一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
NLC2024-7(2024-09)

大規模言語モデルは未知語に対する人間の感情想起を再現できるか？

宮川　由衣[†]　松平　茅隼[†]　加藤　大貴[†]
平山　高嗣[††,†]　駒水　孝裕[†]　井手　一郎[†]

† 名古屋大学　〒464–8601 愛知県名古屋市千種区不老町
†† 人間環境大学　〒444–3505 愛知県岡崎市本宿町上三本松 6–2
E-mail: † {miyakaway,matsuhirac,katoh}@cs.is.i.nagoya-u.ac.jp, taka-coma@acm.org, ide@i.nagoya-u.ac.jp,
†† t-hirayama@uhe.ac.jp

あらまし　未知語とは，実在せず意味を持たない語である．しかし，たとえ意味を持たなくても，人間はしばしば未知語の発音や綴りから様々な印象や感情を想起することが知られている．未知語に対する人間の知覚を理解することは，商品の特性に合わせた商品名やキャラクタ名の考案などに活用できる．既存研究により，このような人間の知覚の推定に対して大規模言語モデルの有用性が示唆されているが，感情的側面については，大規模言語モデルの有用性は明らかでない．そこで本研究では，未知語から想起される感情に着目し，大規模言語モデルと人間による感情想起の関係性を明らかにすることを目的とする．本発表では，GPT-4 を用いて，未知語に対して人手で感情強度値を付与した先行研究の手順を再現して感情強度値を推定した結果，GPT-4 によって推定された感情強度値と，人手で付与した値との間に 0.40 の正の相関が見られたことを報告する．先行研究では，人間のアノテータ間の相関係数は平均で 0.69 であり，既存語に対する感情強度値を訓練データとして訓練した回帰モデルと人間との間の相関係数は 0.17 であった．これらをふまえると，本研究で得られた GPT-4 と人間との間の感情強度値の一致度合いは高く，大規模言語モデルは人間の感情想起をある程度再現できると言える．
キーワード　大規模言語モデル，単語と感情，未知語

Can an LLM Reproduce Emotional Associations to Nonsense Words by Humans?

Yui MIYAKAWA[†], Chihaya MATSUHIRA[†], Hirotaka KATO[†],
Takatsugu HIRAYAMA[††,†], Takahiro KOMAMIZU[†], and Ichiro IDE[†]

† Nagoya University　Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464–8601 Japan
†† University of Human Environments　6–2 Kamisanbonmatsu, Motojuku-cho, Okazaki-shi, Aichi, 444–3505 Japan
E-mail: † {miyakaway,matsuhirac,katoh}@cs.is.i.nagoya-u.ac.jp, taka-coma@acm.org, ide@i.nagoya-u.ac.jp,
†† t-hirayama@uhe.ac.jp

Abstract　Nonsense words are words that do not exist within the vocabulary of a language. Understanding human perception of nonsense words is helpful to devise product and character names that match their characteristics. Existing studies suggest that Large Language Models (LLMs) could estimate this perception, but they did not focus on emotions. Hence, this study explores the relationship of emotions evoked by nonsense words between LLMs and humans. Using GPT-4, we reproduce the manual procedure of a previous study to annotate evoked emotions for nonsense words. As a result, a positive correlation of 0.40 was observed between GPT-4's and humans' emotion intensity ratings. A previous study reported that the correlation among humans was about 0.69 on average and that between a regression model trained on the annotations for real words and humans was 0.17. These scores indicate that GPT-4's agreement with humans is notably strong.
Key words　Large Language Models, word–emotion associations, nonsense words

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©2024 by IEICE

1. はじめに

未知語とは，実際には存在しない語のことである．たとえこのような意味を持たない語であっても，人間はしばしばその発音や綴りから様々な印象や感情を連想することが知られている [1]．一例として，人間は尖った形や丸い形を特定の音に関連付ける傾向があるという Bouba/Kiki 効果 [2] がある．このような未知語に対する人間の知覚を理解することは，様々な領域での応用が期待される．例えば，心理言語学における人間の一般的な言語処理機構の解明 [3] や，商業分野において，商品やキャラクタ，ブランドの特性に合わせた新しい名称を考案するために役立ち，商品の売上向上やキャラクタ名・企業名の迅速な定着に繋がることなどが期待される．しかし，このような人間の知覚を調査するためには，人間を対象とした実験が必要であり，コストと労力がかかる．そのため，統計解析に十分な規模のデータを得ることは困難である．

先行研究において，大規模言語モデル（Large Language Model; LLM）が自然言語の認知モデルとして機能するかどうかについて議論がなされている [4]．Cai ら [5] は，LLM の 1 つである ChatGPT [6] と Vicuna [7] を用いて，未知語の発音とその音から想起される形や性別との間の関連性を調査した．彼らは，Bouba/Kiki 効果のような人間が未知語から受ける印象の推定における LLM の有用性を示唆している．しかし，語から想起される感情 [8] に対して，LLM がどのように機能するかは解明されていない．

Mohammad と Turney [9] は，実在する英単語と感情の関連性に関する大規模で質が高い感情強度辞書を構築している．彼らの辞書を基に，Sabbatino ら [1] は 272 の未知語を対象とした感情強度辞書を構築している．この辞書では，各未知語に対して，Ekman の 6 つの基本感情 [10]（喜び，悲しみ，怒り，嫌悪，恐怖，驚き）のそれぞれに対する強度値が付与されている．120 人の英語母語話者をクラウドソーシングにより集め，ベスト-ワースト・スケーリングによるアノテーションを行なうことで感情強度値を収集している．また，Sabbatino らは，このアノテーション結果を利用し，未知語の感情強度値を推定する回帰モデルも構築している．実在する英単語（既存語）の感情強度値 [9] を用いて訓練された回帰モデルを用いて未知語の感情強度値を推定した結果，人間による感情強度値との間の Pearson の相関係数は高々 0.17 であった．これは，未知語の特徴量設計や教師あり学習を用いた従来の回帰モデルによるアプローチが未知語から想起される感情の推定に不十分であった可能性が高いことを示唆している．

従来，LLM による未知語の感情推定に焦点を当てた研究は存在せず，LLM が人間と同様に未知語から受ける感情を再現できるかどうかは明らかではない．そこで，本研究では，LLM と人間との間の，未知語によって想起される感情の関係を明らかにし，LLM の特性を解明するための 1 つの観点を提案・検証することを目的とする．

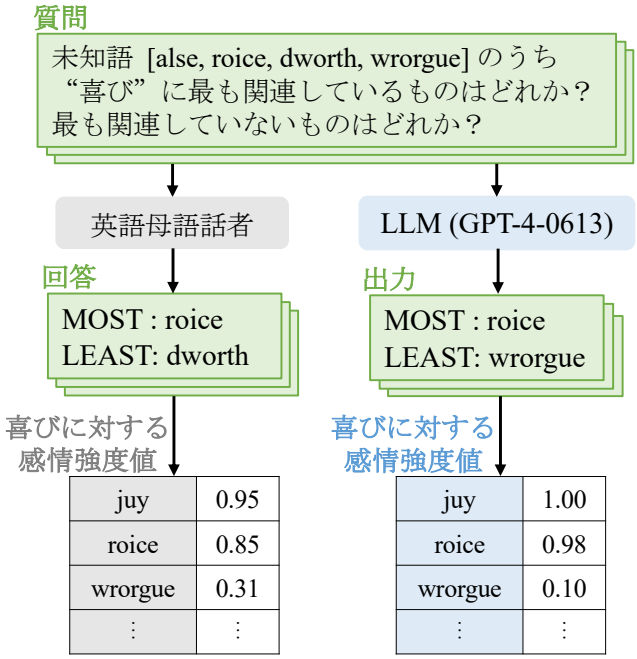

図 1: 人間と LLM それぞれによる未知語に対する感情強度値の付与手順．

Fig. 1 Nonword-emotion annotation procedures by humans and an LLM.

2. 未知語に対する感情評価

Sabbatino ら [1] は，未知語を正書法上正しい綴りであり，単音節のものに限定しており，本研究でも同様の定義を採用する．本節では，LLM を用いて Sabbatino らが行なったベスト-ワースト・スケーリング・アノテーションを再現する手法について説明する．彼らの手法に従い，本研究でも 6 つの基本感情 [10] に注目する．人間及び LLM による未知語に対する感情強度値付与手順を図 1 に示す．これにより，未知語の解釈に関する LLM と人間との間の相関を測定する．

2.1 人間による感情強度値の付与

Sabbatino らによるベスト-ワースト・スケーリング・アノテーションでは，まず，対象となる感情 $e \in \{$ 喜び，悲しみ，怒り，嫌悪，恐怖，驚き $\}$ を 1 つ選択する．次に，272 個の未知語と 68 個の既存語(注1)の中から無作為に 4 語ずつ抽出し，組を作成する．その際に，各単語が 8 個の組に出現するように出現頻度を調整する．作成した各組について，3 人のアノテータは，以下の問いに回答する: “Which of the four words do you associate MOST and which do you associate LEAST with the emotion e?”（「4 つの単語のうち，e という感情に最も関連する単語と最も関連しない単語はどれか？」）母語が英語である 120 人のアノテータから得た回答を用いて，単語 w に対する感情 e の感情強度値 $\mathrm{score}_e(w)$ を以下のように算出する:

$$\mathrm{score}_e(w) = \frac{\mathrm{most}_e(w) - \mathrm{least}_e(w)}{\mathrm{count}_e(w)}. \tag{1}$$

ここで，$\mathrm{most}_e(w)$ と $\mathrm{least}_e(w)$ はそれぞれ，w が感情 e に最も

(注1)：既存語は，先行研究 [9] との比較やアノテータのアテンションチェックのために使用される．

表 1: LLM に入力するプロンプト．各プロンプトにおいて，`Word list` は 4 つの未知語からなる組に対応しており，`EMOTION` は 6 つの感情のうちの 1 つを表している．

Table 1 Input prompt for an LLM. For each prompt, the word list corresponds to a tuple consisting of four nonwords and the EMOTION describes one of the six emotions.

```
Role: System
  Content: You are a native English speaker. Be sure to answer the question within 200 words.
Role: User
  Content: Which of the four nonsense words listed below do you associate most and
           which do you associate least with EMOTION?
           Word list: [WORD1, WORD2, WORD3, WORD4].
           Explain the way you think step by step, and answer with ''MOST:'' for the choice you
           associate most and ''LEAST:'' for the choice you associate least with EMOTION at the end.
```

表 2: (a) LLM による評価と人間による評価の間の各感情に対する Pearson の相関係数（*は p 値が 0.05 未満であることを示す）．
(b) 未知語の感情強度値のアノテーションにおけるスプリット・ハーフ信頼性の値 [1].

Table 2 (a) Pearson's correlation coefficients for the six emotions between the LLM's and the humans' ratings (*: p-value is less than 0.05).
(b) Split-half reliability for annotation of nonword emotion intensity scores [1].

		喜び	悲しみ	怒り	嫌悪	恐怖	驚き	平均
(a)	**LLM（GPT-4）と人間**	0.44*	0.40*	0.41*	0.47*	0.44*	0.26*	0.40*
(b)	**人間同士**	0.72	0.68	0.71	0.72	0.70	0.60	0.69

表 3: 各感情における，未知語に対する LLM による感情強度値と人間による感情強度値の平均と標準偏差．

Table 3 Mean and standard deviation of LLM's ratings and human ratings for six emotions.

	喜び	悲しみ	怒り	嫌悪	恐怖	驚き
LLM（GPT-4）	0.50 ± 0.23	0.50 ± 0.20	0.50 ± 0.23	0.50 ± 0.23	0.50 ± 0.22	0.50 ± 0.20
人間	0.52 ± 0.16	0.48 ± 0.15	0.49 ± 0.16	0.50 ± 0.16	0.50 ± 0.15	0.52 ± 0.13

関連する単語と最も関連しない単語として選択された回数であり，$\mathrm{count}_e(w)$ は e において w が提示された回数である．最後に，この $\mathrm{score}_e(w)$ を [0, 1] に正規化する．この処理を 6 つの感情全てについて行なう．

2.2 LLM による感情強度値の付与

LLM を用いて人間による手順を再現するため，Sabbatino らと同じ 272 個の未知語を用いる．本研究では未知語にのみ関心があり，感情強度値に基づく感情内の単語の相対的な順序は相関分析に影響しないため，既存語は使用しない．この 272 個の未知語から非復元無作為抽出により 4 単語からなる組を 1,632 個作成した．この際，各単語の評価回数を Sabbatino らが行なった 120 人に対する評価と一致させるため，各単語が 24 個の異なる組に出現するように頻度調整を行なった（i.e. $\forall w$; $\mathrm{count}_e(w) = 24$）．

次に，対象となる感情 e と 4 つの未知語の組に対する入力プロンプトを作成する．評価に使用したプロンプトを表 1 に示す．システムロールでは LLM の設定や特性を指定し，ユーザロールでは質問や指示を与える．具体的には，システムロールでは LLM の特徴を Sabbatino らの研究におけるアノテータの属性に近づけるため，英語母語話者を模倣するように指示し，冗長な回答を避けるために 200 語以内で質問に答えるように指示する．ユーザロールでは，Sabbatino らと同様の質問を与え，組内の 4 つの未知語から，感情 e に最も関連する単語と最も関連しない単語を選択するように LLM に指示する．また，LLM に対し，根拠をもって段階的に考えて回答すること [11] と，後段でのテキスト抽出処理を容易にするために最も関連性の強い単語を「MOST:」，最も関連性の弱い単語を「LEAST:」に続けて回答することを指示する．

最後に，このプロンプトを LLM に入力し，2.1 節での説明と同様の方法で式 1 によって感情強度値を計算する．この処理を 6 つの感情全てに対して行なう．

3. 実　　験

未知語から想起される感情について，人間と LLM の相関関係を調査する実験を行なった．

3.1 実験設定

本実験で対象とした LLM は GPT-4 [6]（OpenAI API(注2) の `GPT-4-0613`）である．これは最も強力で，容易に利用可能な LLM の 1 つである．再現性を確保するため，出力の多様性に関係する `temperature` パラメータを 0，`top_p` パラメータを 1 に設定した．LLM と人間の感情想起の相関を測定するため，Pearson の相関係数を用いた．なお，従来の相関分析に倣い，p 値が 0.05 未満であれば，LLM による評価と人間による評価 [1] の間に有意な相関があることが示される．

3.2 結　　果

表 2 の (a) 行は，LLM による評価と人間の評価の相関係数を

(注2)：https://platform.openai.com/（2024/7/24 参照）

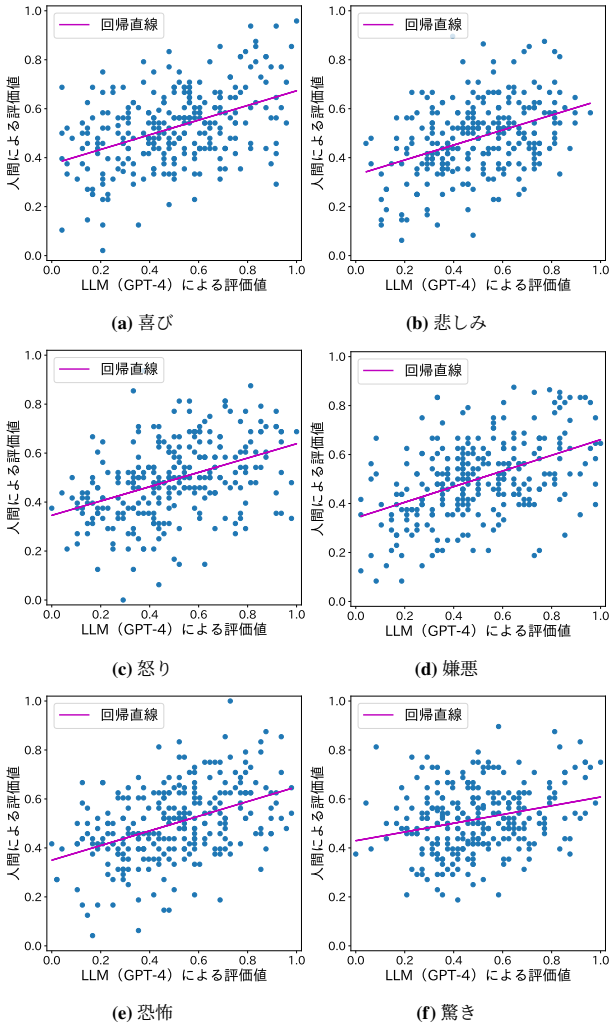

図 2: 各感情に対する LLM と人間の感情強度値の散布図.

Fig. 2 Scatter plots of LLM and human emotion intensity scores for the six emotions.

示しており，その平均値は 0.40 であった．最も高い相関が得られたのは嫌悪の感情であり，その相関係数は 0.47，最も低い相関となったのは驚きの感情であり，その相関係数は 0.26 であった．注目すべきは，すべての感情において p 値が 0.05 未満となったことである．6 感情に対する LLM と人間の感情強度値の散布図を図 2 に示す．全ての散布図に右上がりの傾向があり，LLM と人間の評価の間には正の相関があることを示している．また，表 3 は，GPT-4 による感情強度値と人間による感情強度値の平均値と標準偏差を示している．人間による評価では，驚きに対する標準偏差が最小であった．表 4 は，各感情における感情強度値の上位 10 語を示している．例えば，喜びにおいて，未知語「juy」は GPT-4 による評価と人間による評価の両方で最も高い感情強度値を得た．

表 5 は，喜びの感情において，[alse, roice, dworth, wrorgue] の未知語を与えた際の GPT-4 の出力結果の例であり，各未知語に対するベスト-ワースト・スケーリングの理由を示している．この記述から，発音が似ている既存語の意味が未知語から想起される感情の選択に影響を与えたことが分かる．このような影響が表れた出力は他にも以下のような例があった．

```
MOST: "Juy" - It's phonetically and visually
similar to "joy", which makes it the most
associated with joy in my opinion.

MOST: "Voathe" - Because it sounds similar to
"loathe", a word that signifies strong dislike
or disgust.
```

1 つ目の例は，喜びに対して未知語「juy」を含む組を評価した際の出力文の一部である．「juy」は発音と綴りが「joy」に似ているため，喜びとの関連性が高いと評価された．2 つ目の例は，嫌悪に対して未知語「voathe」を含む組を評価した際の出力文の一部である．「voathe」は，激しく嫌悪することを意味する「loathe」と響きが似ているため，嫌悪との関連性が高いと評価された．

以下は喜びに対して未知語「roice」を含む組を評価した際の出力文の一部である．

```
MOST: "Roice" - Because it sounds similar to
"rejoice", a word that is directly associated
with joy.  It also has a soft sound due to the
"r" and "oi" sounds.
```

この例の「`It also has a soft sound due to the "r" and "oi" sounds`」の部分に見られるように，GPT-4 は未知語を既存語と関連付けるだけでなく，音象徴 [2], [12], [13] に基づいてその意味を解釈した可能性がある．

3.3 考　　察

Sabbatino ら [1] の研究において，人間のアノテータ間の相関係数の平均値は 0.69 であった（表 2 の (b) 行参照）．これは，LLM を用いて得られた相関（平均 0.40）が人間のアノテータ間の相関に及ばず，人間の感情想起を十分には再現できていないことを意味する．一方，Sabbatino らの研究では，既存語で訓練した回帰モデルを未知語に適用した場合での評価も行なっており，その相関係数は 0.17 であった．本実験と彼らの実験ではテストデータの数が異なるが，相関が大きく向上していることから，回帰モデルによるアプローチよりも GPT-4 の方が人間の未知語からの感情想起をより良く再現できると考えられる．

嫌悪に対する相関が最も高くなった理由として，他の感情に比べて関連する既存語の数が多く（例えば，未知語「voathe」に対する「loathe」，未知語「fliche」に対する「filth」，未知語「groose」に対する「gross」など），GPT-4 と人間が共に嫌悪に紐付ける未知語が多いことが考えられる．驚きの相関が最も低くなった理由としては，人間による感情強度値の分散が小さいことが考えられる（表 3 参照）．図 2(f) によると，人間による感情強度値が 0.2 未満または 0.9 以上となった未知語がほぼ存在しないことから，英語話者が明確に驚きを想起する未知語または明確に驚きを想起しない未知語がほぼ存在しないことがわかる．これにより人間のアノテータ間でも意見が割れ，GPT-4 との相関も低くなったと考えられる．

3.4 結　　論

本研究では，LLM と人間の未知語理解の間の関係性を明ら

表 4: LLM と人間による各感情における感情強度値の上位 10 語.
Table 4　Words with top 10 LLM and human emotion intensity scores for each emotion.

喜び				悲しみ				怒り			
LLM（GPT-4）		人間 [1]		LLM（GPT-4）		人間 [1]		LLM（GPT-4）		人間 [1]	
juy	1.00	juy	0.96	scrown	0.96	vomp	0.90	scouge	1.00	terve	0.94
roice	0.98	flike	0.94	drouch	0.90	phlump	0.88	gnazz	0.98	shait	0.88
flalf	0.96	splink	0.94	dwull	0.90	dis	0.87	scrarsh	0.98	phrouth	0.85
swisque	0.96	glaim	0.88	sloarse	0.90	losh	0.85	bange	0.94	broin	0.81
marve	0.94	roice	0.85	wheuth	0.88	drasque	0.83	grulch	0.94	psench	0.81
spraif	0.94	shrizz	0.85	ghoothe	0.85	weathe	0.83	scrown	0.94	slanc	0.81
chymn	0.92	spreece	0.85	dwaunt	0.83	dwaunt	0.81	scruch	0.94	straif	0.81
croose	0.92	snusp	0.83	dworth	0.83	phlerm	0.79	scunch	0.94	thwealt	0.79
gream	0.92	spirp	0.83	dwoth	0.83	phreum	0.79	gnang	0.92	zorce	0.79
gwirl	0.92	drean	0.82	sulse	0.83	sout	0.79	rhulch	0.92	boarse	0.77

嫌悪				恐怖				驚き			
LLM（GPT-4）		人間 [1]		LLM（GPT-4）		人間 [1]		LLM（GPT-4）		人間 [1]	
groose	1.00	druss	0.88	scrare	0.98	phrouth	1.00	scrare	1.00	throoch	0.90
phlam	0.98	pheague	0.87	scrarsh	0.98	ghoothe	0.88	splourth	0.98	shrizz	0.88
phlelm	0.98	boarse	0.85	shrier	0.96	boarse	0.85	whoal	0.96	shrier	0.83
phlerm	0.98	snulge	0.85	dwaunt	0.94	wrorgue	0.85	bange	0.94	spreil	0.81
voathe	0.98	foathe	0.83	scouge	0.94	drasque	0.83	splink	0.94	strem	0.81
grulch	0.96	gneave	0.83	wrorgue	0.94	dwalt	0.83	splirth	0.94	swunt	0.79
psoathe	0.96	gream	0.83	grulch	0.92	keff	0.81	spraif	0.94	kease	0.77
shait	0.96	phlerm	0.83	skurnt	0.92	bange	0.79	shrier	0.92	purf	0.77
psuick	0.94	phlonch	0.83	gwotch	0.90	frete	0.79	spladd	0.92	bange	0.75
foathe	0.92	vomp	0.83	lurge	0.90	psoathe	0.77	spirp	0.90	droosh	0.75

表 5: 「喜び」を評価の対象とし，[alse, roice, dworth, wrorgue] の組を与えた場合の LLM（GPT-4）の出力.
Table 5　LLM (GPT-4) output for the tuple of [alse, roice, dworth, wrorgue] targeting the emotion “joy”.

```
The association of nonsense words with a concept like ‘‘joy’’ is subjective and depends on personal interpretation.
However, I’ll try to explain my thought process.
‘‘Alse’’ sounds similar to ‘‘else’’ and doesn’t evoke any particular emotion.
‘‘Dworth’’ sounds a bit like ‘‘dwarf’’, which could be associated with fantasy and joy for some, but it also
sounds like ‘‘worth’’ which is neutral.
‘‘Wrorgue’’ sounds harsh and complex, not typically associated with joy.
‘‘Roice’’ sounds similar to ‘‘rejoice’’, a word strongly associated with joy and happiness.
Therefore, it might be the word most associated with joy due to this phonetic similarity.
So, my choices would be:
MOST: Roice - because it sounds similar to ‘‘rejoice’’, a word that is often used to express joy or happiness.
LEAST: Wrorgue - because it sounds harsh and complex, which are not typically characteristics associated with
the feeling of joy.
```

かにすることを目的とし，GPT-4 を用いて Sabbatino らの研究の感情強度値付与手順を再現した．その結果，GPT-4 と人間による感情強度値の間には約 0.40 の正の相関があることが分かった．これは，LLM が人間の感情想起をある程度再現でき，人間が未知語から想起する感情を推定するのに LLM が有用であることを示している．GPT-4 では，綴りが類似した既存語の意味が未知語の解釈に大きく影響していることや，感情に関する音象徴の知識を利用している可能性が示唆された．

4. おわりに

本発表では，LLM の 1 つである GPT-4 [6] を用いて，未知語に対して人手で感情強度値を付与した先行研究の手順を再現して感情強度値を推定した結果，人間との間の感情強度値の一致度合いが高いことを示したが，PaLM [14] や LLaMA [15] など他の LLM では未知語に対する振る舞いが異なる可能性がある．また，Sabbatino ら [1] による分析では，120 人のアノテータによる評価を収集しているが，そのペルソナはそれぞれ異なるはずである．本研究では，LLM に対して英語の母語話者であるというペルソナを与えていたが，今後は，より多様なペルソナを割り当て，国籍や年齢などによる評価のばらつきの可能性を探る予定である．更に，LLM による未知語の解釈に影響を与える要因についても更なる調査を行ないたい．また，本研究は英

語の未知語に対する英語話者の知覚を対象としているため，異なる環境，例えば，他の言語話者の知覚で評価を行なった場合，主に英語のデータで学習されている LLM がその言語話者の評価を再現できるかどうかに関しても更なる分析が必要である．

謝辞　本研究の一部は JSPS 科研費 23K24868 の支援による．

文　　献

[1] V. Sabbatino, E. Troiano, A. Schweitzer, and R. Klinger, ""splink" is happy and "phrouth" is scary: Emotion intensity analysis for nonsense words," Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pp.37–50, 2022.

[2] W. Köhler, Gestalt Psychology, Liveright, New York, NY, USA, 1929.

[3] M.J. Traxler and M.A. Gernsbacher, Handbook of Psycholinguistics, Elsevier, Amsterdam, the Netherlands, 2006.

[4] K. Mahowald, A.A. Ivanova, I.A. Blank, N. Kanwisher, J.B. Tenenbaum, and E. Fedorenko, "Dissociating language and thought in large language models," Trends in Cognitive Sciences, vol.28, no.6, pp.517–540, 2024.

[5] Z.G. Cai, X. Duan, D.A. Haslett, S. Wang, and M.J. Pickering, "Do large language models resemble humans in language use?," Computing Research Repository, arXiv:2303.08014, 2024.

[6] OpenAI, "Introducing ChatGPT," 2022. (Accessed July 24, 2024). https://openai.com/blog/chatgpt

[7] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J.E. Gonzalez, I. Stoica, and E.P. Xing, "Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality," 2023. (Accessed July 24, 2024). https://lmsys.org/blog/2023-03-30-vicuna/

[8] S. Mohammad, "Word affect intensities," Proceedings of the 11th International Conference on Language Resources and Evaluation, pp.174–183, 2018.

[9] S. Mohammad and P. Turney, "Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon," Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp.26–34, 2010.

[10] P. Ekman, "Universal and cultural differences in facial expression of emotions," Nebraska Symposium on Motivation, vol.19, pp.207–283, 1972.

[11] T. Kojima, S.S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," Advances in Neural Information Processing Systems, vol.35, pp.22199–22213, 2022.

[12] L. Hinton, J. Nichols, and J.J. Ohala, Sound Symbolism, Cambridge University Press, Cambridge, England, UK, 1995.

[13] E. Sapir, "A study in phonetic symbolism," Journal of Experimental Psychology, vol.12, no.3, pp.225–239, 1929.

[14] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H.W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A.M. Dai, T.S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "PaLM: Scaling language modeling with pathways," Journal of Machine Learning Research, vol.24, no.240, pp.1–113, 2023.

[15] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," Computing Research Repository, arXiv:2302.13971, 2023.