大規模言語モデルの音象徴ベンチマーク

稲垣 賢斗 [1]　神藤 駿介 [2]　高道 慎之介 [1,2]

[1] 慶應義塾大学　[2] 東京大学

概要

音象徴とは語の音形と意味とが結びつく現象であり，近年では人間の言語獲得において重要な役割を果たしている可能性が示唆されている．本論文では，音象徴が言語モデルにおいて現れるかを検証するためのベンチマークを提案する．音象徴は，人間の複雑な情報処理に起因する現象であり，言語モデルにおいてこれを検証することは，そのメカニズムを解明するための手がかりとなると考えられる．検証方法としては，従来広く用いられてきた無意味語と他の要素との対応づけを，最小対比較の枠組みとして言語モデルに適用する．具体的には，英語の象徴素を対象に検証を行い，その音象徴が言語の統計的学習のみでは説明できないことを示した．

1　はじめに

大規模言語モデル（large language model; LLM）は，高い水準で自然言語を理解し生成する能力を示している．しかし，その理解・生成の過程にはブラックボックス的な側面があり，その振る舞いをどのように説明するかが重要な課題となっている [1]．こうした状況を踏まえ，LLM の言語処理が，人間と比べてどのような点が共通し，どのような点が異なるのかを体系的に調査することは極めて重要である [2]．そうして得られた知見は，LLM が人間社会にどのように関わり得るのかを考えるための基盤となる [3]．また，LLM と人間の差異の分析は，「人間の言語学習から得られる洞察は，言語モデルの改良にどのように活かせるか？」といった BabyLM [4] のような研究動向とも密接に関連する．

人間の言語処理において特徴的な要素のひとつに音象徴（sound symbolism）がある [5]．音象徴とは，語の音形がその意味と体系的に結びつく現象であり，例えばブーバ・キキ効果 [6] では，角ばった形には「kiki」，丸い形には「bouba」という名前を結びつける傾向があり，様々な言語地域でこれが実証されている．音象徴は生物学的・身体的に基づく言語非依存な要素と，言語体系や文化的習慣から現れる言語依存な要素があり，秋田らは，それらを言語発達及び言語進化の両方において現れるアイコニシティリングモデルとして説明している [7]．また，今井ら [8] は音象徴ブートストラッピング仮説（sound-symbolism bootstrapping hypothesis）を提唱し，音象徴が乳児・幼児の語彙学習を多様な方法で支えると主張した．これは発達心理学，脳科学においての多くの実証研究 [9, 10, 11] により支持されており，音象徴は人間の言語獲得において重要な役割を持つ可能性が示唆される．

本研究では，音象徴が言語モデルにおいて現れるかどうかを検証するためのベンチマークを提案する．具体的には，無意味語が文中でどのように用いられているかに基づき，音象徴的に整合する文と整合しない文からなるミニマルペアのデータセットを構築する．そして，言語モデルがいずれの文に対してより高い尤度を割り当てるかを評価する．音象徴の具体的事例は多岐に渡るが，今回はその中の一つである象徴素（phonestheme）[1)] について検証する．

本ベンチマークにより，言語モデルが音象徴をどのようにとらえるのかを明らかにし，人間との差異を検証できる．これにより，音象徴のうち，当該言語の体系に内在する要因と，言語非依存な音声やマルチモーダル処理に由来する要因とを分離して捉えるための新たな視座を提供する．また，この評価セットを，言語モデルの人間らしさを評価する指標の一つとして提案する．

2　データセットの作成手順とベンチマークとしての利用方法

2.1　設計

音象徴的文と非音象徴的文からなるミニマルペアのデータセットを構築する．音象徴的文は，特定の

1）　形態素未満の語の構成単位で，特定の意味を想起させるもの．

象徴素を含む無意味語を一つ含み，その無意味語に対して，象徴素から想起される意味（グロス）句を後続させた文である．一方，非音象徴的文は，対応する音象徴的文の無意味語の象徴素のみを別のものに置き換えた文である．こうして，無意味語の象徴素のグロスと意味句が整合する対照文（control）とそうでない処置文（treatment）のペアを最小限の差異で作成する．対象言語は英語とする．

2.2 作成方法

象徴素は先行研究 [12] で検証されている接頭・接尾位置に現れる象徴素 44 個のうち，作成過程においてデータベース内出現頻度のフィルターで省かれた*-awl* を除いた，43 個（接頭:22, 接尾:21）を用いる．複数の象徴素がまとめて一つとして扱われている場合（*scr/skr*, [V]*ng*:*ang/eng/ing/ong/ung* など）それらはデータセット内で均等に表れるようにした．文は無意味語と意味句を LLM を用いて別々に生成し組み合わせて作成した．使用した LLM は ChatGPT-4o-mini [2] である．データセットは Github で公開している [3]．

2.2.1 無意味語の生成

無意味語は，象徴素が接頭位置に現れるものと接尾位置に現れるものとを分けて，以下の手順に従って生成した．

手順 1. LLM を用いて，各象徴素を含む無意味語を生成し，そこから象徴素以外の部分（以下，基底形）を抽出・収集する．

手順 2. 全象徴素で収集した基底形をランダムに各象徴素へ再分配し結合する．

手順 3. 結合によって得られた語から既知語を排除するため，英単語データベースである wordfreq[4] において出現頻度が 0 でない語を除外する．さらに，英語としての自然さを担保するため，出現頻度上位 10,000 語に一度も出現しない 3-gram を含む語も除外する．この手順により，先行研究で報告されている象徴素のうち *skr-*,*-awl, -osk, -usk* は除外された．

手順 4. 手順 3 までで残った無意味語を対照語とする．各対照語について，象徴素部分を他の象徴素に置き換えることで処置語群を生成する．なお，置き換える象徴素には，元の象徴素と文字の重複を含まないもののみを用いる．処置語群についても，手順 3 と同様に既知語および英語として不自然な語を除外し，その結果，処置語群が空集合となった対照語は削除する．

手順 5. 残った各対照語に対応する処置語群からランダムに 1 語を選択し，対照語–処置語のペアを構成する．これを各象徴素につき 1000 個作成する．

なお，手順 1 および手順 2 において基底形を全集計して再分配するのは，無意味語生成時に特定の象徴素に適合した基底形が生成されることで，語の自然さの点において対照語が処置語よりも有利になるというバイアスを排除するためである．

2.2.2 意味句の生成

文テンプレート `[Left] [Word] [Right]` に対し，`[Word]` が指定したグロスを持つような `[Left]` と `[Right]` を LLM に生成させる．その後，`[Word]` を無意味語に置換することで，データセットを作成する．

データセットは，43 の象徴素について各 1000 の音象徴的／非音象徴的な文のミニマルペアとなった．データセットの例を表 1 に示す．

2.3 ベンチマークにおける利用

本データセットを言語モデル評価に用いる方法を述べる．

文全体の尤度比較. ミニマルペアデータセットによるベンチマークである BLiMP [13] と同様にペアをそれぞれモデルに入力し，音象徴的文／非音象徴文どちらに高い尤度を示したかを計算する．

意味句の尤度比較. 意味句のみの尤度を比較する方法を提案する．説明のために以下の文例を考える．

$$\underbrace{\text{The bright}}_{\text{Left}}\ \underbrace{\text{glenthor}}_{\text{Word}}\ \underbrace{\text{shines at night.}}_{\text{Right}}$$

この文全体に対する対数尤度は以下のようになる．

$$\begin{aligned}\text{score} &= \log P(\text{Left}, \text{Word}, \text{Right}) \qquad (1)\\ &= \underbrace{\log P(\text{Left})}_{\text{Constant}}\\ &\quad + \underbrace{\log P(\text{Word}|\text{Left})}_{\text{Likelihood of word}}\\ &\quad + \underbrace{\log P(\text{Right}|\text{Left}, \text{Word})}_{\text{Likelihood of gloss given word}} \qquad (2)\end{aligned}$$

この第 2 項には無意味語のトークン系列の出現しやすさという要素が含まれる．これは，検証対象であ

2) https://openai.com/ja-JP/index/hello-gpt-4o/
3) https://github.com/takamichi-lab/PhoMP
4) https://github.com/rspeer/wordfreq

表 1　象徴素とデータセットの例

Phonestheme	Gloss	Sample of Control / Treatment
gl-	“having to do with light or with vision; or something visually salient”	“The bright {*glenthor* / *sprenthor*} shines at night.”
sn-	“related to the nose, or breathing; or by metaphorical extension to snobbishness, inquisitiveness”	“The sharp {*snalbert* / *tralbert*} caused trouble while breathing.”
-ack	“collision creating noise or action with abrupt end”	“The loud {*flimbtack* / *flimbtump*} shattered the silence quickly.”

る無意味語とグロスの関連とは無関係であり，評価の阻害要因となりうる．

そこで，無意味語に対する意味句の対数尤度である第 3 項のみを計算する．なお，ミニマルペアを比較する際に第 1 項は定数となるため，実装上は第 2 項のみを除外すればペア間の尤度の高低は変化しない．

以上の計算法において，データセットの中で音象徴的文に対して高い尤度を示した割合をベンチマークスコアとして扱う．

3　データセットの評価

3.1　人手評価

本データセットが音象徴的文／非音象徴的文のペアとして妥当か判断するため，人手評価を実施した．評価者は Prolific[5] を通じて募集した，第一言語を英語とする 100 名である．評価では文のペアを提示し，それぞれに含まれる無意味語が文脈に適合していると感じる方を選択させた．各評価者は 43 の象徴素につき 1 回ずつ評価し，提示されるペアは 1000 個の中から試行ごとにランダムに選出した．以下，音象徴的文を正答として扱う．

評価の結果，象徴素レベルの正答率の分布は 図 1 のようになった．中央値は 0.564 となり二者択一課題におけるチャンスレベルである 0.5 を上回っている．このことは，音象徴的な文の選択が偶然以上の頻度で行われたことを示している．

この結果を先行研究 [12] において本実験と最も近い，グロスに合う無意味語を 4 者択一する実験と比較する．全象徴素のカッパ係数の平均は，先行研究の実験では 0.257 であったのに対し，本実験では 0.091 であった．本実験の方が低い主な理由としては，提示データの違いが考えられる．先行研究では無意味語とグロスを対応させるのに対して，本実験では無意味語と，グロスが想起される意味句を対応させるという間接的な方法をとっている．また，先行研究では言語学者がデータを作成している一方で，本実験では LLM が作成しているため，一部のデータの文脈がうまくグロスを表現できていない可能性がある．

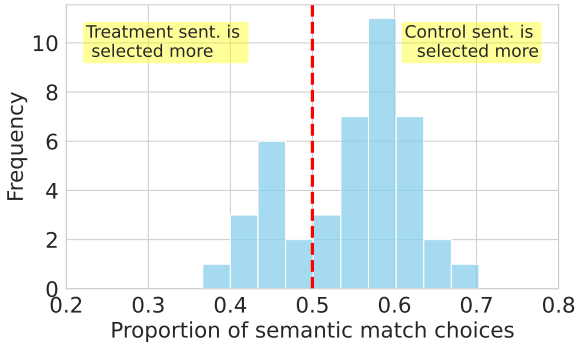

図 1　象徴素レベルの正答率の分布

先行研究と本研究で正答率の高い象徴素の傾向がどの程度一致するかを調べるため，象徴素レベルの正答率のスピアマンの順位相関係数を算出した．先行研究の実験結果と本実験結果の順位相関係数は ρ は 0.024 となり，ほぼ無相関であった．選択肢として提示する無意味語の形や，課題の提示方法などの実験条件が大きく異なっていたことによる可能性が高い．この結果から，音象徴性の大きさを定量的に評価することの難しさが示唆される．

3.2　言語モデル評価

データセットの全象徴素のうち人手評価の正答率が中央値（0.564）以上のものを言語モデルに用いる．

以下の 2 種類の言語モデルを対象とする．

学習済みモデル． 他のベンチマーク論文（例えば [14]）と同様に，一般的に入手可能な Pythia [15], Qwen [16], Llama [6] を用いた．

トークナイザの異なる新規学習モデル． 象徴素は統計的学習に因る側面が強い [17] ため，モデル学習時に象徴素を含む語がどのようにトークナイズされていたかは，得られる結果に影響を与えると考えられる．言語情報がどの粒度で学習された場合に，人間の音象徴判断に最も近い振る舞いが現れるのか

5）https://www.prolific.com/

6）https://github.com/meta-llama/llama-models

を検証する．

本研究では，先行研究 [18] に従い，GPT-2 [19] を用いて，サブワード，文字ベース，および音素ベースという異なるトークナイズ方法で学習したモデルを構築し，比較検証を行った．学習には先行研究と同様に BabyLM データセット [7] を用い，音素ベースの場合には，学習・検証データセットを G2P+ [20] によって音素化したものを使用した．すべてのモデルは 400,000 ステップ学習を行った．

紙幅の都合のため，各象徴素についての言語モデルの評価結果は付録 A に示す．

3.2.1 意味句の尤度を測る効果

Section 2.3 で示した意味句の尤度比較の効果を検証するため，サブワード学習を行った GPT2 に対して，無意味語のみをモデルに入力した場合のスコアと，文全体を入力した場合のスコアの順位相関係数を算出した．その結果 0.856 という高い値が得られ，文全体の尤度比較でのスコアは，無意味語とグロスの関連とは無関係な無意味語のトークン系列に強く支配されることが示された．

対して意味句の尤度比較でスコアを算出したところ，無意味語のみを入力した場合のスコアとの順位相関係数は 0.113 まで低下した．この結果は，本手法によって無意味語自体の尤度の影響を効果的に除去できていることを示している．

以降，意味句の尤度比較の手法を用いて結果を算出する．

3.2.2 学習済みモデルの結果

学習済み言語モデル間のスコアの順位相関係数を表 2 に示す．これらの結果から，異なる言語モデル間において評価傾向は概ね一致していることが分かる．また，平均スコアに関してもモデルのアーキテクチャや規模による大きな差が見られなかった．これらの結果から，本ベンチマークは言語を統計的に学習する過程で共通して獲得される性質を反映している可能性がある．

一方，全体として言語モデルは人間の結果とは低い相関を示し，スコアにおいて異なる部分が多くあった．これは，象徴素の音象徴がテキストの統計学習（具体的には next token prediction）のみでは表現できない，言語非依存な音声的要素やマルチモーダル的な要素を含んでいることを示唆する．

7） https://huggingface.co/datasets/vesteinn/babylm

表 2 学習済み言語モデルの平均スコアと相関
※ 0.3 以下を **赤字**，0.7 以上を **青字** で示している．

	Mean score	Human	Pythia-1B	Qwen2.5-3B	Llama3.2-3B	Llama3.1-8B
Human	0.605	1.000				
Pythia-1B	0.550	**0.145**	1.000			
Qwen2.5-3B	0.542	**0.247**	**0.923**	1.000		
Llama3.2-3B	0.551	**0.109**	**0.866**	**0.892**	1.000	
Llama3.1-8B	0.553	**−0.236**	**0.749**	**0.775**	**0.814**	1.000

表 3 異なるトークナイザの平均スコアと相関

	Mean score	Human	Subword	Character	Phoneme
Human	0.605	1.000			
Subword	0.550	**−0.101**	1.000		
Character	0.563	**0.104**	0.686	1.000	
Phoneme	0.538	**−0.206**	0.607	0.345	1.000

3.2.3 トークナイザの異なるモデルの結果

トークナイザの異なるモデル間のスコアの順位相関係数は，表 3 に示すとおり，互いにやや異なる傾向を示した．文字ベースのトークナイザでは，サブワードトークナイザよりも高い平均スコアを示した．これは，サブワードトークナイズでは，無意味語をトークナイズする際に象徴素が常に一まとまりの単位として分割されるとは限らないためであると考えられる．

4 今後の展望

本研究では，音象徴を言語モデルがどの程度把握しているかを検証するためのデータセットの作成方法と，そのベンチマークとしての利用方法を提案した．本研究で検証対象としたのは，音象徴の中でも統計的学習に因る側面が強い象徴素であるが，本手法は他の種類の音象徴に対しても適用可能であり，拡張性を有すると考えられる．音象徴は，ブーバ・キキ効果に代表されるように，複数のモダリティ間の対応関係として現れる現象である．そのため，本手法をマルチモーダル言語モデルに適用して検証を行うことで，人間と機械学習モデルの差異に関して，より意義深い知見が得られる可能性がある．

謝辞

本研究は，JST 創発的研究支援事業 JPMJFR226V，JSPS 科研費 23K24895 の支援を受けて実施した．

参考文献

[1] H. Luo, L. Specia, "From understanding to utilization: A survey on explainability for large language models," 2024. [Online]. Available: https://arxiv.org/abs/2401.12874

[2] Q. Niu et al., "Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges," 2025. [Online]. Available: https://arxiv.org/abs/2409.02387

[3] J. Grieve et al., "The sociolinguistic foundations of language modeling," 2024. [Online]. Available: https://arxiv.org/abs/2407.09241

[4] M. Y. Hu et al., "Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora," in **The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning**, M. Y. Hu et al., Eds. Miami, FL, USA: Association for Computational Linguistics, Nov. 2024, pp. 1–21.

[5] P. Perniss et al., "Iconicity as a general property of language: Evidence from spoken and signed languages," **Frontiers in Psychology**, vol. Volume 1 - 2010, 2010. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2010.00227

[6] W. Köhler, "Gestalt psychology," in **An Introduction to New Concepts in Modern Psychology (Liveright Publishing Corporation, New York)**, 1929.

[7] K. Akita, M. Imai, **The iconicity ring model for sound symbolism**, 11 2022, pp. 27–46.

[8] M. Imai, S. Kita, "The sound symbolism bootstrapping hypothesis for language acquisition and language evolution," in **Philosophical Transactions of the Royal Society B**, vol. 369: 20130298, 2014. [Online]. Available: https://cogpsy.sfc.keio.ac.jp/imailab/journalpapers/Imai_and_Kita_The_sound_symbolism_Phil_Trans.pdf

[9] M. Imai et al., "Sound symbolism facilitates early verb learning," **Cognition**, vol. 109, no. 1, pp. 54–65, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010027708001807

[10] ——, "Sound symbolism facilitates word learning in 14-month-olds," **PLOS ONE**, vol. 10, no. 2, pp. 1–17, 02 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0116494

[11] M. Asano et al., "Sound symbolism scaffolds language development in preverbal infants," **Cortex**, vol. 63, pp. 196–205, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010945214002883

[12] S. S. HUTCHINS, "The psychological reality, variability, and compositionality of english phonesthemes," in **Atlanta: Emory University dissertation.**, 1998.

[13] A. Warstadt et al., "Blimp: The benchmark of linguistic minimal pairs for english," **Transactions of the Association for Computational Linguistics**, vol. 8, pp. 377–392, 2020. [Online]. Available: https://doi.org/10.1162/tacl_a_00321

[14] K. T. Chitty-Venkata et al., "Llm-inference-bench: Inference benchmarking of large language models on ai accelerators," 2024. [Online]. Available: https://arxiv.org/abs/2411.00136

[15] S. Biderman et al., "Pythia: A suite for analyzing large language models across training and scaling," in **International Conference on Machine Learning**. PMLR, 2023, pp. 2397–2430.

[16] J. Bai et al., "Qwen technical report," **arXiv preprint arXiv:2309.16609**, 2023.

[17] B. K. Bergen, "The psychological reality of phonaesthemes," **Linguistic Society of America**, vol. 80, no. 2, pp. 290–311, 2004.

[18] Z. Goriely et al., "From babble to words: Pre-training language models on continuous streams of phonemes," in **The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning**, M. Y. Hu et al., Eds. Miami, FL, USA: Association for Computational Linguistics, Nov. 2024, pp. 37–53. [Online]. Available: https://aclanthology.org/2024.conll-babylm.4/

[19] A. Radford et al., "Language models are unsupervised multitask learners," 2019.

[20] Z. Goriely, P. Buttery, "Ipa-childes g2p+: Feature-rich resources for cross-lingual phonology and phonemic language modeling," 2025. [Online]. Available: https://arxiv.org/abs/2504.03036

A　詳細な実験結果

本文に載せきれない各象徴素ごとのベンチマーク結果を表 4 に示す.

表 4　各象徴素ごとの，学習済み LLM・異なるトークナイズの GPT2 における本ベンチマーク結果
※チャンスレベル（0.5）よりも統計的に有意な値（$p < 0.05$）を色付きで示している．（Human: $n = 100$，その他: $n = 1000$）

		Open source LLMs					GPT-2		
Phonestheme	Human	Pythia-1B	Qwen2.5-3B	Llama3.2-1B	Llama3.2-3B	Llama3.1-8B	Subword	Character	Phoneme
cl-	**0.653**	0.491	0.501	**0.542**	**0.579**	0.525	0.526	0.505	**0.424**
cr-	0.594	0.480	0.473	0.485	0.511	0.502	0.515	0.498	**0.455**
dr	0.574	**0.567**	**0.589**	**0.546**	**0.561**	**0.549**	**0.590**	**0.572**	0.520
fl-	**0.624**	**0.553**	**0.554**	**0.584**	**0.612**	**0.573**	**0.566**	**0.535**	**0.575**
gr-	0.584	**0.570**	0.508	**0.547**	**0.555**	**0.538**	**0.546**	**0.570**	**0.535**
sc-/sk-	0.574	0.488	0.493	0.470	0.490	0.519	0.500	0.503	0.491
scr-	0.574	0.493	0.528	**0.551**	0.524	**0.569**	0.499	**0.537**	0.490
sn-	0.574	0.484	**0.592**	**0.576**	**0.610**	**0.645**	**0.536**	**0.535**	**0.532**
spl-	0.574	0.478	**0.555**	**0.616**	**0.599**	**0.619**	**0.574**	**0.532**	**0.649**
str-	0.574	0.490	**0.459**	0.487	0.471	**0.541**	**0.606**	**0.572**	**0.609**
tr-	0.584	0.496	**0.464**	0.485	0.514	0.490	**0.547**	0.522	**0.611**
-ack	**0.624**	**0.678**	**0.613**	**0.694**	**0.627**	**0.670**	**0.718**	**0.773**	**0.582**
-am	**0.624**	**0.454**	**0.438**	**0.457**	**0.446**	**0.371**	**0.346**	0.517	0.498
-ap	**0.624**	**0.374**	0.478	**0.454**	**0.410**	**0.404**	**0.349**	**0.443**	**0.601**
-ash	**0.703**	**0.600**	**0.596**	**0.601**	**0.590**	**0.585**	**0.683**	**0.721**	**0.665**
-ick	**0.624**	**0.711**	**0.663**	**0.640**	**0.652**	**0.718**	**0.581**	**0.699**	**0.220**
-ip	**0.634**	**0.328**	**0.320**	**0.347**	**0.390**	**0.365**	**0.352**	0.510	0.490
-[V]ng	**0.653**	**0.554**	**0.540**	**0.586**	**0.566**	0.524	**0.453**	**0.465**	**0.320**
-oil	0.594	**0.818**	**0.813**	**0.817**	**0.804**	**0.852**	**0.697**	**0.816**	**0.707**
-olt	**0.624**	**0.671**	**0.586**	0.501	0.519	**0.537**	**0.737**	**0.789**	**0.664**
-owl	0.574	**0.748**	**0.694**	**0.622**	**0.610**	0.530	**0.671**	**0.378**	**0.602**
-ust	0.564	**0.389**	0.471	**0.452**	0.489	**0.537**	0.521	**0.409**	**0.592**
all	0.605	0.550	0.542	0.548	0.551	0.553	0.550	0.563	0.538