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
HIP2024-55(2024-12)

大規模言語モデルによる五感表現の適合性評価の比較

小池 充† 橋本 翔†† 張 帆† 都賀 美有紀† 山崎 陽一††† 長田 典子†

†関西学院大学理工学部/感性価値創造インスティテュート 〒669-1330 兵庫県三田市学園上ヶ原 1
††西南学院大学商学部 〒814-8511 福岡市早良区西新 6-2-92
†††長崎県立大学情報システム学部 〒851-2195 長崎県西彼杵郡長与町まなび野 1-1-1

E-mail: †{koike-mk, zhangfan, toga.m, nagata}@kwansei.ac.jp ††s-hashimoto@seinan-gu.ac.jp
†††yamazaki@sun.ac.jp

あらまし 本研究では，大規模言語モデル（LLM）によって感性評価を代替する目的で，五感に関する評価表現の適合性評価と評価語間の距離測定を LLM で行った．視覚，聴覚，嗅覚の 3 つの感覚について主観評価の結果と LLM による評価値を比較し，また触覚，味覚を加えた五感すべての適合性評価の LLM による評価値を感覚別で比較して，LLM による適合性評価の特性を明らかにした．また，聴覚において音色の評価表現の階層構造を再現する目的で，4 つの尺度（1）具体 - 抽象，（2）単純 - 複雑，（3）客観 - 主観，（4）想起が困難 – 想起が容易，による語間の距離測定を LLM で行い，評価表現の階層構造を推定した．これらの結果から，一部の感覚では LLM 評価の代替可能性が示された．

キーワード 大規模言語モデル，GPT4o, LLM as a judge，主観評価

Comparison of suitability evaluation of sensory expression using Large Language Models

Makoto KOIKE† Sho HASHIMOTO†† Fan ZHANG† Miyuki TOGA†
Yoichi YAMAZAKI††† Noriko NAGATA†

†School of Science and Technology / Kwansei Gakuin Institute of Kansei Value Creation, Kwansei Gakuin University 1 Uegahara, Gakuen, Sanda-shi, Hyogo, 669-1330 Japan
††Faculty of Commerce, Seinan Gakuin University , Nishiara, Hukuoka-shi, Hukuoka, 814-8511 Japan
†††Faculty of Information Systems, University of Nagasaki 1-1-1 Manabino, Nagayo-cho, Nishisonogi-gun, Nagasaki 851-2195, Japan

E-mail: †koike-mk@kwansei.ac.jp

Abstract In this study, for the purpose of alternative evaluation of sensibility by a large-scale language model (LLM), we evaluated the conformity of evaluation expressions related to the five senses and measured the distance between evaluation words by the LLM. We compared the results of subjective evaluations of three senses (sight, hearing, and smell) with the values evaluated by LLM, and also compared the values evaluated by LLM for all five senses including touch and taste, to clarify the characteristics of conformity evaluation by LLM for each sense. In addition, in order to reproduce the hierarchical structure of auditory evaluation of tones, we measured the distance between words on four scales (1) concrete - abstract, (2) simple - complex, (3) objective - subjective, and (4) difficult to recall - easy to recall, using LLM to estimate the hierarchical structure of evaluation expressions. These results indicate the possibility of using LLM evaluation for some senses.

Keywords Large Language Models, GPT4o, LLM as a judge, human evaluation

1. はじめに

プロダクトデザインなどの商品・サービスの開発において，機能や価格といった従来の価値に加え，わくわく感や高級感といった感性的な価値が注目されている[1]．この感性的な価値を扱う手法として感性工学のアプローチが有用とされ，その具体的手法の 1 つに感性評価手法がある．この手法は，さまざまなプロダクトに関して五感を用いた主観評価を行い，機械学習や

統計学によって物理要因から印象や感性的価値までを関連付け，客観的・定量的評価を与える方法である．デザイン分野をはじめとしてさまざまな分野で基盤技術として適用されている[2]. しかし，感性評価手法では，対象の主観評価はもちろんのこと，主観評価を行うための評価語の選定や対象刺激の選定などの準備としても，一対比較法や SD 法などで多数回の主観評価を必要とする[3]. さらにドメイン（対象分野）が変わったり，モダリティ（感性的価値を感じる感覚器）が変わったりするたびに，それらに合わせた評価を行わなければならないため，人的・時間的な負荷が高い．

一方で自然言語処理分野では Transformer ベースの大規模言語モデル（LLM）の発展が多方面から注目されており，LLM を利用した対話システムなどによってパフォーマンスの向上が図られている[4]. とりわけ LLM-as-a-judge で知られるような LLM に判断や意志決定を行わせる試みが急速に拡がっており，一部では人間と同等の評価能力が確認されている[5].

そこで本研究では，LLM に感性評価を代替する方法を提案する．LLM による感性評価の傾向を人による主観評価の結果と比較することで，主観評価を代替できる可能性を検討する．

2. 関連研究

感性工学において用いられる感性評価手法[6-8]には，以下の 3 つの段階が必要である[8].

1. 対象とするプロダクトやサービスの印象や価値を記述する評価語の収集・選定を行う．このとき「自由記述」とよぶ評価語を自由に回答してもらう手続きを行い，次にそれらの評価語が対象を評価する語としてふさわしいかを判定する「適合性評価」とよぶ手続きを行い評価語を選別する．さらに評価語間の距離を定義する「距離測定」という手続きを行い，MDS（多次元尺度構成法）やクラスタリングを介して代表性・網羅性の高い評価語セットを作成する．
2. 刺激であるプロダクトに対しても評価語セットの作成と同様に，刺激の収集・選定，刺激空間の構造化を目的とした実験・分析を行うことで，代表性・網羅性の高い刺激セットを作成する．
3. 上記で得られた評価語セットを用いて，刺激セットを評価する SD（Semantic Differential）法などによる主観評価を実施し，得られた評価データに対して因子分析を行う．これにより対象の評価における主要因子を抽出し，各刺激の因子得点を算出する．結果として，因子の意味する性質（印象や価値）を各プロダクトがどの程度有しているかがわかる．

このように，感性評価では指標化の各段階において複数の実験や分析が必要となり，人的および時間的な負荷が高いという課題が生じる

また近年の機械学習による自然言語処理では Transformer ベースの LLM の発展が目覚ましく，中でも Open AI から発表された対話型 AI サービス ChatGPT が大きな関心を集めている．ChatGPT を用いてテキスト生成モデルの品質評価を代替できるかの研究[9]では，ChatGPT が人の評価と一致した傾向を示し，特に文法の正確さや，特定のテーマ・状況との関連性というタスクで高い一致度が確認された．

他にも感性的なニーズ・ゴール・ペルソナ・シナリオなどを設定する“人に寄り添う AI”の研究が進みつつある．無論 AI が真に感性を持つためには，記号接地問題（Harnad, 1990）や身体性の課題の解決を待つ必要がある．しかし工学応用としては「なぞり感性」[3]であっても大きな社会的貢献の可能性を秘めている．

そこで本研究では，LLM による感性評価の代替を目的として，感性評価手法における適合性評価および評価語間距離測定を五感別の評価語に対して行い，主観評価との比較を行い，代替可能性を検討する．

3. LLM による感性評価手法の提案

本研究では，人による感性評価を LLM に代替する方法を提案し，その可能性を検討する．感性評価の例として，五感表現の評価語に関する適合性評価および評価語間距離測定を課題として扱う．先行研究で用いた実際の主観評価データを用いて検証する．

3.1. 感性評価

感性評価は 2 章で述べたように，基本的には呈示された対象刺激に関する主観評価を回答するものである．しかし現在の AI には一般的に刺激を観察する感覚器は備えられておらず，また刺激に対する反応データが Web やデジタル空間から収集できるわけでもない．

そこで本研究では，印象や価値を記述する評価語の感性評価に着目する．感性的な評価語は記号接地はされていないものの，一定量がデジタル空間に存在しており，感性的な印象，価値，体験などの推定ができる可能性がある．今回は適合性評価課題と距離測定課題を扱う．

3.2. LLM のプロンプトチューニング

感性評価を適切に行えるよう，プロンプトチューニングを行う．プロンプトの例を図 1 に示す．プロンプト全体は，主観評価の際に参加者に呈示した教示内容とできるだけ同じ内容で構成し，主観評価と比較できるようにする．条件 2 では，（評価語）という言葉の得点をつけるよう，また回答はリッカート尺度の得点のように n 段階でつけるよう求める．ただし条件 1 では，

通常参加者への教示は「ふさわしいか」のように単極で尋ねるが，LLM へは「ふさわしいか，ふさわしくないかお伺いします」と両極で訊ね,1 語に引き摺られないようにした．

また先行研究で，GPT において Step by step で思考過程を出力させることで様々なタスクの正答率が上昇することが指摘されている[10]．そのため評価値を算出した理由を出力させるようにプロンプトを追加した．

評価語の入力については，1 度に複数の評価語を入力した場合，同時に入力した評価語の違いから評価値にバイアスがかかる可能性が考えられるため，1 回の入力で複数の評価語を入力する場合と評価語 1 語を入力した場合のそれぞれの場合で実験を行った．

以上のように作成したプロンプトを，入力の際に，各感覚や評価語にあわせて()内に入力し実行した．このプロンプトによる LLM の出力 10 回の平均を取ることで LLM 評価による評価値とした．

以下の条件の下，（評価語）という言葉の得点をつけてください．

条件 1:提示した言葉について，（各感覚）の表現としてどの程度ふさわしいか，ふさわしくないかをお伺いします．言葉が香りの表現としてふさわしいかどうかを表す得点を"step by step"で回答してください．

条件 2:回答は次の 7 段階でお答えください．
1. 全くふさわしくない
2. ふさわしくない
3. ややふさわしくない
4. どちらともいえない
5. ややふさわしい
6. ふさわしい
7. とてもふさわしい

<制約>
出力は必ず 言葉,得点を算出した理由,得点 の形式で行ってください．
出力は必ず","で区切って，"言葉，理由，得点(数字のみ)" を出力してください．
理由は","を使用せずに簡潔に出力してください．
</制約>

図 1. 感性評価のためのプロンプト
（適合性評価課題の例）

3.3. データセット

3.3.1. 適合度評価課題

主観評価による適合度評価については，視覚（プリント柄の質感）[11]，聴覚（音質）[12]，嗅覚（香料）[13]の 3 感覚に関しては，先行研究で収集した評価語および評価値を利用した．

残りの触覚と味覚に関する評価語については，それぞれの感覚を表現する言葉を LLM でプロンプトチューニングし収集した.表 1 に感覚ごとの実験参加者数，収集された評価語数，および主観評価時に参考データとして呈示した刺激数について示す．

LLM 評価による適合度評価ついては，3.2 で述べたふさわしさに関するプロンプトによって，全評価語の評価値が出力された．

なお，主観評価と LLM 評価の比較に際しては，ふさわしい評価語，すなわち評価値平均が 5 以上，標準偏差が 2 以下の語のみを用いる．これらについての適合率（precision），再現率（recall），および F-1Score を算出した．

表 1　収集した評価表現数

	視覚（柄評価）	聴覚（音質評価）	嗅覚（香料評価）	触覚	味覚
実験参加者数	10	50	10	N/A	N/A
評価語数	345	53	627	65	174
提示刺激数	20	0	52	N/A	N/A

4. 実験結果

3 つの実験を実施した．1 つ目は適合度評価課題において，視覚，聴覚，嗅覚の 3 感覚について，主観評価と LLM 評価の比較を行った．2 つ目は，同じく適合性評価課題において，五感全てについて，LLM 評価において評価値の高い語を抜粋し確認した．3 つ目は距離測定課題について，主観評価と LLM 評価の比較を行った．

4.1. 適合性評価課題(3 感覚)による H-L 比較

まず，主観評価,1 回の出力で複数の評価語を一括評価した LLM（whole）評価,1 回の出力で評価語 1 語を評価した LLM（each）評価の 10 回実行時の得点の頻度分布を図 2 に示す．

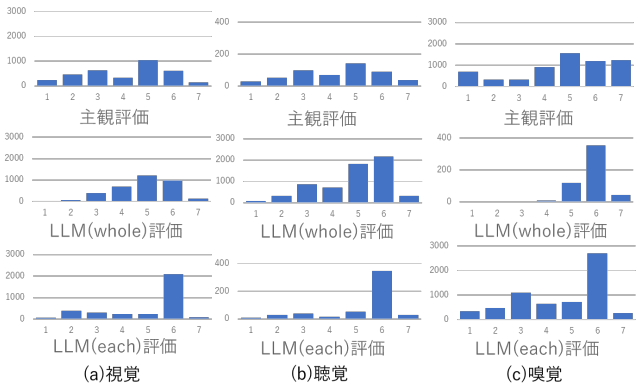

図 2. 評価値の頻度分布

次に，主観評価と LLM(whole)評価，主観評価と LLM（each）評価の評価語ごとの平均値の分布図とそ

の相関係数を図 3 に示す．また，主観評価（human evaluation）の分散値（H）と，主観評価実験の参加者間の相関平均（H），主観評価実験参加者と LLM 評価の相関平均(H- L)について表 2 に示す．

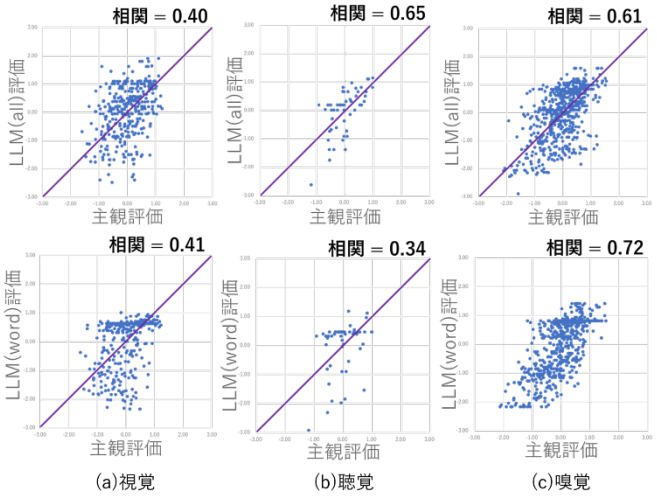

図 3. 評価語ごとの平均値の分布（主観 -LLM）

表 2. 適合性評価 比較

		視覚（柄評価）	聴覚（音質評価）	嗅覚（香料評価）
主観評価	分散値(H)	2.464	2.387	2.144
	相関平均(H)	0.270	0.223	0.365
LLM(whole)評価	分散値(L)	1.216	0.317	1.822
	平均値の相関(H-L)	0.404	0.645	0.613
	相関平均(H-L)	0.210	0.248	0.355
LLM(each)評価	分散値(L)	2.488	1.904	2.808
	平均値の相関(H-L)	0.406	0.342	0.717
	相関平均(H-L)	0.209	0.153	0.431

表 4. 感覚別 LLM 評価の評価指標

	視覚(柄評価)	聴覚(音質評価)	嗅覚(香料評価)
Precision	0.238	0.244	0.647
Recall	0.869	0.909	0.822
F1-Score	0.373	0.385	0.724

さらに LLM の評価傾向を明らかにするため,主観評価と LLM 評価の評価値平均の差が大きい，もしくは小さい語について, LLM の出力理由例をまとめた（表 3）．

最後に，LLM 評価の結果からふさわしいと評価された語を選定し，適合率（precision），再現率（recall），F-1Score を算出した結果を表 4 に示す．

4.2. 適合性評価課題（五感）による LLM 評価

五感を表す評価表現すべてについて，LLM 評価の評価値の上位 5 語を挙げ，表 5 に示す．

4.3. 距離測定課題（階層 4 尺度）による H-L 比較

距離測定課題では LLM(each)を用いて，LLM 評価を 10 回実行した．主観評価と LLM 評価の評価語ごとの平均値の分布図とその相関係数を図 4 に示す．

表 5. LLM 評価の上位表現

	評価語	LLM平均	LLM 出力理由例
視覚	木目調の	6.60	木目柄の表現として非常にふさわしいため
	対称的な	6.50	柄画像において左右や上下が対称である場合に適しているため
	彩度の高い	6.50	色鮮やかな柄画像に適しているため
	美しい	6.40	柄画像の評価として適しているため
	派手な	6.40	柄画像は視覚的に目立つことが多いため
聴覚	静かな	7.00	音質が低音量や無音に近い状態を示すためふさわしい
	騒々しい	6.90	音質に対して適切な表現であるため
	うるさい	6.70	音質の評価に適しているため
	かん高い	6.30	高音を表現するのに適している
	こもった	6.10	音質がこもっているという表現は一般的に音が明瞭でないことを意味するため音質の評価にふさわしい
嗅覚	柑橘系	7.00	柑橘系の香りは一般的に知られているため香りの表現として非常に適切
	ハッカ	7.00	清涼感のある香りが一般的に認識されているため
	花の香り	7.00	香りの表現として非常に一般的であるため
	花の芳しい香り	7.00	花の香りを具体的に表現しているため
	ペパーミントのような	7.00	ペパーミントは一般的に香りの強いハーブでありその香りは多くの人に認識されているため香りの表現としてふさわしい
触覚	柔らかい	7.00	触覚の表現として非常に一般的であり多くの人が共感できるため
	固い	7.00	触覚の表現として非常に一般的であり物体の硬さを直接示す言葉
	ふわふわした	7.00	柔らかく軽い触感を連想させるため
	つるつるした	7.00	触覚の表現として非常に適切であり滑らかで摩擦が少ない感覚を直接的に表現しているため
	でこぼこした	7.00	触覚的な凹凸を感じる表現だから
味覚	甘い	7.00	味覚の表現として非常に一般的であり、砂糖や果物などの味を直接連想させるため
	苦い	7.00	味覚の表現として非常に一般的であり、特定の味を直接的に示す
	酸っぱい	7.00	味覚の表現として非常に一般的であり、特に酸味を表現する際に使用されるため
	塩辛い	7.00	味覚の表現として非常に一般的であり、特に塩味を強く感じる食べ物に対して使われるため
	辛い	7.00	味覚の表現として一般的に使用されるため

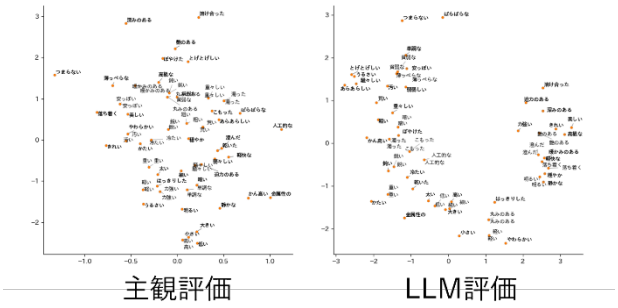

図 4. 評価語ごとの平均値の分布（主観 -LLM）

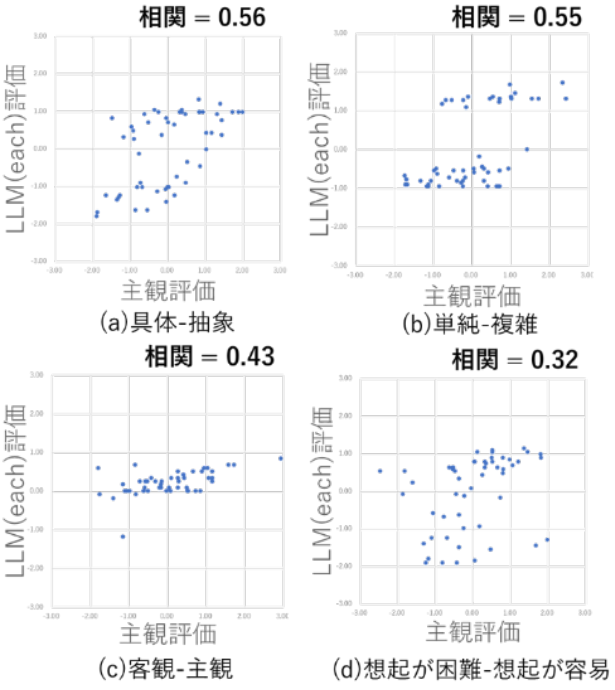

図 5. 距離測定からの階層構造推定結果

表 3. LLM 評価 比較結果

五感	差の大小	差の符号	評価語	評定平均値			LLM出力理由例
				主観評価	LLM評価	差	
視覚	大	正	嫌いな	4.90	1.30	3.60	ネガティブな感情を持つ言葉であり柄画像には適さない
			婆臭い	4.20	1.30	2.90	年配の女性に対する侮蔑的な表現であり柄画像の評価としては適切でない
			うるさい	5.30	2.50	2.80	音や騒音を連想させるため適切ではない
		負	ギャルっぽい	2.00	5.90	-3.90	若者文化や派手なデザインを連想させるため
			喜ばしい	2.10	5.70	-3.60	ポジティブな感情を表現する言葉で柄画像にも適している
			キラキラしている	2.60	6.10	-3.50	視覚的に輝くイメージが強いため柄画像に適している
	小	正	冷静な	3.80	3.80	0.00	柄画像の表現としてはあまり感情を表さないため
			寒色の	6.00	5.90	0.10	柄画像の評価として色彩の冷たさや落ち着きを表現するのに適しているため
			漸進的な	4.50	4.40	0.10	柄画像の表現としては徐々に変化する様子が含まれるため
		負	きれいな	5.90	6.00	-0.10	柄画像の評価としての表現が抽象的で直接的な視覚的連想が難しいため
			古臭い	4.70	4.80	-0.10	視覚的に不明瞭な柄を連想させるため
			優しい	4.70	4.80	-0.10	柄画像が古風なデザインの場合に適している可能性があるため
聴覚	大	正	大きい	5.58	3.20	2.38	音質の評価にはあまり直接的ではないため
			小さい	4.47	2.80	1.67	音質の評価としては曖昧で具体性に欠ける
			弱々しい	4.24	2.60	1.64	音質の評価としてはあまり適していないため
		負	溶け合った	3.02	5.80	-2.78	音が滑らかに融合するイメージがあるため
			高級な	3.41	5.90	-2.49	音質が上品で質が高い印象を与えるため
			静かな	4.60	7.00	-2.40	音質が低音量や無音に近い状態を示すためふさわしい
	小	正	軽い	4.63	4.60	0.03	音質が明るく軽快な感じを連想させるため
			あらあらしい	5.49	5.40	0.09	音質が粗く感じられるため
			冷たい	3.60	3.50	0.10	音質の評価としては抽象的で温度を連想するため
		負	迫力のある	5.94	6.00	-0.06	音質の表現として適切であり力強さを感じる
			力強い	5.90	6.00	-0.10	音質の表現として適切であるため
			低い	5.58	6.00	-0.42	音質の評価において低音を指す際に適している
嗅覚	大	正	渋みのある	5.50	2.33	3.17	味覚に関連する表現で香りには直結しないため
			苦味のある	5.30	2.20	3.10	味覚の表現で香りには直接関係しないため
			青い	4.60	1.60	3.00	色に関する表現であり香りの表現としては一般的ではない
		負	アンジェリカ	3.00	6.00	-3.00	ハーブとしての香りが知られているため
			ふくよかな	3.30	5.90	-2.60	香りの層が多いと感じるため
			淑女	3.30	5.50	-2.20	上品なイメージが香りと結びつくため
	小	0	田舎の香り	4.90	4.90	0.00	具体的な香りが想像しにくいため
			祈り	2.40	2.40	0.00	香りの具体的なイメージと結びつきにくい
			エネルギーを感じる	4.90	4.90	0.00	動的で活気のある香りを想像させるため
		正	スーっとする	6.10	6.00	0.10	清涼感があり特定の香りに関連付けやすい
			深みのある	6.10	6.00	0.10	香りに対して一般的に否定的な意味を持つため
			頭がすっきり	5.80	5.70	0.10	清涼感やリフレッシュ感を連想させる香りの表現として比較的ふさわしい
		負	ムスク	6.20	6.30	-0.10	ミントは一般的に香りの表現として使用されるため
			官能的な	5.90	6.00	-0.10	感覚的な魅力を持つ香りに対して使われることがあるため
			クール	5.90	6.00	-0.10	冷たさや爽快感を連想させるため香りの表現として適している

さらに LLM による距離測定の結果から評価語間の心理的距離を求め，多次元尺度法によって評価語をクラスタリングし図上に付置した結果の主観評価との比較を図 5 に示す．

5. 考察

5.1. 適合性評価課題(3 感覚)による H-L 比較

図 2 から全般的に主観評価による視覚・聴覚・嗅覚の 3 感覚の分布が相似しているのに対して，LLM 評価による分布は尖度，歪度が高い異なった分布になっていた．また表 2 で示したように LLM(whole)評価では，主観評価に比べ，分散値が極端に低くなっていたが，これは LLM(each)評価を用いることで，主観評価に近い分散値をとるように改善された．しかし，LLM 評価は全体的に主観評価に比べ過大評価傾向が見られ，視覚 77.1%，聴覚 100%，嗅覚 56.8%の語が過大評価された．このことは，LLM(whole)評価において，高い評価値を中心に評価値が分布していることや，LLM(each)評価において評価値 6 の出現頻度が極端に多くなっていることなどからも読み取れる．

次に，主観評価と LLM 評価の評価語ごとの平均値の相関は，すべて正の相関を示したため，両者の似た傾向を発見はできたが，LLM の評価値に頻度の多い値が出ている点から，なにかしらのバイアスがかかっている可能性が示された．

評価値の差から判断できる LLM 評価の傾向については，LLM にネガティブなイメージを持つ語の評価値を低くし，逆にポジティブなイメージを持つ語は高くする傾向が見られた．また，比喩的，共感覚的な意味を持つ言葉は主観評価との差が大きく，辞書的な意味しか持たない言葉は主観評価と一致した評価を行えることが明らかになった．LLM は身体性を持たず，言葉の意味の連鎖によって出力を生成するため，複数のモダリティを表す表現や主観的な体験に基づく比喩的な

表現などの評価は主観評価と大きく異なる可能性を考慮する必要がある.

最後に, LLM 評価の評価指標では, 全ての感覚において Precision < Recall となっており, 分布からの考察同様に, 過大評価傾向が見られた. しかし, Recall の値は高く, 主観評価で選定された語の多くは LLM 評価でも選定されるため, 主観評価実験を行う前の補助的な活用であれば LLM 評価を十分に利用できると考える. また, 嗅覚においては F1-Score が 0.724 と高い値をとっていることから, 収集した評価語の傾向によっては, 主観評価を代替できる可能性が示唆された.

5.2. 適合性評価課題(五感)による LLM 評価

五感別の LLM 評価の評価値の上位の評価語については, 嗅覚以外の感覚において, 物理的な特徴を示す評価語だけでなく主観的な印象を示す評価語が上位に現れており, 感性の指標化手法において活用できる評価だと判断できる.

5.3. 距離測定課題(階層 4 尺度)による H-L 比較

LLM による言葉の距離測定では,「想起が困難-想起が容易」という指標において, 相関が顕著に低くなり, 概念の抽象性が増した場合に, Step by Step で理由を出力させたとしても LLM 評価で主観評価の傾向を再現することは難しいことが明らかになった. しかし, 図 6 に示した階層構造では, 分布はかなり異なっているものの, 主観評価で近くに付置された似た意味を持つ言葉は LLM 評価でも近くに付置しており, LLM 評価の際に概念を具体的に記述できれば, 階層性を推定できる可能性が示唆された.

6. 結論

本研究では, 大規模言語モデル(LLM)を用いて, 感性評価に必要な評価表現の適合性評価, 評価語間の距離測定を, 五感表現別に行い, LLM 評価の特性を明らかにし, 主観評価の代替可能性を示した.

まず, LLM による評価においては,感覚や課題に依らず過大評価する傾向にあり, 特定の出力の頻度が高くなる傾向を確認した. しかし, 嗅覚において主観評価の結果との間に強い正の相関がみられ, 評価語の選定で高い再現率を示したことから, 主観評価を代替できる可能性が示唆された.

また, ネガティブもしくはポジティブなイメージが強い表現や, 比喩的, 共感覚的な表現の評価において人と異なる傾向にあった. LLM が身体性を持たず,辞書的意味に基づく出力を行っている可能性が示された.

今後は, さらに人と同じ傾向を示すように, 過大評価傾向を解決した LLM 評価のためのプロンプトチューニングや評価語収集の手法などを検討したい.

謝辞

本研究は科学研究費 22H03681 の支援を受けた.

文　献

[1] 和泉志穂, and 赤岡仁之, "消費者行動における感性価値の研究—複数の感覚項目の関係性および性差・世代差からの検討—." *繊維製品消費科学* 56.7, pp. 613-619, 2015

[2] 横尾俊輔, 柳澤秀吉, 村上存, 大富浩一, 穂坂倫佳, “製品音のデザインにおける和音性特徴量の感性評価”, 日本デザイン学会研究発表大会概要集, 第 57 回研究発表大会, 2010.

[3] 井口征士, 猪田克美, 小林重順, 田辺新一, 長田典子, and 中村敏枝. “感性情報処理”, in 電子情報通信学会編ヒューマンコミュニケーション工学シリーズ, オーム社, 1994.

[4] 坂本有紀, 内田貴久, 石黒浩, “大規模言語モデルを用いたユーザモデル推定機能を持つ対話システムの検討,” 人工知能学会全国大会論文集, vol. 37, p.3O1OS2c01, 2023.

[5] Zheng, Lianmin, et al. "Judging llm-as-a-judge with mt-bench and chatbot arena." Advances in Neural Information Processing Systems 36 (2023): 46595-46623.

[6] 長町三生, 商品開発と感性. 海文堂出版, 2005.

[7] 柳澤秀吉, 村上存, 大富浩一, and 穂坂倫佳, “感性の多様性を考慮した感性品質の定量化手法”, 日本機械学会論文集 C 編, vol. 74, no. 746, pp. 2607-2616, 2008.

[8] 山田篤拓, 橋本翔, & 長田典子, “レビューデータを用いた評価表現辞書に基づく印象の自動指標化”, 日本感性工学会論文誌, 17(5), 567-576, 2018

[9] Chiang, Cheng-Han, and Hung-yi Lee. "Can large language models be an alternative to human evaluations?.", arXiv preprint arXiv:2305.01937, 2023.

[10] WU, Yang, et al. Improving cross-task generalization with step-by-step instructions. arXiv preprint arXiv:2305.04429, 2023.

[11] N. Sunda, K. Tobitani, I. Tani, Y. Tani, N. Nagata, & N. Morita. “Impression estimation model for clothing patterns using neural style features.” In HCI International 2020-Posters: 22nd International Conference, HCII 2020, Part III 22 (pp. 689-697). Springer International Publishing, Copenhagen, Denmark, July, 2020.

[12] 浅川香, 矢野敦仁, 木村勝, 片平建史, 山﨑陽一, & 長田典子. “車室内エンジン加速音及び定速走行音の聴取時における感情評価の個人特性”, 日本音響学会誌, 77(11), 694-697, 2021

[13] 村上柚香, 都賀美有紀, 長田典子, 綿村豪, 三瓶和也, 寺本圭吾, 天然・合成香料のための感性指標の構築と好ましさとの関係性に基づく個人の類型化, 信学技報, 122(367), MVE2022-37, 15-20, 2023