言語モデルにおける語と色の連想関係のアライメント

石倉誠也　荒瀬由紀
東京科学大学
ishikura.s.1771@m.isct.ac.jp　arase@c.titech.ac.jp

概要

「暑い」と聞くと赤色を思い浮かべるように，語は特定の色のイメージを喚起することがある．本研究は 6000 語以上のデータを用いて，言語モデルの語–色の連想関係を人間の連想傾向と比較した．一般的なタスクとは異なり，語–色連想の一致度はモデルの大規模化に伴って必ずしも改善せず，モデルファミリーによっては低下する例も見られた．語の特性に基づいた分析から，形容詞・名詞，および具体的な語ほど人間の傾向と一致しやすいことがわかった．さらに，人間の語–色連想を追加学習すると，学習前と比較して感情分類や比喩の意味理解の性能が向上する傾向があることを確認した．

1　はじめに

「りんご」から赤色を連想し，「冷たい」から青色を連想するように，その言葉が具体的か抽象的かにかかわらず特定の色のイメージを喚起することがある．このような語と色の連想関係は，情報提示や可視化においてメッセージ理解を助けたり感情反応を誘発したりする要因になっている [1]．

語と色の連想関係を扱う先行研究として，Mohammad はクラウドソーシングにより英語の大規模な語–色連想データを構築した [1]．Harashima らはこれを日本語へ拡張し，2,903 語の語–色連想関係を収集するとともに，語を文脈付きで提示した場合と文脈なしで提示した場合を比較した [2]．また，語と色の対応を人手回答ではなく実世界から推定する試みとして，書籍の表紙画像に含まれる文字の色から単語ごとの色分布を推定した研究 [3] や Web 上の画像をもとに語から連想される色を推定する研究 [4] も報告されている．近年は，大規模言語モデル（LLM）や視覚言語モデル（VLM）が人間の語–色連想をどの程度再現するかの検証も進みつつあり，Fukushima らは日本語 80 語に対する人間の回答と GPT 系列モデルの回答を比較した [5]．

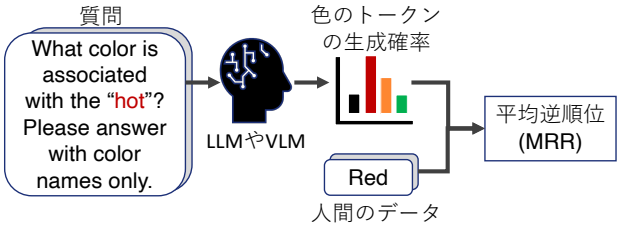

図 1　LLM/VLM での語と色の連想関係調査の概略図

これらの先行研究を踏まえ，本研究には次の 2 つの新規性がある．(i) より大規模な英語の語–色連想データを用いて，LLM/VLM がもつ連想傾向を人間の傾向と比較する点，(ii) 比較にとどまらず，人間の連想傾向を追加学習させた場合の下流タスクへの影響を検証する点である．

本研究は言語モデル内での語と色の連想関係に関して，次の 2 つのリサーチクエスチョンを設定する．(1) モデルのサイズや LLM/VLM の違いが，人間の語–色連想との一致度にどのように影響し，また一致しやすい語にはどのような特徴があるのか．(2) 人間の語–色連想関係をモデルに追加学習させることで，語が喚起する感覚的イメージの理解が補強され，人間の感性が強く影響するタスク（例：感情分類，比喩の意味理解）の性能向上につながるのか．実験の結果，(1) に関してサイズ別や LLM と VLM の比較から人間の連想関係と一致しやすいモデルの一貫した特徴は見られなかった．しかし，語の特性に基づいた分析から，形容詞や具体的な語は人間の連想傾向と一致しやすいということがわかった．また (2) に関して，人間の語–色連想を学習させることで，学習前と比較して感情分類および比喩の意味理解タスクの性能が向上することを確認した．

2　言語モデルの語–色連想関係調査

2.1　実験設定

本研究では，英語の語–色連想関係を大規模に収集した Mohammad のデータセットを用いる．こ

のデータセットはクラウドソーシングにより構築され，各評価者は提示された単語に対して，11 色 (white, black, red, green, yellow, blue, brown, pink, purple, orange, grey) の色名の中から最も連想される色を 1 つ選択することで，語と色の対応を付与している．データセットには，各語について投票総数や最多票を得た色とその得票数，語義が記録されている．すべての語が明確な色連想を持つとは限らない．そこで本研究では，人手投票において半数以上が同一の色に投票した語のみを抽出し，モデルの連想傾向の評価対象とした．また，語義が複数種類存在する語も除外した．最終的にモデルの評価に使用した語数は 6,298 語である．

具体的な調査方法は図 1 のとおりである．各単語 `word` に対して，ユーザプロンプト `What color is associated with the "{word}"? Please answer with color names only.` をモデルへ入力し，直後に生成される 11 色の色名の生成確率を抽出する．抽出した確率分布に基づき，11 色を降順にランキングし，モデルの語–色連想の順位を得る．

評価指標には平均逆順位（MRR）を用いる．各語について，人手投票で最多票を得た色をもとに MRR を計算する．なお，評価語彙は「半数以上が同一色に投票した語」に限定しているが，語によっては条件を満たす色が 2 色存在する場合がある．この場合，評価時には 2 色のうちモデルのランキングでより上位に現れた色を用いて MRR を算出する．

パラメータサイズの増加に伴う語–色連想の変化を観察するため，4 種類のモデルファミリーを対象とした．Llama-3.1 [6]，Qwen2.5 [7]，Qwen2.5-VL [8]，gemma-3 [9] を用いる．さらに，LLM と VLM の比較を行うために，言語モデルのデコーダが共通である次の 3 組のペアを用意し比較する：(a) Llama-3.1-8B-Instruct vs Llama-3.2-11B-Vision-Instruct (Llama-3.x-LV)，(b) Phi-4-mini-instruct [10] vs Phi-4-multimodal-instruct (Phi-4-LV)，(c) Qwen2.5-7B-Instruct vs Qwen2.5-VL-7B-Instruct (Qwen2.5-LV).

2.2 結果

モデルサイズによる影響 モデルサイズ別に MRR の推移を確認すると（図 2），Qwen2.5 と Qwen2.5-VL ではモデル規模の増加に伴い MRR が上昇し，語–色連想関係が人間の傾向へ近づくことが観察された．一方で，他のモデルファミリーでは，規模が大きいほど一貫して人間の連想に近づくという傾向は確認できなかった．Llama-3.1 では，パラメータ数に大きな差があるものの，大きいモデルのほうが MRR は低く，より人間の連想傾向から遠ざかる結果となった．また，モデルファミリー間の比較をすると，gemma-3 は小規模モデルであっても，他の小規模モデルと比べて人間の語–色連想傾向と一致しやすいことを示した．

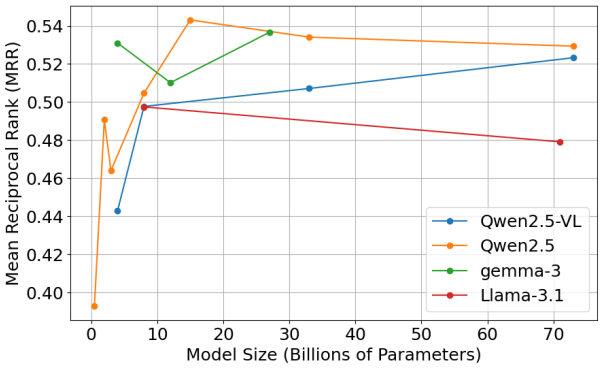

図 2 モデルファミリーやサイズ別の MRR の推移図

表 1 LLM と VLM の語–色連想関係の比較

	Llama-3.x-LV	Phi-4-LV	Qwen2.5-LV
LLM	0.497	0.494	**0.505**
VLM	**0.507**	**0.495**	0.498

LLM と VLM の比較 次に，デコーダが同一である LLM と VLM のペアを比較した（表 1）．どのモデルファミリーにおいても LLM と VLM で顕著な差は確認できなかった．このことは，画像を用いた事後学習が，人間の語–色連想関係の再現に対して，一貫した改善効果を示さないことを示唆している．

2.3 語の特性に基づいた分析

本節では，語の属性が語–色連想の一致度に与える影響を調べるため，品詞，語の具体性に基づく分析を行う．いずれの分析でも，各モデルに対して対象語彙を該当カテゴリに分割し，カテゴリごとに MRR を算出したうえで，これまで調査した 19 種類のモデルにおける MRR の平均値を報告する．

まず品詞別の分析では，各語の品詞を WordNet [11] に基づいて特定した．複数の品詞を持つ語については，それぞれの品詞に属する語として重複を許してカウントし，品詞ごとに MRR を算出した．その結果は表 2 のとおりである．形容詞が最も人間の語–色連想傾向と一致しやすく，次に名詞の MRR が高いことが分かった．

語の具体性に基づく分析では，具体性スコアが付与されたデータセット [12] を用いた．このデータ

表 2　語のカテゴリ別の MRR

カテゴリ	語数	MRR
名詞	4472	0.503
動詞	2303	0.484
形容詞	1762	**0.513**
副詞	168	0.466
具体性が高い語	1580	**0.566**
具体性が低い語	855	0.477

表 3　語–色の連想関係学習の結果

学習の前後	Qwen2.5-7B	Qwen-VL-7B	gemma-3-12b
学習前	0.514	0.515	0.503
学習後	**0.656**	**0.658**	**0.652**

セットには，人間の五感に基づいた語の具体性スコアが 1–5 の範囲で記録されている．具体性スコアが 4 以上の語と 2 以下の語の 2 群に分割して MRR を比較した．結果は表 2 のとおりである．具体的な語のほうが LLM や VLM も人間に近い色を連想しやすいことを確認した．海などの液体に関わる語や，野菜のように実体をもつ語が，人間の連想傾向と一致しやすいことが確認された（詳細は付録 A）．

3　語と色の連想関係の学習

データセット　学習には，連想関係調査と同様に，人手投票において半数以上が同一の色に投票した語のみを用いる．ただし，教師信号を単一ラベルとして扱うため，最多得票の色が 2 つ存在する語は除外した．このフィルタリング後の語彙に対して，色ラベルの分布を維持しながら，学習: 検証: 評価を 8:1:1 の比率で分割した．

モデル　2.2 の結果を受けて，ベースモデルで人間との連想関係の一致率が高い Qwen2.5 と Qwen2.5-VL，gemma-3 から以下の 3 つのモデルを学習対象として用いた．

- Qwen2.5-7B-Instruct (Qwen2.5-7B)
- Qwen2.5-VL-7B-Instruct (Qwen-VL-7B)
- gemma-3-12b-it (gemma-3-12b)

学習方法　学習は指示チューニングの枠組みで行い，QLoRA [13] を用いて学習した．詳細な設定は付録 B に記載した．

結果　評価データ上での学習前後の MRR の結果は表 3 の通りである．すべてのモデルにおいて学習後に MRR が上昇した．この結果は，LLM および VLM が，人間が付与した語–色連想関係を追加学習によって獲得し得ることを示している．

4　語と色の連想関係学習の影響調査

本節では，モデルに人間の語–色連想関係を学習させることが，人間の感性や知覚の寄与が大きいタスクにどのような影響を与えるかを検証する．具体的には，感情分類タスク 2 種と比喩の意味理解タスクを対象に，連想学習の前後での性能を比較することで，語–色連想の追加学習が下流タスク判断に及ぼす効果を定量的に評価する．

比較のため，次の 3 つの学習設定を用意する．

- **ベースモデル**：語–色連想関係に関する追加学習をしなかったモデル
- **ランダムな色での学習**：各語に対して，人間が連想しない色を割り当てて追加学習したモデル
- **人間の色での学習**：各語に対して，人間が連想する色を用いて追加学習したモデル

4.1　GoEmotions

実験設定　感情分類タスクにおける性能比較のため，GoEmotions [14] を用いる．GoEmotions は Reddit のコメントを対象としたデータセットであり，各コメントに対して 28 種類（neutral を含む）から 1 つ以上の感情ラベルが人手で付与されている．本研究では，評価の単純化と比較のために，28 種類の感情ラベルを Ekman [15] の感情（anger，disgust，fear，joy，sadness，surprise）と neutral へ変換したうえで，変換後に 2 つ以上の感情ラベルが付与されるサンプルを除外し，最終的に 4,968 件の評価データを作成した．

評価時には入力テキストと候補となる感情ラベルの選択肢をモデルに提示し，最も適切なラベルを 1 つ選択させる形式で評価する．プロンプトは評価データと同様のフィルタリングを行った検証データでの Qwen のベースモデルで一番性能が良かったものを採用した．モデルが `A`，`(A)`，あるいは `(A) anger` のように，回答を生成した場合には，その記号に対応するラベルを予測結果として採用する．一方，これらの形式を満たさない場合には，出力の先頭トークンにおける記号集合の生成確率分布を確認し，最も確率が高い記号に対応するラベルをモデルの予測として決定する．元の論文と同様に評価指標として macro-F1 を採用した．

結果　表 4 に GoEmotions での結果を示す．Qwen2.5-7B ではベースモデルが最も良い性能を示したものの，Qwen2.5-VL-7B と gemma-3-12b で

表 4 感情分類タスクでの結果（Qwen=Qwen2.5-7B, Qwen-V=Qwen2.5-VL-7B, gemma=gemma-3-12b）

	GoEmotions			enISEAR		
学習設定	Qwen	Qwen-V	gemma	Qwen	Qwen-V	gemma
ベース	**0.404**	0.340	0.396	0.765	**0.809**	0.823
ランダム色	0.387	0.343	0.390	0.769	0.800	0.825
人間色	0.395	**0.351**	**0.410**	**0.788**	0.802	**0.842**

は人間が連想する語と色の連想関係を用いた学習が最も良い結果となった．また，ランダム色と人間色を比較すると，人間色のほうが高い性能を示した．

4.2 enISEAR

実験設定 語–色連想関係の追加学習が感情分類に与える影響を，追加で検証するために enISEAR [16] を用いた評価を行う．enISEAR は，ある感情に関連するイベントの記述を収集したデータセットであり，作成時に 2 段階のアノテーションを行っている．第 1 段階では，アノテーターに特定の感情を提示し，その感情を経験したイベントを記述させる．第 2 段階では，第 1 段階での記述を別のアノテーターに提示し，どの感情のイベントかを推測させる．本研究では，第 2 段階の評価を正解ラベルとして用いた．ラベルは anger，joy，shame，sadness，fear，guilt，disgust の 7 種類であり，neutral に相当するラベルを含まない．

データ前処理として，ある感情ラベルへの投票数が全体の過半数を占めるサンプルのみを抽出した．その後，ラベル分布を考慮しつつ，検証用と評価用データを 1:1 で分割した．分割後の評価データは 481 件である．プロンプトの選択および所定形式以外の出力の処理は GoEmotions での実験と同じ方針で行った．評価指標としては元の論文と同じ micro-F1 を採用した．

結果 表 4 に enISEAR における結果を示す．人間色での学習はベースモデルと比較して，Qwen2.5-VL-7B では性能がわずかに低下したが，Qwen2.5-7B および gemma-3-12b では改善が見られた．GoEmotions 同様に，2 行目と 3 行目を比較すると，人間色のほうが高い性能を示した．

GoEmotions の結果と合わせると，人間の語–色連想関係を追加学習することで，語が喚起する抽象的イメージが人間の傾向に近づき，感情分類の判断が改善する傾向があることがわかる．

4.3 MUNCH

実験設定 比喩表現は，色などの人間が知覚できるものを抽象概念へ写像して理解される場合がある（例：「腹黒い」）．したがって，語–色連想関係という感覚的知識は，比喩の解釈に寄与し得ると考えられる．モデルの比喩の意味理解の性能を測定するために，MUNCH [17] を使用する．このデータセットは，比喩を含む文章と，文中の比喩表現に対する言い換え候補 2 つから構成される．各候補には，言い換えとして適切または不適切のラベルが付与されており，モデルには各文章について 4 通りの選択肢（片方のみ適切が 2 通り，両方適切，両方不適）が提示される．評価データは 1,492 件である．

表 5 MUNCH での結果

学習設定	Qwen2.5-7B	Qwen-VL-7B	gemma-3-12b
ベース	0.333	**0.414**	0.471
ランダム色	0.302	0.349	0.486
人間色	**0.351**	0.388	**0.523**

元論文では比喩への言及の度合いに基づく 3 タイプの設定があり，各タイプに 3 種類のプロンプトが存在する（合計 9 プロンプト）．本研究では，これら 9 つのプロンプトを対話形式に整形し評価で使用した．評価指標は元論文と同じ正解率を採用する．

結果 表 5 に結果を示す．数値は 9 つのプロンプトから得られた正解率の平均値である．人間色での学習はランダム色で学習するよりも良い性能を示す．一方で人間色とベースモデルを比較すると，Qwen2.5-VL-7B では性能低下がみられ，Qwen2.5-7B と gemma-3-12b では性能が向上した．

5 おわりに

本研究では，言語モデルの語–色連想に関して，モデルサイズ，LLM/VLM の違い，語の特性に基づき，人間の連想傾向との一致度を比較した．その結果，語–色連想の一致度はモデルの大規模化に伴って必ずしも改善しないことや，LLM/VLM の間に一貫した優劣がないことを確認した．語の特性に基づいた分析から，形容詞や具体的な語ほど人間の傾向と一致しやすいことがわかった．さらに，人間の語–色連想を追加学習すると，学習前より感情分類や比喩理解の性能が向上する傾向があることがわかった．

今後は日本語データや CLIP でのテキストと画像を用いた実験など，言語やモダリティを拡大して，語–色連想学習の影響をより広範に調査していく．また，語のニュアンスを正確に捉えるために単一ラベルの語–色の連想データだけではなく，色連想の分布を考慮したデータでの学習を検討していく．

謝辞

本研究は JSPS 科研費 JP25H01149 の助成を受けたものです．また，本研究は東京科学大学のスーパーコンピュータ TSUBAME4.0 を利用して実施した．

参考文献

[1] Saif Mohammad. Even the abstract have color: Consensus in word-colour associations. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, 2011.

[2] Jun Harashima. Japanese Word—Color associations with and without contexts. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**, 2016.

[3] Masaya Ikoma, Brian Kenji Iwana, and Seiichi Uchida. Effect of text color on word embeddings. In **Document Analysis Systems**, 2020.

[4] Bhargav Srinivasa Desikan, Tasker Hull, Ethan Nadler, Douglas Guilbeault, Aabir Abubakar Kar, Mark Chu, and Donald Ruggiero Lo Sardo. comp-syn: Perceptually grounded word embeddings with color. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 1744–1751, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[5] Makoto Fukushima, Shusuke Eshita, and Hiroshige Fukuhara. Advancements and limitations of llms in replicating human color-word associations. **Discover Artificial Intelligence**, Vol. 5, p. 64, 2025.

[6] Llama Team. The llama 3 herd of models, 2024.

[7] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[8] Qwen Team. Qwen2.5-vl, January 2025.

[9] Gemma Team. Gemma 3 technical report, 2025.

[10] Microsoft. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025.

[11] George A. Miller. WordNet: A lexical database for English. In **Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992**, 1992.

[12] Marc Brysbaert, Amy Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. **Behavior research methods**, Vol. 46, , 2013.

[13] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In **Advances in Neural Information Processing Systems**, 2023.

[14] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, 2020.

[15] Paul Ekman. An argument for basic emotions. **Cognition & Emotion**, Vol. 6, pp. 169–200, 1992.

[16] Enrica Troiano, Sebastian Padó, and Roman Klinger. Crowdsourcing and validating event-focused emotion corpora for German and English. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, 2019.

[17] Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. Metaphor understanding challenge dataset for LLMs. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, 2024.

表 6　語の意味別の MRR

順位	意味	MRR
1	blueness	0.971
2	blackness	0.963
3	fluidity	0.961
4	whiteness	0.951
5	purple	0.908
6	yellowness	0.904
7	ocean	0.883
8	vegetable	0.881
9	mariner	0.874
10	redness	0.874

表 7　GoEmotions で評価時に使用したプロンプト

Read the TEXT and choose exactly one emotion that best captures the overall expressed emotion.
If no clear emotion is expressed, choose neutral.

Labels:
(A) anger
(B) disgust
(C) fear
(D) joy
(E) sadness
(F) surprise
(G) neutral

Output the single letter corresponding to the chosen emotion.

TEXT: {{ text }}

A　語の意味別の MRR

語の意味別に計算した MRR は表 6 のとおりである．fluidity や ocean，mariner など液体に関連しているような語も人間の連想傾向と一致しやすい．

B　モデルの学習設定

学習率は 6×10^{-7} から 9×10^{-6} の範囲で探索し，各設定について 10 エポック学習した．各学習について検証データで最も良い性能を示したエポック終了時点のモデルを保存した．最適な学習率は，検証データでの MRR が最も高いものとした．

C　GoEmotions の補足説明

評価用のプロンプトは表 7 のとおりである．

D　enISEAR の補足説明

評価用のプロンプトは表 8 のとおりである．

表 8　enISEAR で評価時に使用したプロンプト

TEXT: {{ text }}

Based on the TEXT above, choose the single emotion that you think the writer most likely felt.
Output only the corresponding letter of your selected option.

Options:
A = Anger
B = Disgust
C = Fear
D = Guilt
E = Joy
F = Sadness
G = Shame

表 9　MUNCH で評価時に使用したプロンプト

Choose the word(s) that can replace the highlighted word in the given sentence without changing the meaning of the sentence.
Output only the corresponding letter of your selected option.

Sentence: {{ original_sentence }}

Option A: {{ substitution_a }}
Option B: {{ substitution_b }}
Option C: Both Option A and Option B
Option D: Neither Option A nor Option B

E　MUNCH の補足説明

モデルの性能評価時に使用したプロンプトは表 9 のとおりである．一文目はプロンプトの種類によって変わる．元論文のプロンプトの一文目と同じもの[1)]を使用した．

1）https://github.com/xiaoyuisrain/metaphor-understanding-challenge/blob/main/tasks/prompts.md の Word judgement の 9 種類のプロンプト