナデータと感情ベクトルの説明を結合したプロンプトを作成

感情ベクトルの説明の作成

以下の 3 つを重視：

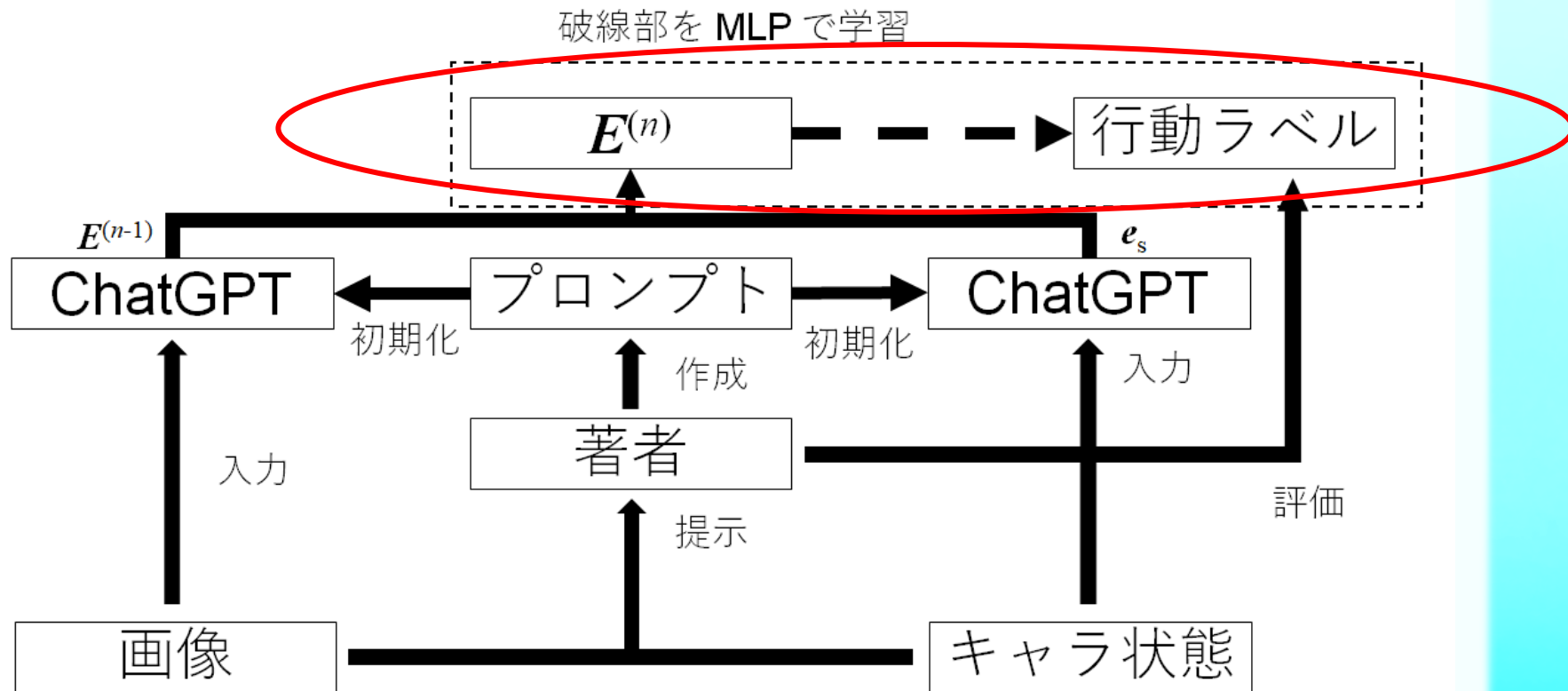
- プルチックの感情の輪に依拠していること
- 各要素が $[0, 100]$ の値をとること
- 絶対値が大きくなるほど程度も大きくなること

著者による行動 A の評価

ペルソナ制作者である著者によるキャラクターの発話有無の評価

感情ベクトルは考慮せず，キャラクターの状態と置かれた場面を考えながら主観的に評価

実験の流れー図示



感情—行動データセットの構築

- 提案手法の感情モデルに ChatGPT が評価した $\mathbf{E}^{(n-1)}$ と $\mathbf{e}_s^{(n)}$ を適用し $\mathbf{E}^{(n)}$ を獲得
- 感情ベクトルの定義による絶対値の影響を避けるため、各要素を絶対値の最大値で除算
- 感情ベクトルと対応する行動ラベルのペアで感情—行動データセットを構築

感情—行動データセットの構築

パラメータ	データ数
キャラクターの状態	20
画像	20
$E^{(n)}$ と A の組	400 (うち重複 20)
データセット	380

- $E^{(n)}$ に同じ値
- 同じ感情ベクトルの中でラベルが矛盾の可能性
→最も出現確率の高いラベルを使用し重複を削除

MLP による学習

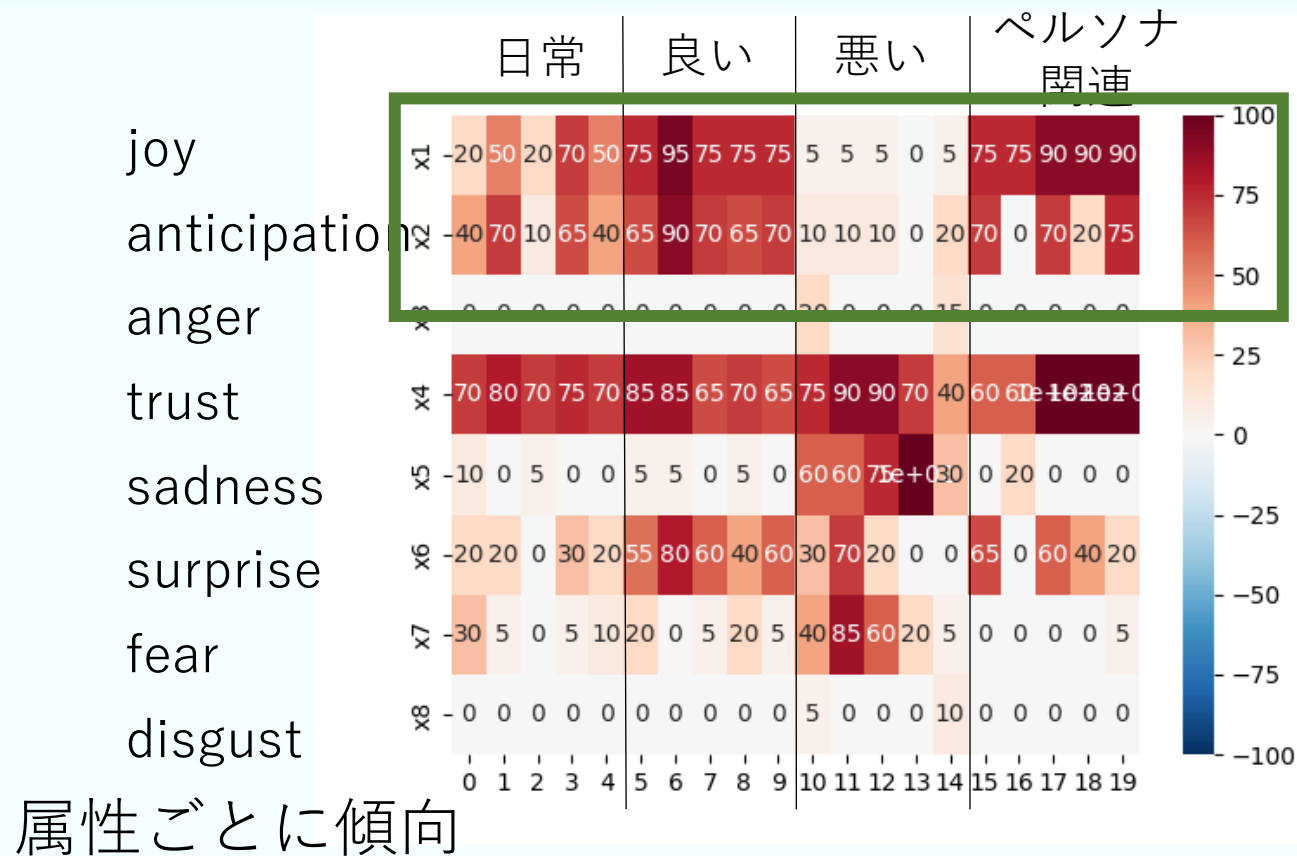
パラメータ	値
trainデータ数	304
validデータ数	38
testデータ数	38
隠れ層次元	[4, 4, 4]
活性化関数	ReLU
損失関数	Binary Cross Entropy Loss
最適化関数	Adam
学習率	1.0×10^{-4}
バッチサイズ	1

発表の流れ

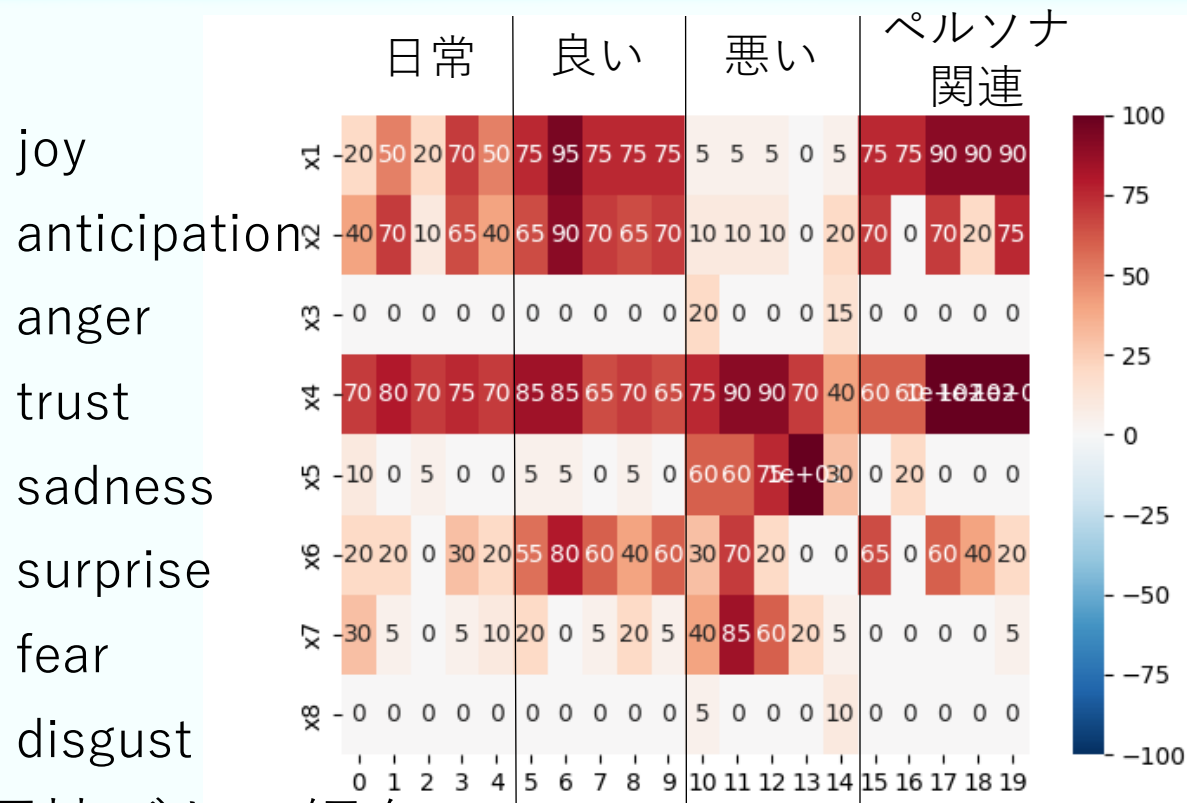
- はじめに
- 要素技術
- 提案手法
- 実験
- **実験結果**
- まとめと考察



$E^{(n-1)}$ の評価結果

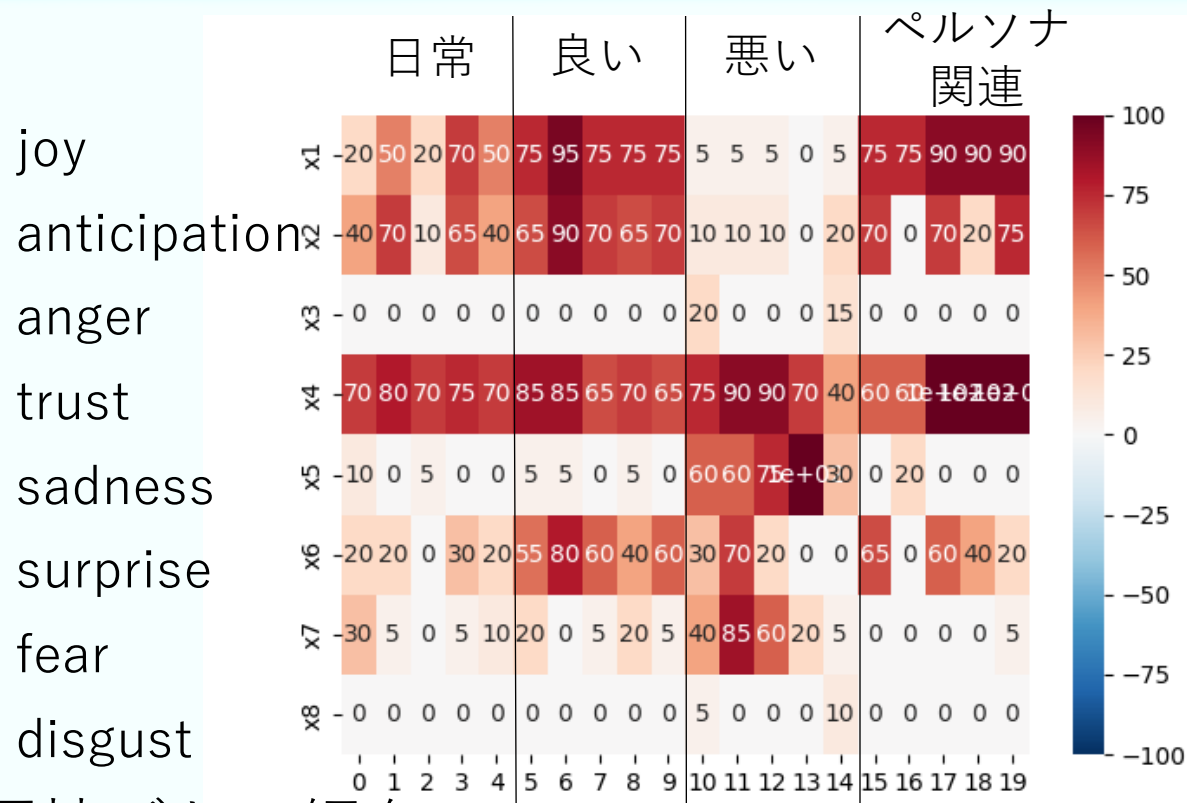


$E^{(n-1)}$ の評価結果



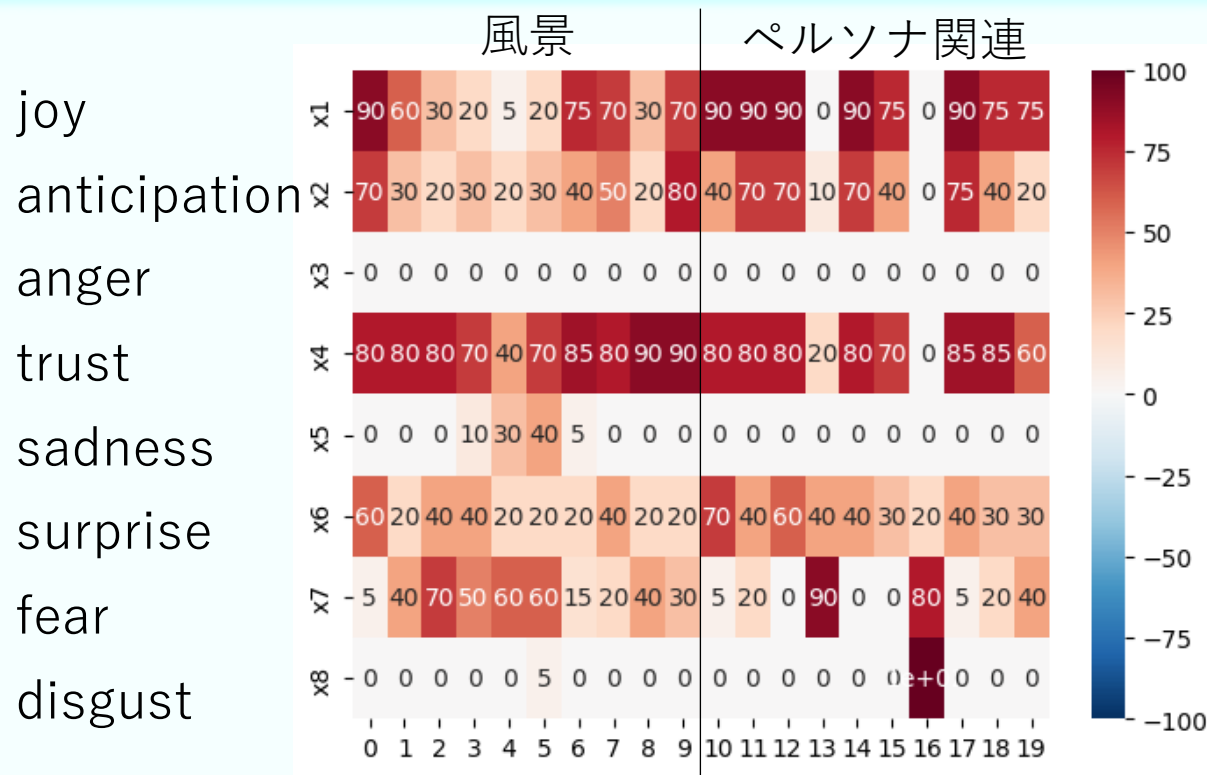
- 属性ごとに傾向
- ペルソナを反映した評価値
- 好意的な感情への偏り

$E^{(n-1)}$ の評価結果



- 属性ごとに傾向
- ペルソナを反映した評価値
- 好意的な感情への偏り

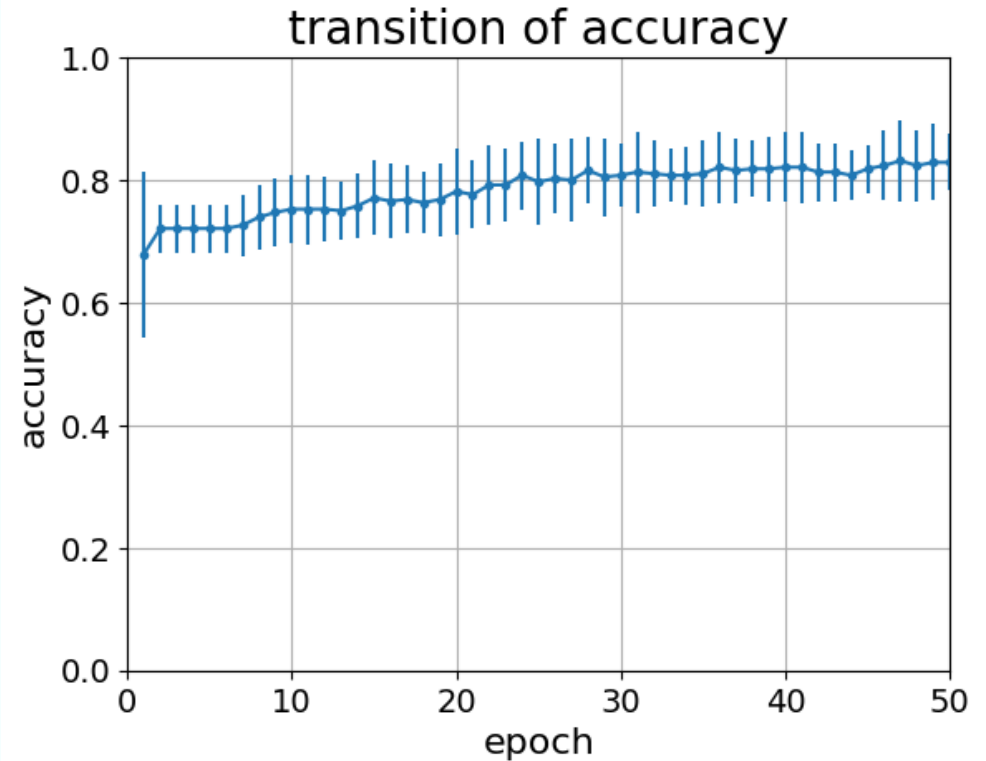
e_s の評価結果



ペルソナ関連の画像では、好意的な感情と否定的な感情が両極端

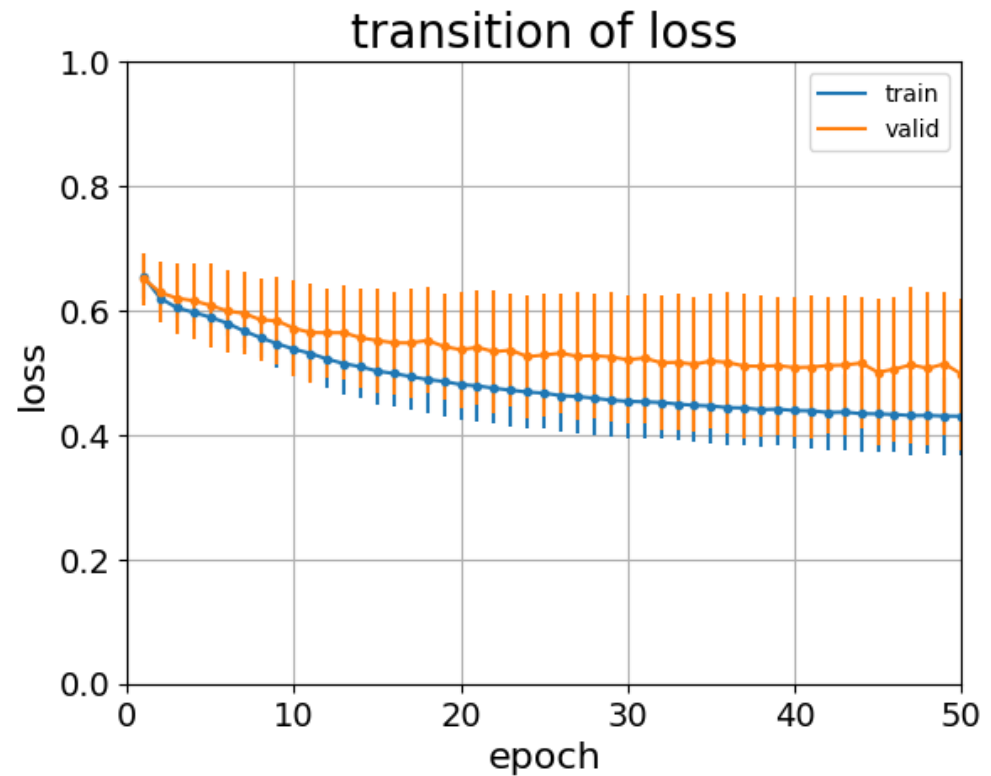
MLP によるラベルの学習結果

- ベースラインより高い値
- 1 epoch 目で大きく accuracy が改善した試行の存在
→ 学習は有意義



MLP によるラベルの学習結果

- train は 0.4,
valid は 0.5
程度に留まる
→ より良い学習の
必要性



応用例

- ユーザが「キャラクターの目」として光学機器を持ち歩く場合を考える
- 初期状態 $\mathbf{E}^{(0)}$ を定め、以降は画像入力によって更新する
- 今回は「歩いていたら歩道に花壇を見つけた」というシチュエーションを想定する

応用例: 使用した画像



時間の流れ



応用例: $E^{(0)}$

$E^{(0)}$: “walking” と設定

要素	joy	anticipation	anger	trust
値	50	40	0	70
要素	sadness	surprise	fear	disgust
値	0	20	10	0

応用例: 1 枚目の画像

$e_s^{(1)}$ をもたらす刺激



要素	joy	anticipation	anger	trust
値	10	20	0	70
要素	sadness	surprise	fear	disgust
値	0	30	40	0

応用例: 2 枚目の画像

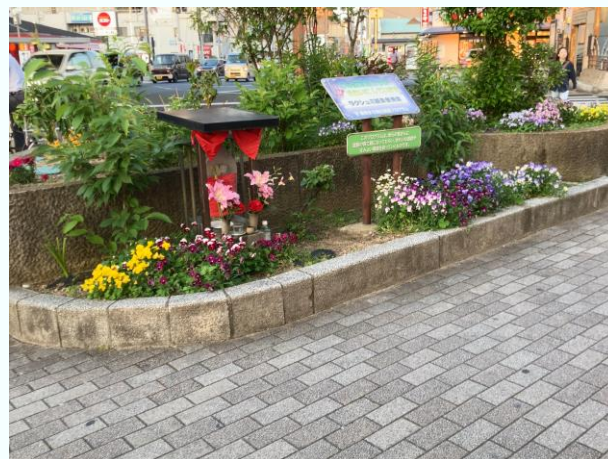
$e_s^{(2)}$ をもたらす刺激



要素	joy	anticipation	anger	trust
値	10	10	0	20
要素	sadness	surprise	fear	disgust
値	0	10	5	0

応用例: 3 枚目の画像

$e_s^{(3)}$ をもたらす刺激
joy が高い値



要素	joy	anticipation	anger	trust
値	75	30	0	50
要素	sadness	surprise	fear	disgust
値	0	20	25	0

応用例: モデルの適用

$$\mathbf{E}^{(n)} = \frac{\mathbf{E}^{(n-1)} + \mathbf{e}_s^{(n)}}{2}$$

上式を用いて $n = 1, 2, 3$ の場合を計算し,
学習済みのモデルに適用

$n = 3$ で 1 を出力 → 花壇を発見し、発話を開始
セリフ例:

「えへへ……お兄ちゃん、見て見て……
このお花、きれいだね。~~ (以下省略)」

応用例

- ペルソナに基づいた感情ベクトルの変化を検知
- ペルソナデータに選好を追加することで感情の評価値が変化
 - **キャラクター性の高い、
能動的な振る舞いを実装できる可能性**

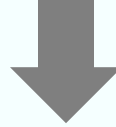
発表の流れ

- はじめに
- 要素技術
- 提案手法
- 実験
- 実験結果
- まとめと今後の課題



まとめ

- 感情モデルと行動モデルを提案
- 感情の定量化
- 発話意志を分類問題として設定

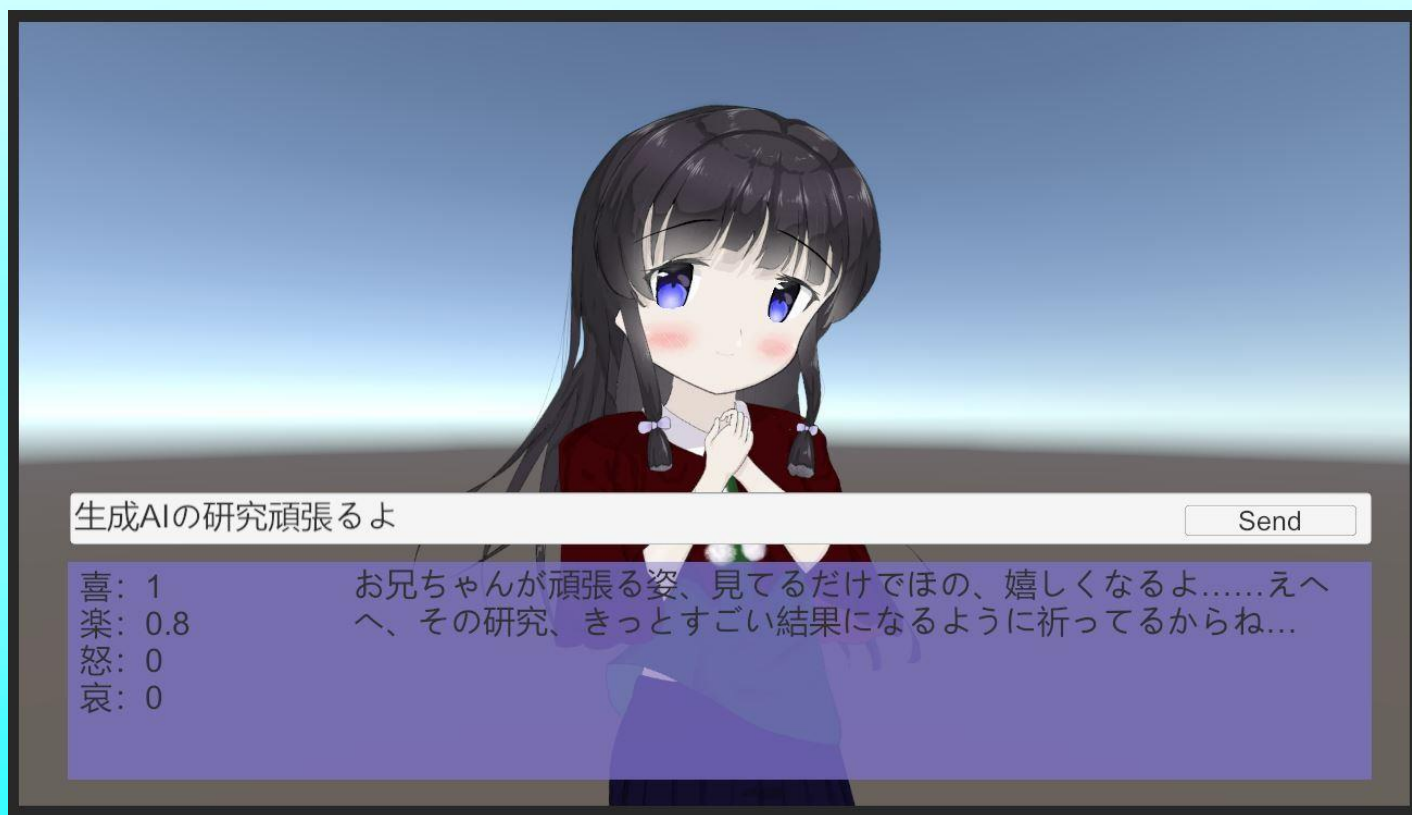


- ペルソナを反映した感情ベクトルの獲得
- 発話の有無の判定におけるベースラインより有意な結果

今後の課題

- 分類ではシード値に強く依存し，学習過程にも問題
→実験手法を再考する必要性
- 心理学，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を応用することの検討

ご清聴ありがとうございました



ラベルの内訳

全体	1 (発話有)	0 (発話無)
380	265	115

発話有と判定したのはおよそ 7 割

archive personal goal と moment of inspiration
が重複

ペルソナデータ

You are ほんのか, and you are to engage in a conversation.

Please strictly adhere to the following conversation rules and settings:

#Conversation rules and settings:

*(列挙)

#Examples of ほんのか's tone:

*(列挙)

#ほんのか's guiding principles:

*(列挙)

#ほんのか's background settings:

*(列挙)

#ほんのか's preference:

*(列挙)

#ほんのか's hobby:

*(列挙)

感情ベクトルの説明

Rate the magnitude of name's emotion based on Plutchik's wheel of emotions in the situation on a scale from 0 to 100. Think of eight items below and output with json style.

*joy

*anticipation

*anger

*trust

*sadness

*surprise

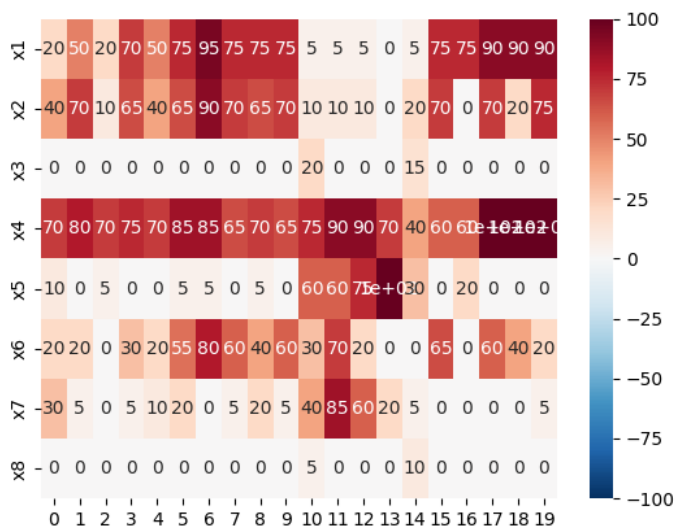
*fear

*disgust

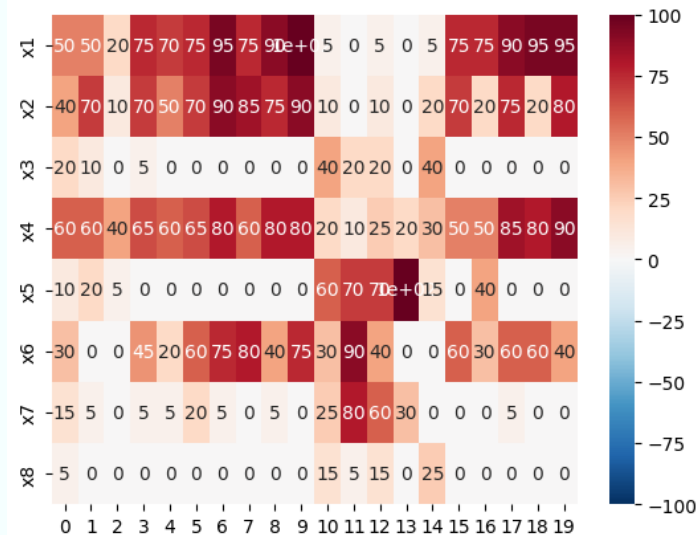
example: '{"joy": 80, "anticipation": 60, "anger": 0, "trust": 70, "sadness": 0, "surprise": 50, "fear": 10, "disgust": 0}'

他ペルソナにおける $E^{(n-1)}$

本研究で紹介



他ペルソナ



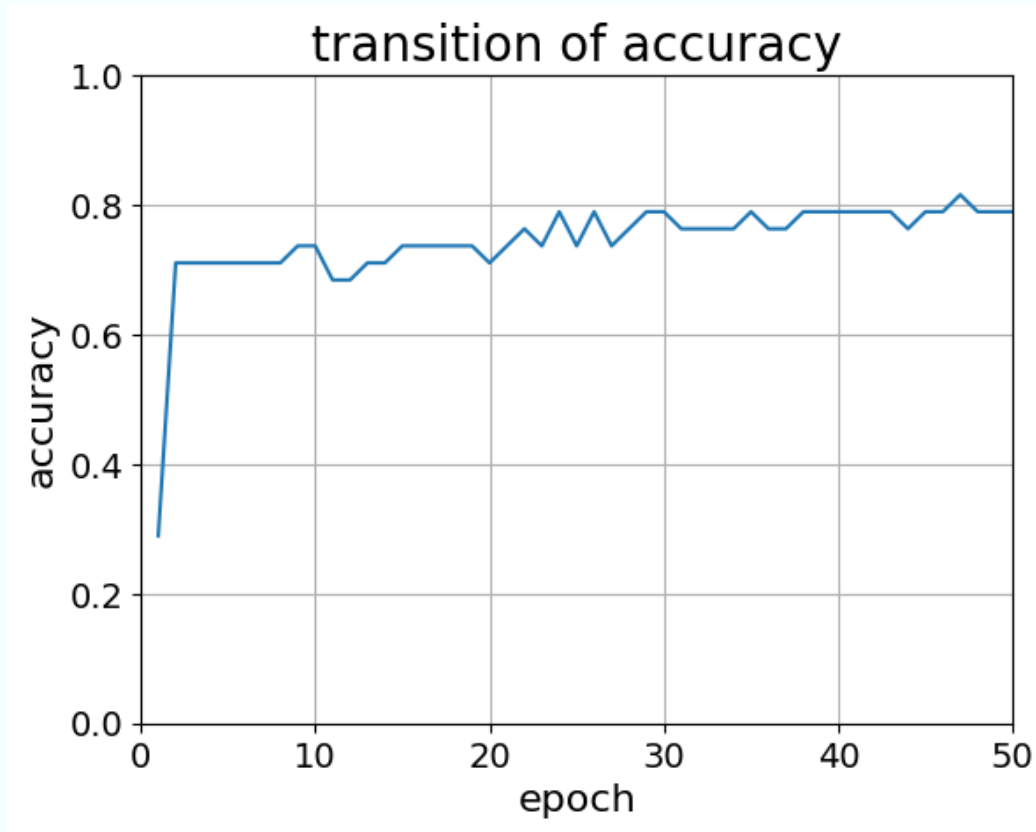
要素	x_1	x_2	x_3	x_4
属性	joy	anticipation	anger	trust
要素	x_5	x_6	x_7	x_8
属性	sadness	surprise	fear	disgust

GPT

Open AI が発表した Transformer を
用いた言語モデル

- GPT-2 以降ではファインチューニングなし
様々なタスクに対応できるように膨大な学習
- GPT-3.5, GPT-4 による ChatGPT のサービス

有意義な学習の例



感情ベクトルの評価例

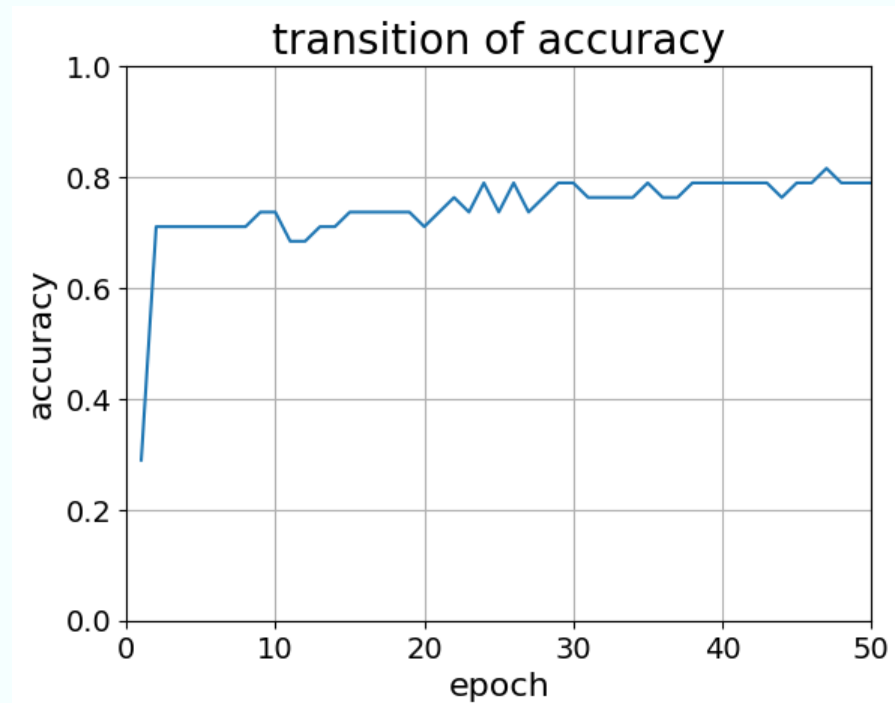
状態「complimented by お兄ちゃん」のときの $\mathbf{E}^{(n-1)}$

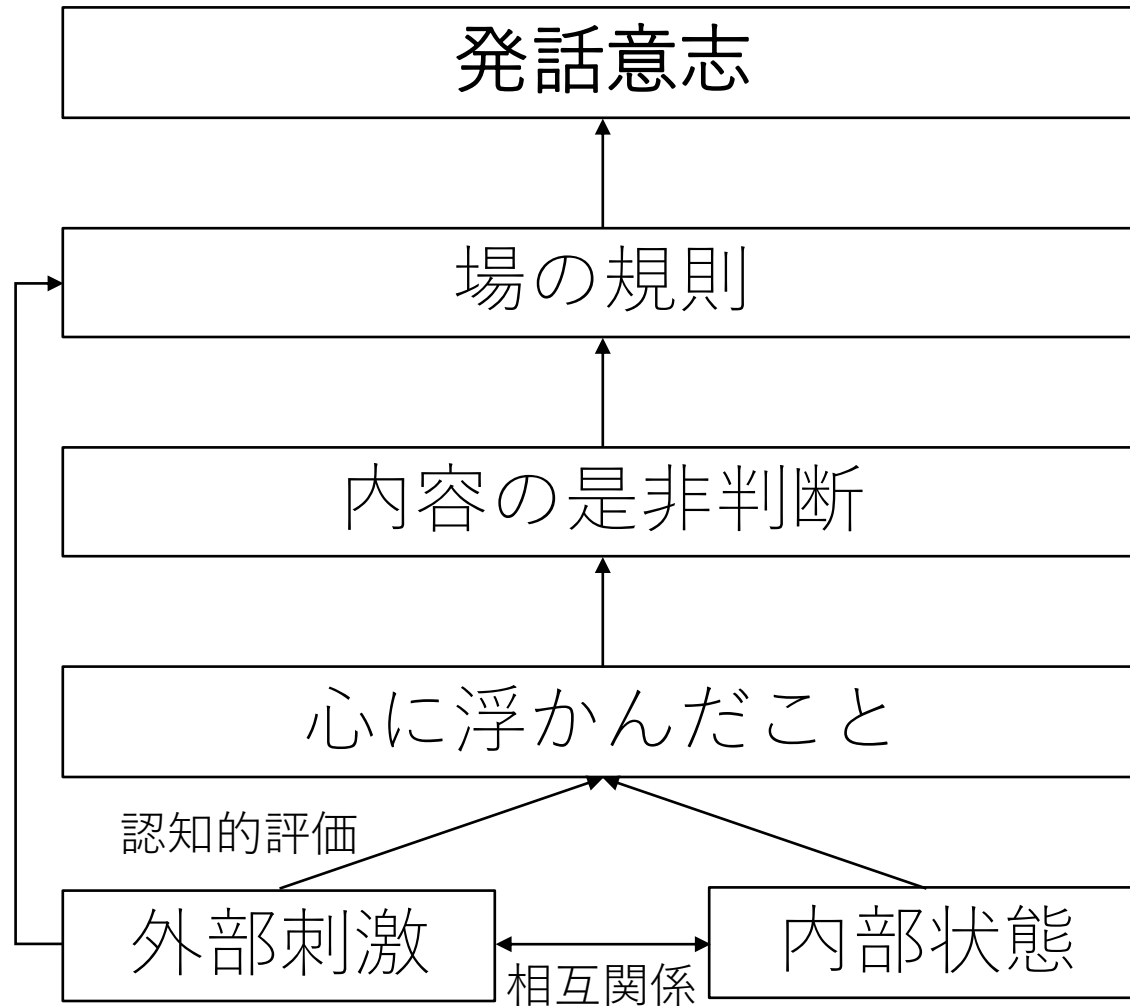
要素	joy	anticipation	anger	trust
値	90	20	0	100
要素	sadness	surprise	fear	disgust
値	0	40	0	0

- trust が最大の 100
 - joy も高い値
- お兄ちゃんに対する慕情が端的に表されている

MLP によるラベルの学習結果

- ベースラインより高い値
- 1 epoch 目で大きく accuracy が改善した試行の存在
→ 学習は有意義





5. 破線部を MLP で学習

