CMCL-17

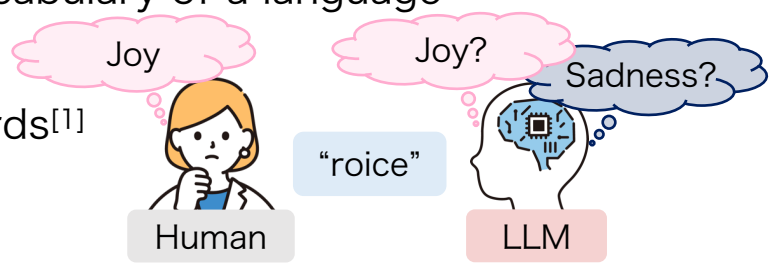
Do LLMs Agree with Humans on Emotional Associations to Nonsense Words?

Yui Miyakawa[1], Hirotaka Kato[1], Chihaya Matsuhira[1], Takatsugu Hirayama[2,1], Takahiro Komamizu[1], Ichiro Ide[1]
[1]Nagoya University, Japan, [2]University of Human Environments, Japan
Email: miyakaway@cs.is.i.nagoya-u.ac.jp

Background

- Nonsense words (nonwords) := Nonexistent words in the vocabulary of a language
- Understanding human perception of nonwords
  - Aid in devising new product and character names
- LLMs are useful for predicting human perception of nonwords[1]
  - How LLMs work for emotions is not revealed

[1] Cai et al., Do large language models resemble humans in language use?, CMCL, 2024.



Experiment: Comparing Emotion Ratings by Humans and LLMs

Human annotation[2]

- Emotion intensity lexicon of nonwords (on the right table)
  - Collected 120 human ratings for 272 English nonwords
  - Each word is assigned a score for each of 6 emotions

Nonwords list (Randomly selected 4 words)

1	[alse, roice, dworth, wrorgue]
2	[drouch, marve, durp, theight]
⋮	⋮
1,632	[chuick, marve, rheint, splink]

LLM rates the emotion intensity of nonwords

- Follow the annotation procedure by Sabbatino et al.[2]
- Instruct the LLM to imitate a native English speaker

[2] Sabbatino et al., "splink" is happy and "phrouth" is scary: Emotion intensity analysis for nonsense words, WASSA, 2022.

Question

Which of the 4 nonwords listed below do you associate most and which do you associate least with "joy"?
Word list: [alse, roice, dworth, wrorgue]

Native English speakers

Answer

MOST : roice
LEAST: dworth

Emotion ratings for joy[2]

juy	0.95
roice	0.85
wrorgue	0.31
⋮	⋮

GPT-4-0613

Output

MOST : roice
LEAST: wrorgue

Emotion ratings for joy

juy	1.00
roice	0.98
wrorgue	0.10
⋮	⋮

$$\text{Emotion ratings of nonword } w = \frac{(\#\text{ of times } w \text{ was selected as MOST}) - (\#\text{ of times } w \text{ was selected as LEAST})}{(\#\text{ of times } w \text{ was presented in the list})}$$

(Normalized to [0, 1])

Result: LLMs somewhat agree with humans

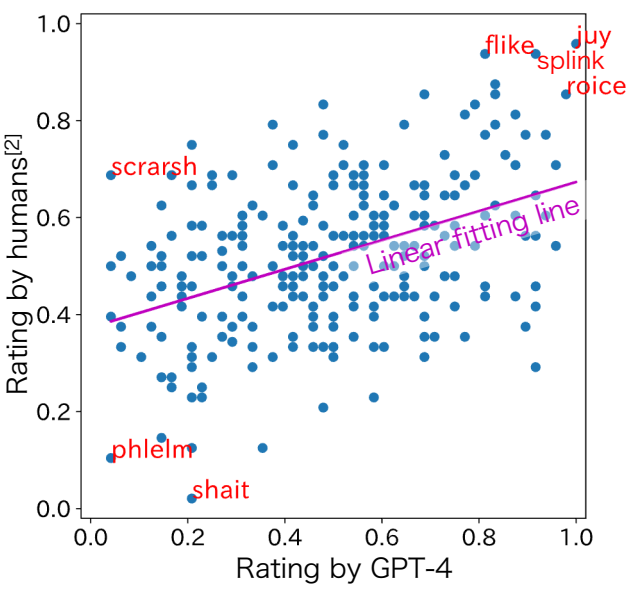
Pearson's correlation coefficients

Regression model[2] < GPT-4 < Humans[2]

0.17 | 0.40 | 0.69

Joy	Sadness	Anger	Disgust	Fear	Surprise	Mean
0.44	0.40	0.41	0.47	0.44	0.26	0.40

- Positive correlations between GPT-4 and human ratings
- Lowest correlation for Surprise
  - Surprise has a smaller variance in human ratings than the other emotions[2]
    - Few nonwords evoke obvious surprise in English speakers



Scatter plot for joy

How GPT-4 interprets nonwords?

Why did GPT-4 associate "roice" with joy?

*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.*

- Large influence of similarly spelled real words' meanings
- GPT-4 may utilize knowledge of sound symbolism regarding emotion