Do LLMs Agree with Humans on Emotional Associations to Nonsense Words?

Yui Miyakawa[1], Chihaya Matsuhira[1], Hirotaka Kato[1], Takatsugu Hirayama[2,1], Takahiro Komamizu[1], Ichiro Ide[1]

[1]Nagoya University, [2]University of Human Environments

Correspondence: miyakaway@cs.is.i.nagoya-u.ac.jp

Abstract

Understanding human perception of nonsense words is helpful to devise product and character names that match their characteristics. Previous studies have suggested the usefulness of Large Language Models (LLMs) for estimating such human perception, but they did not focus on its emotional aspects. Hence, this study aims to elucidate the relationship of emotions evoked by nonsense words between humans and LLMs. Using a representative LLM, GPT-4, we reproduce the procedure of an existing study to analyze evoked emotions of humans for nonsense words. A positive correlation of 0.40 was found between the emotion intensity scores reproduced by GPT-4 and those manually annotated by humans. Although the correlation is not very high, this demonstrates that GPT-4 may agree with humans on emotional associations to nonsense words. Considering that the previous study reported that the correlation among human annotators was about 0.68 on average and that between a regression model trained on the annotations for real words and humans was 0.17, GPT-4's agreement with humans is notably strong.

1 Introduction

Nonsense words (hereinafter called "nonwords") are words that do not exist within the vocabulary of a language. Although these words do not have any meaning, humans often associate specific impressions and feelings to their pronunciation and spelling (Sabbatino et al., 2022). A well-known example is the Bouba/Kiki effect (Köhler, 1929), in which people tend to associate pointy and round shapes with certain sounds. Understanding such human perception of nonwords brings benefits especially in commerce, as it helps to devise new product, character, and brand names that match their characteristics. Also, it can contribute to discovering how humans process words in general (Traxler and Gernsbacher, 2006). However, investigating such human perception requires experiments on humans, which is costly and labor-intensive, making it difficult to obtain large-scale data sufficient for statistical analysis.

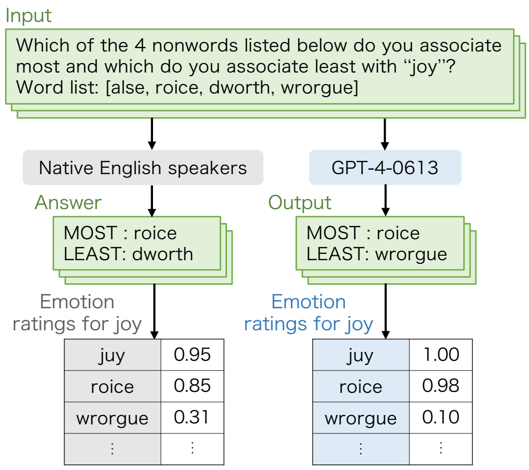


Figure 1: Nonword-emotion annotation procedures by humans and an LLM.

Previous studies have discussed whether Large Language Models (LLMs) can function as cognitive models of natural language (Mahowald et al., 2024), suggesting their usefulness in estimating the evoked impressions of nonwords in humans. Cai et al. (2024) evaluated the association between the sound and form of a nonword and the association between sound and gender in LLMs, namely ChatGPT (OpenAI, 2022) and Vicuna (Chiang et al., 2023). They suggest the usefulness of LLMs for estimating the nonword impressions in humans such as the Bouba/Kiki effect. However, they have not revealed how LLMs work for emotions which is a core component for the meaning of a language vocabulary (Mohammad, 2018).

Mohammad and Turney (2010) constructed a large, high-quality, word–emotion association lexicon to contribute to the study of emotion analysis. Based on their lexicon, Sabbatino et al. (2022) constructed an emotion intensity lexicon targeting 272

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 81–85
August 15, 2024

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.

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

nonwords. For each nonword, six emotion ratings (joy, sadness, anger, disgust, fear, and surprise) were assigned according to Ekman's basic emotions (Ekman, 1972). Crowdsourced best–worst-scaling annotations were used to collect ratings by 120 native English speakers. Sabbatino et al. (2022) constructed a regression model to estimate the emotion ratings of nonwords. In the training phase, the regressor was trained on the emotion ratings of real words, then it was tested on those of nonwords. This regression model showed Pearson's correlation coefficient of 0.17 at best. This indicates the traditional regression approach is insufficient to deal with the emotion rating prediction of nonwords.

Furthermore, since there has been no research focusing on emotion predictions through LLMs, how they can associate nonwords with emotions is elucidated. Hence, the purpose of this study is to elucidate the relationship of the emotions evoked by nonwords between an LLM and humans. Our contributions can be summarized as:

- This paper is the first to evaluate the correlation between an LLM and humans regarding the emotions evoked by nonwords.
- Following the procedure of the annotation by Sabbatino et al. (2022) as in Fig. 1, we found a positive correlation of about 0.40 between an LLM and humans.
- Evaluation demonstrates that an LLM (in particular GPT-4) agrees with humans to some extent on emotional associations to nonwords.

2 Emotion Ratings for Nonwords

To measure the correlation of nonword interpretation between an LLM and humans, this section proposes a method to reproduce Sabbatino et al. (2022)'s best–worst-scaling annotations using an LLM. Following their methodology, we focus on the six basic emotions. The annotation procedures by humans and an LLM are contrasted in Fig. 1.

2.1 Emotion Ratings by Humans

In the best–worst-scaling annotations by Sabbatino et al. (2022), first, they selected a target emotion $e \in \{joy, sadness, anger, disgust, fear, surprise\}$. Then, four words were randomly selected from the 272 nonwords and 68 real words to create a tuple. These nonwords have an orthographically correct spelling, and a monosyllabic pronunciation. The real words were used for comparison to previous studies and for attention checks of the annotators. Each word was selected eight times to create tuples, and for each tuple, three annotators answered the question: "Which of the four words do you associate MOST and which do you associate LEAST with the emotion e?" After 120 annotators had selected the word most and least associated with e, the emotion intensity score $\mathrm{score}_e(w)$ of word w was calculated as follows:

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

where $\mathrm{most}_e(w)$ and $\mathrm{least}_e(w)$ are the numbers of times w was selected as MOST and LEAST, respectively, and $\mathrm{count}_e(w)$ is the number of times w was presented. Lastly, they normalized this $\mathrm{score}_e(w)$ to $[0, 1]$. This process was performed for all six emotions.

2.2 Emotion Ratings by LLM

To reproduce this procedure using an LLM, we use the same 272 nonwords. We do not use real words because we are only interested in nonwords, and the relative order of words in an emotion based on the emotion intensity score is not affected to the correlation analysis. Using the 272 nonwords, we randomly created 1,632 tuples consisting of

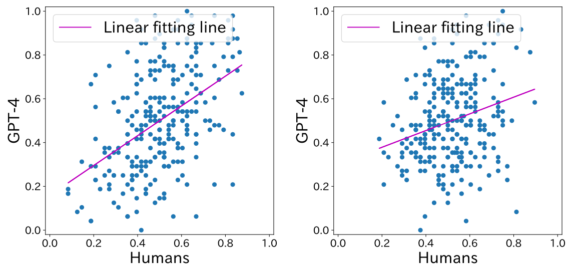
		Joy	Sadness	Anger	Disgust	Fear	Surprise	Mean
(a)	**GPT-4 & Humans**	0.44*	0.40*	0.41*	0.47*	0.44*	0.26*	0.40*
(b)	**Among humans**	0.72	0.68	0.71	0.72	0.70	0.60	0.69

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 nonword annotation (Sabbatino et al., 2022).

four nonwords. For this, we made sure that each word appeared in 24 different tuples and was not selected more than once within the same tuple to match the number of times each word was rated with the ratings by Sabbatino et al. (2022) (i.e., $\forall w; \mathrm{count}_e(w) = 24$). Then, we create an input prompt for a target emotion e and a tuple of four words.

Table 1 shows the prompt used in our evaluation. The system role indicates the role of the LLM, while the user role asks questions and instructions. In the system role, we instruct the LLM to imitate a native English speaker to make its characteristics closer to the attributes of the annotators in Sabbatino et al. (2022)'s work and to answer the question within 200 words to avoid redundant responses. In the user role, we instruct the LLM to answer the most and least relevant words to the emotion e from the four nonwords in the tuple. In addition, we instruct the LLM to think step by step to answer the question with evidence (Kojima et al., 2022), and to answer with the most relevant word followed by "MOST:" and the least relevant word followed by "LEAST:" to facilitate text processing on it.

Next, we input this prompt into the LLM. This is repeated for all tuples. The nonwords selected as most and least relevant to emotion e are then extracted from the LLM's output sentences, and the emotion intensity scores are calculated by Eqn. (1). This process is performed for all six emotions.



(a) Disgust (b) Surprise

Figure 2: Scatter plots for disgust and surprise. The x-axis represents ratings by humans, while the y-axis represents those reproduced by GPT-4.

3 Correlation Analysis: An Experiment

We conducted an experiment to investigate the correlation between humans and LLMs regarding the emotions evoked by nonwords.

3.1 Experimental Setup

`GPT-4-0613` from OpenAI API[1] was selected as the target LLM because it was one of the most predominant and powerful LLMs easily available. We set the temperature parameter for output diversity to 0 and the `top_p` parameter to 1 to ensure reproducibility.

To measure the correlation between emotional associations by the LLM and humans, Pearson's correlation coefficient was used. Note that a p-value less than 0.05 indicates a significant correlation between the LLM's and humans' ratings (Sabbatino et al., 2022).

3.2 Results

Table 2(a) shows the correlation coefficients between the LLM's and the humans' ratings with an average correlation of 0.40. The highest correlation was 0.47 for disgust and the lowest was 0.26 for surprise. Notably, p-values were less than 0.05 for all emotions. The scatter plots for disgust and surprise are displayed in Fig. 2. Both plots show an ascending trend to the right which indicates a positive correlation.

Table 3 shows an example of the GPT-4 output including the reason for its best–worst judgements of nonwords. These descriptions indicate a large influence of the meanings of real words with similar spellings on its choice. Examples of the outputs showing such an influence are shown below.

- Part of the evaluation of a tuple containing the nonword "juy" for the emotion "joy":
  `MOST: "Juy" - It's phonetically and visually similar to "joy", which makes it the most associated with joy in my opinion.`

  "Juy" was highly associated with joy because

[1] https://platform.openai.com/ (Accessed April 28, 2024)

`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.`

Table 3: GPT-4 output for the tuple of [alse, roice, dworth, wrorgue] targeting the emotion “joy”.

its pronunciation resembled “joy”.

- Part of the evaluation of “voathe” for the emotion “disgust”:
  `MOST: “Voathe” - Because it sounds similar to “loathe”, a word that signifies strong dislike or disgust.`
  “Voathe” was highly associated with disgust because it sounded similar to “loathe”, which means to intensely dislike.
- Another example for the evaluations of “roice” for joy:
  `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.`
  As seen in the part “`It also has a soft sound due to the “r” and “oi” sounds`”, GPT-4 may not have only associated the nonword with a real word, but also grasped its meaning based on sound symbolism (Hinton et al., 1995; Köhler, 1929; Sapir, 1929) related to the emotion.

3.3 Discussion

In the study of Sabbatino et al. (2022), the average correlation calculated among human annotators was 0.69 (Table 2(b)) This means that the correlation coefficients obtained using an LLM were lower than those among humans. However, the correlation when the regression model trained on real words was applied on nonwords, reported in the study by Sabbatino et al. (2022), was 0.17. Although the number of test data differs between our experiment and theirs, the large gain in correlation suggests that GPT-4 reproduces human evaluation better than the regression model.

The highest correlation for disgust may be due to its larger pool of associated real words (e.g., “loathe” for the nonword “voathe”, “filth” for the nonword “fliche”, “gross” for the nonword “groose”) compared to the other emotions.

In contrast, a possible reason for the lowest correlation for surprise could be that it has a smaller variance in human ratings (See Fig. 2(b)). Since almost no nonword has a human rating of less than 0.2 or more than 0.9, few nonwords obviously evoke or do not evoke surprise in English speakers. This may have made the annotation task difficult and the correlation low.

4 Conclusion and Future Work

Do LLMs agree with humans on emotional associations to nonsense words? —Yes, LLMs somewhat agree with humans. With the aim of elucidating the correlation between an LLM’s and humans’ understanding of nonwords, our study used GPT-4 to reproduce the emotion ratings of Sabbatino et al. (2022)’s study. We found a positive correlation of approximately 0.40 between GPT-4 and human ratings. This indicates that an LLM can be useful to estimate the emotions evoked by nonwords for humans. GPT-4 suggests that the meaning of real words with similar spellings largely influences its interpretation of nonwords, and that it may utilize knowledge of sound symbolism regarding emotion.

The existing analysis gathered from 120 persons, surely have different personae. In the future, we plan to assign more diverse personae to explore potential variations in ratings based on factors such as gender, age, and nationality. Furthermore, investigating factors affecting the LLM’s nonword interpretation will also be promising.

Limitations

Although our results show that GPT-4 can reproduce the nonword emotion ratings by humans, other LLMs, such as PaLM (Chowdhery et al., 2023) and LLaMA (Touvron et al., 2023), may behave differently to nonwords. Additionally, our experiment targeted English speakers' perception of English nonwords. If tested in different settings, e.g., another language speakers' perception, it is still an open question whether LLMs mainly trained on English data can reproduce their ratings.

Acknowledgments

Part of this work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (23K24868).

References

Zhenguang G. Cai, Xufeng Duan, David A. Haslett, Shuqi Wang, and Martin J. Pickering. 2024. Do large language models resemble humans in language use? *Computing Research Repository*, arXiv:2303.08014.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. (Accessed May 16, 2024).

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Paul Ekman. 1972. Universal and cultural differences in facial expression of emotions. *Nebraska Symposium on Motivation*, 19:207–283.

Leanne Hinton, Johanna Nichols, and John J. Ohala. 1995. *Sound Symbolism*. Cambridge University Press, Cambridge, England, UK.

W. Köhler. 1929. *Gestalt Psychology*. Liveright, New York, NY, USA.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, In press.

Saif Mohammad. 2018. Word affect intensities. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 174–183, Miyazaki, Japan.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA, USA.

OpenAI. 2022. Introducing ChatGPT. (Accessed May 16, 2024).

Valentino Sabbatino, Enrica Troiano, Antje Schweitzer, and Roman Klinger. 2022. "splink" is happy and "phrouth" is scary: Emotion intensity analysis for nonsense words. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 37–50, Dublin, Ireland.

Edward Sapir. 1929. A study in phonetic symbolism. *Journal of Experimental Psychology*, 12(3):225–239.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *Computing Research Repository*, arXiv:2302.13971.

Matthew J. Traxler and Morton A. Gernsbacher. 2006. *Handbook of Psycholinguistics*. Elsevier, Amsterdam, the Netherlands.