Published version in: https://doi.org/10.3758/s13428-024-02515-z
Open Access View-Only version: https://rdcu.be/d2ufu

Martínez, G., Molero, J.D., González, S. et al. Using large language models to estimate features of multi-word expressions: Concreteness, valence, arousal. Behav Res 57, 5 (2025). https://doi.org/10.3758/s13428-024-02515-z

Using large language models to estimate features of multi-word expressions: Concreteness, valence, arousal

Gonzalo Martínez [1] Juan Diego Molero [2] Sandra González [2] Javier Conde [2]

Marc Brysbaert [3] Pedro Reviriego [2]

[1] Universidad Carlos III de Madrid, Spain

[2] ETSI de Telecomunicación, Universidad Politécnica de Madrid, Spain

[3] Department of Experimental Psychology, Ghent University, Belgium

Accepted for publication in *Behavior Research Methods*

Keywords: word norms, concreteness, valence, arousal, multi-word expressions, large language model

Correspondence address: Marc Brysbaert
Department of Experimental Psychology
Ghent University
9000 Ghent, Belgium
marc.brysbaert@ugent.be

Abstract

This study investigates the potential of large language models (LLMs) to provide accurate estimates of concreteness, valence and arousal for multi-word expressions. Unlike previous artificial intelligence (AI) methods, LLMs can capture the nuanced meanings of multi-word expressions. We systematically evaluated GPT-4o's ability to predict concreteness, valence and arousal. In Study 1, GPT-4o showed strong correlations with human concreteness ratings (r = .8) for multi-word expressions. In Study 2, these findings were repeated for valence and arousal ratings of individual words, matching or outperforming previous AI models. Studies 3-5 extended the valence and arousal analysis to multi-word expressions and showed good validity of the LLM-generated estimates for these stimuli as well. To help researchers with stimulus selection, we provide datasets with LLM-generated norms of concreteness, valence and arousal for 126,397 English single words and 63,680 multi-word expressions.

Some words/expressions are easier to understand or produce than others. Language researchers examine the variables that affect processing ease, to build theories about the processes involved in understanding and producing language. A number of variables can be measured objectively, such as word frequency, part of speech or word length. Other variables are more subjective and are usually measured by asking people to give ratings. These variables include impressions of concreteness, age of acquisition, familiarity, or feelings associated with words/expressions.

Online testing has greatly facilitated the collection of ratings, so that it is now possible to collect ratings for thousands of words/expressions within weeks and at an affordable price (if one has access to grant money). As a result, many large-scale datasets have been published in recent years (e.g., Brysbaert et al., 2014; Diez-Alamo et al., 2019; Hinojosa et al., 2023; Proos & Aigro, 2023; Warriner et al., 2013).

At the same time, it has become clear that good estimates of ratings can be obtained with artificial intelligence (AI). At first, it did not work well because researchers tried to approximate ratings by taking average values of closely related words (Mandera et al., 2015), but soon researchers discovered that better results were obtained when they worked directly with the semantic vectors of individual words. Hollis et al. (2017; see also Westbury et al., 2015) used the Google Word2vec semantic vectors (Mikolov et al., 2013) as predictors of human ratings of concreteness, valence and arousal in a linear regression and showed that the predictions correlated 0.8 with the ratings obtained. More importantly, the predictions generalized to items not used in the initial model estimation (cross-validation), so the model could be used to estimate values for all 78,286 words with semantic vectors in the English word list they used.

The work of Hollis et al. (2017) was extended to large language models based on deep learning with multiple hidden layers between input and output (Plisiecki & Sobieszek, 2023; Solovyev et al., 2022; Wang & Xu, 2023). The approach has also been used successfully to estimate values for under-resourced languages through translation (Buechel et al., 2020; Thompson & Lupyan, 2018).

Almost all available research is limited to single words. This is a restriction as half of language utterances consist of multi-word expressions (Biber et al., 2004; Conklin & Schmitt, 2008; Muraki et al., 2023). These include compound nouns (bird watcher, blind luck), particle verbs (throw up, zone in), and fixed expressions (a drop in the ocean, good morning). There is evidence that multi-word expressions are not simply understood by parsing the words, but are stored as separate entities in the mental lexicon, just like single words. This can be concluded from the finding that multi-word expressions are processed faster than other matched word sequences, and are influenced by factors such as the frequency and age of acquisition of the multi-word expression (Arnon & Snider, 2020; Senaldi et al., 2022; Yi & Zhong, 2024). One reason why multi-word expressions are likely be stored in the mental lexicon is that their meaning often is idiomatic and cannot be derived (fully) from the meaning of the words in the expression (Sprenger et al., 2024).

Muraki et al. (2023) put multi-word expressions on the map of big-data psycholinguistic norms by collecting concreteness ratings for 62,000 English multi-word expressions (for other big data studies of multi-word expressions, see Haagsma et al, 2020; Saxena & Paul, 2020; for data on smaller sets of idiomatic expressions, see also Bonin et al., 2013; Citron et al., 2016; Costa et al., 2022; Dashtipour et al., 2022; Gavilán et al., 2021; Hubers et al., 2019; Lada et al., 2024; Nordmann & Jambazova, 2017). A logical next step would be to collect ratings for more variables, but given the evidence that large language models predict human ratings of individual words, it is worth exploring how well they estimate

multi-word expressions. A major limitation up to the introduction of large language models was that information was limited to single words (e.g., semantic vectors). Since large language models no longer work with single words as input and output, they can potentially provide useful information for word sequences, even if these sequences can take many different forms (such as wash yourself, washing oneself, washes himself, washed themselves, ...).

The present study has two goals. First, we want to see how well ratings from large language models can predict human ratings of multi-word expressions as a proof of concept. We do this by comparing how well GPT-4o estimates of concreteness approximate the human ratings collected by Muraki et al. (2023). Second, we aim to provide estimates of valence and arousal to researchers. Sentiment analysis and research into the processes involved in emotional language processing occupy a prominent place in current research (de Zubicaray & Hinojosa, 2024; Wankhade et al., 2022), and the possibility of extending this research to multi-word expressions will help make progress. The first goal is addressed in Study 1; the second goal in Studies 3-5.

Study 1: Predicting the Muraki et al. (2023) ratings with GPT-4o

We conducted pilot tests on small samples of expressions with several large language models (including models freely available for research), but we discuss only the results of GPT-4o (Open AI, 2023) because they were better than the other models we tried. We used the latest versions[1] that were available via the Application Programming Interface (API) in July-August 2024.

We also tried several types of instructions. At first, we used the instructions provided by Muraki et al. (2023), but found they were too verbose (363 words) to be repeated each time a rating for an expression was asked from the model. We also tried a version in which the model was provided with two expressions and asked to indicate which one was more concrete. In the end, we found the best results with the instructions in prompt 1 below (the three examples of stimuli at each end of the scale were taken from Brysbaert et al. (2014) and slightly improve the correlations with human ratings):

> Prompt1
>
> "Could you please rate the concreteness of the following multi-word expression on a scale from 1 to 5, where 1 means very abstract and 5 means very concrete? Examples of words that would get a rating of 1 are essentialness, although and hope. Examples of words that would get a rating of 5 are bat, frangipane, and blackbird. The expression is: [insert expression here]. Only answer a number from 1 to 5. Please limit your answer to numbers."

LLMs work by predicting the next token (a sequence of letters that may correspond to a word)[2], more precisely they estimate the probabilities (known as logprobs) of each possible token in their "dictionary" being the next and use those probabilities to select the next token. The selection process can simply take the token with the highest estimated probability or sample randomly among the tokens with highest probabilities. LLMs provide parameters such as the temperature to control this selection process. The

[1] The version we used for the master lists of norms was "gpt-4o-2024-08-06". We noticed that the estimates of this version differed slightly from the gpt-4o-2024-05-13 version used in Studies 1-3, without changing the overall quality of the estimates.
[2] An introduction to LLMs and the output they provide is also given in https://poloclub.github.io/transformer-explainer/.

temperature is a measure of the randomness added to the model to obtain a varied output on different trials; it is also summarized as the degree of creativity given to the model. We set the temperature to zero, so that the same results would be obtained in replications. The prompt was repeated for each word to avoid dilution across trials. We used python programs developed by Martínez et al. (2023) to automatize the queries via an Application Programming Interface (API) and asked for the logprobs made available by ChaptGPT-4 (Hills & Anadkat, 2023).[3] This returns the estimated probability of each response alternative, from which the dominant response could be derived and a more precise overall rating can be computed by combining the feature values with the token probabilities (Ivanova et al, 2024). For instance, the rating with the highest probability for "shoot a film" was 4 (prob = .646), followed by 3 (prob = .346), 5 (prob = .006), and 2 (prob = .001). This gave an estimated overall rating of 4*.646 + 3*.346 + 5*.006 + 2*.001 = 3.66.

Only the 62,889 expressions known by the participants of Muraki et al. (2023) were included, in order not to introduce noise due to obscure or uninterpretable entries[4]. The instructions were repeated for each expression to prevent response dilution.

	Muraki	GPT rating	GPT probs
Muraki		.798	.812
GPT rating	.793		.988
GPT probs	.807	.978	

Table 1: Correlations between the human ratings of Muraki and the GPT-4o estimates (rating with the highest probability and sum of the ratings times their probabilities). Above the diagonal: Pearson correlations; below the diagonal: Spearman correlations. (N = 62,889)

Table 1 shows the correlations between the GPT-4o estimates and the human ratings of Muraki et al. (2023). Given that the reliability of the Muraki et al. (2023) ratings is estimated at r = .84, the observed correlations of .8 comes close to the maximum value that can be expected. The sum of the ratings times the probabilities gave slightly more information than the rating with the highest probability (the same was true for all other analyses we ran). So, we will use this measure in all analyses below.

A more stringent test of whether the LLM estimates are equivalent to human ratings is by looking at the score distributions. As shown in Figure 1, they are not completely comparable. GPT-4o gives more extreme values than humans and has modes around the integers of the Likert scale. Therefore, a better way to use LLM ratings as an alternative for human ratings may be to use rank scores rather than raw scores. Indeed, the Spearman correlation between the two variables is close to the Pearson correlation (compare the lower half to the upper half of Table 1).

[3] An introduction to LLMs and the output they provide is also given in https://poloclub.github.io/transformer-explainer/.
[4] A risk of working with large databases is that some of stimuli are not known to participants or do not make sense to them. Examples of such stimuli are "1 timothy", "a cat can look at a king" or "à la provençale". Researchers interested in these specific expressions can use the instructions we used to get estimates from GPT-4o (or other LLMs).

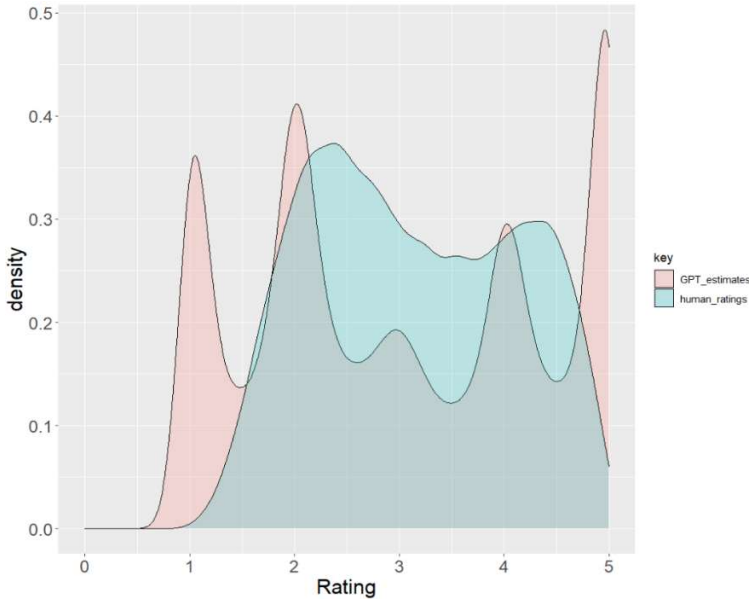

Figure 1: Distribution of the human concreteness ratings and the GPT-4o estimates for the 62,889 multi-word expressions of Muraki et al. (2023).

We would expect GPT-4o to have the most difficulties with idioms in which the meaning of the multi-word expression is much more figurative (abstract) than the meanings of the words used, such as "a golden key can open any door". The human rating of this expression was 1 (Muraki et al., 2023); the GPT-4o estimate was 2, which is closer than estimates obtained with previous AI tools.

To more systematically map the estimates of opaque idioms, we used the list of 525 most frequent idioms collected by Hsu (2020). There were 486 idioms with exactly the same wordings in our list of GPT-4o estimates[5]. These gave a correlation of r = .56, considerably lower than the .81 observed in Table 1, partly due to range restriction (most estimates were low). Interestingly, of the expressions with a difference larger than 1.75 between the human ratings and the GPT-4o estimates, there were nine in which the GPT estimate was lower than the human ratings (serve notice, for my money, think twice, stay the course, hold water, fourth estate, above board, pull no punches, sit on the fence), whereas there was only one in which GPT was higher (fat cat). So, it is not the case that GPT-4o overestimates the concreteness of opaque idioms that use vivid metaphors to express (abstract) ideas. All values for the Hsu idioms can be found on https://osf.io/k5a4x/.

Another advantage of GPT-4o assessments is that they can be extended easily to new stimuli. Some expressions not in the list of Muraki et al. (2023) are "accede to", "high air pressure", and "coughing up blood." GPT estimates for these new stimuli can be obtained without much effort (1.00, 3.12 and 4.85, respectively) and added to the list.

[5] As indicated in the introduction, an issue with multiple word expressions is that they can be summarized in different wordings. For instance, Hsu (2020) listed "a blind spot", whereas in our list we had "blind spot" (human rating = 2.7; GPT estimate = 3.17).

Study 2: Obtaining GPT-4o word estimates for valence and arousal

Now that we know that useful estimates of word features of multi-word expressions can be obtained with GPT-4o, it becomes interesting to collect other norms. Two norms often used in research are valence and arousal. Valence refers to whether an expression communicates a positive or negative emotion. In the valence ratings of Warriner et al. (2013), words like pedophile and rapist got valence ratings of 1.3, whereas words like happiness and vacation got valence ratings of 8.5 (on a Likert scale of 1-9). Arousal refers to the degree of excitement that a stimulus evokes. Words with the lowest arousal ratings in Warriner et al. (2023) were grain, calm and dull; words with the highest ratings were sex, gun and insanity.

Information on valence and arousal is needed to determine the extent to which information processing is affected by the emotional nature of the stimuli (e.g., de Zubicaray & Hinojosa, 2024; Ferré et al., 2024; Kyröläinen et al., 2021). It is also interesting for sentiment analysis of texts and messages (Birjali et al., 2021; Wankhade et al., 2022).

We used a two-step procedure to obtain valence and arousal estimates for our list of multi-word expressions. First, we built on the 13,914 single-word ratings of Warriner et al. (2013) to decide which instructions were the best. As Warriner et al. used a rating scale from 1 to 9, we used the same scale. These were the final instructions we used for valence (Prompt 2):[6]

> Prompt 2
>
> "Could you please rate how reading the following multi-word expression makes a person feel. Use a scale from 1 to 9, where 1 means very negative, bad and 9 means very positive, good. Examples of words that would get a rating of 1 are pedophile, AIDS and wreck. Examples of words that would get a rating of 9 are vacation, fantastic, and laugh. The expression is: [insert expression here]. Only answer a number from 1 to 9. Please limit your answer to numbers".

An advantage of valence and arousal estimates is that we have several sources to compare the output with. The ones we used were:

1. Warriner et al. (2013): human ratings on 1-9 Likert scale (N = 13,914).
2. Scott et al. (2019): human ratings on 1-9 Likert scale (N = 4,083 words in common with Warriner et al.).
3. Mohammad (2018): human ratings based on most–least (best–worst) rankings (N = 13,864 words in common with Warriner et al.).
4. Recchia & Louwers (2015): estimates based on similarity of ratings to ANEW words (N = 13,783 in common).
5. Hollis et al. (2017): estimates based on semantic vectors and Warriner et al. (2013) (N = 13,789 in common).
6. Buechel et al. (2020): estimates based on semantic vectors fine-tuned on Warriner et al. (2013) (N = 13,810 in common).

[6] We also tried the reverse order of ratings, going from 9 to 1. This gave slightly lower (.01-.02) correlations with the human ratings. The average of both instructions did not have higher validity either.

7. Plisiecki & Sobieszek (2024): estimates based on GPT-3 fine-tuned on Warriner et al. (2013) (N = 13,680 in common).

Table 2 shows the correlations we obtained. They are on par with the correlations obtained in other studies with large language models and better than the early estimates of Recchia & Louwers (2015) and Hollis et al. (2017). The Buechel and Plisiecki estimates did better on the Warriner et al. data on which they were fine-tuned, but not on the other two sets of human data.

	Warriner	Scott	Mohammad	Recchia	Hollis	Buechel	Plisiecki	GPT4
Warriner		.92	.86	.80	.84	.93	.97	.90
Scott	.91		.90	.80	.84	.91	.91	.93
Mohammad	.84	.88		.77	.81	.87	.86	.87
Recchia	.77	.76	.74		.78	.83	.80	.79
Hollis	.81	.82	.79	.74		.89	.84	.83
Buechel	.91	.90	.85	.80	.87		.93	.90
Plisiecki	.87	.89	.83	.78	.81	.91		.89
GPT4	.88	.92	.86	.76	.81	.89	.88	

Table 2: Valence: Correlations between the human data (Warriner, Scott, Mohammad), estimates based on semantic vectors (Recchia, Hollis), and estimates based on LLM (Buechel, Plisiecki, GPT4). Upper half: Pearson correlations; lower half: Spearman correlations.

The high correlation between AI estimates and human ratings hides the fact that the distributions of values differ considerably, as can be seen in Figure 2. Again, for some purposes it may be better to use GPT-4o ranks rather than the raw values obtained. The lower half of Table 2 shows the Spearman correlations, which are only slightly lower than the Pearson correlations.

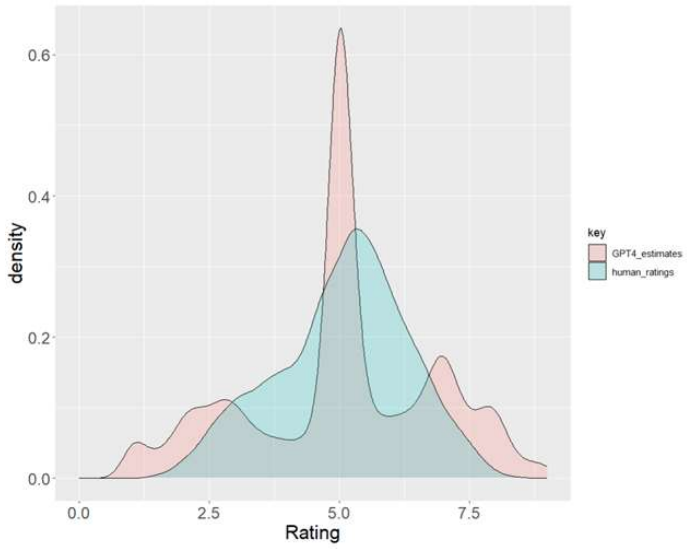

Figure 2: Distribution of the human valence ratings and the GPT-4o estimates for the 13,914 words tested in Warriner et al. (2013).

Table 3 and Figure 3 show the same data for the arousal ratings. The prompt used was:

Prompt 3

"Could you please rate how reading the following multi-word expression makes a person feel. Use a scale from 1 to 9, where 1 means very calm, relaxed and 9 means very aroused, energized. Examples of words that would get a rating of 1 are grain, dull and rest. Examples of words that would get a rating of 9 are gun, lover, and thrill. The expression is: [insert expression here]. Only answer a number from 1 to 9. Please limit your answer to numbers".

	Warriner	Scott	Mohammad	Recchia	Hollis	Buechel	Plisiecki	GPT4
Warriner		.61	.68	.62	.68	.79	.92	.74
Scott	.60		.54	.51	.56	.58	.60	.56
Mohammad	.66	.53		.68	.71	.76	.69	.81
Recchia	.59	.49	.64		.72	.76	.65	.71
Hollis	.66	.55	.68	.68		.83	.71	.78
Buechel	.78	.60	.74	.73	.81		.81	.85
Plisiecki	.92	.59	.66	.60	.68	.80		.76
GPT4	.73	.57	.79	.68	.77	.83	.74	

Table 3: Arousal: Correlations between the human data (Warriner, Scott, Mohammad), estimates based on semantic vectors (Recchia, Hollis), and estimates based on LLM (Buechel, Plisiecki, GPT4). Upper half: Pearson correlations; lower half: Spearman correlations.

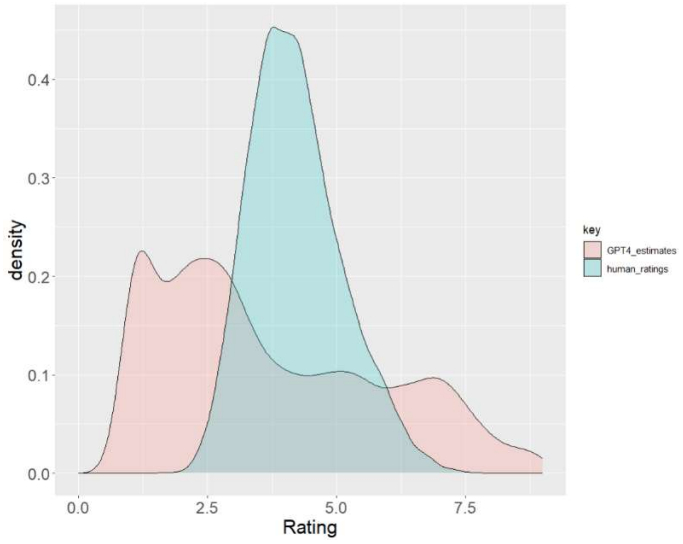

Figure 3: Distribution of the human arousal ratings and the GPT-4o estimates for the 13,914 words tested in Warriner et al. (2013).

There is less agreement about the degree of arousal evoked by words, as can be seen in the lower correlations between the three sets of human data (Warriner, Scott, Mohammad). GPT-4o does well relative to the other sources, in particular for the data of Scott and Mohammad, which were not used to fine-tune the models. The GPT estimates were more spread than the Warriner ratings and situated more at the low end of the arousal scale.

Study 3: Obtaining GPT-4o multi-word expression estimates for valence and arousal

Now that we have good instructions, we can obtain GPT estimates of valence and arousal for the multi-word expressions of Muraki et al. The instructions for valence were the same as prompt 2 above; the instructions for arousal were those of prompt 3.

The 10 expressions with lowest valence ratings were: child pornography, racial extermination, child molester, gang rape, paedophile ring, suicide bombing, ethnic cleansing, nazi party, child abuse, and white supremacy.

The 10 expressions with the highest valence ratings were: I love you, summer vacation, best friend, pure joy, Merry Christmas and a Happy New Year, on cloud nine, totally awesome, vacation time, Heaven on earth, perfect in every way.

The 10 expressions with the lowest arousal ratings were: of a, soybean oil, of an, rye seed, oat grass, such as, was to, oat bran, haricot bean, there are.

The 10 expressions with the highest arousal ratings were: gang rape, suicide bomber, racial extermination, suicide bombing, child pornography, child molester, terrorist act, terrorist attack, mass murder, suicide terrorist.

A further way to verify the quality of the estimates is to look at the relationship between valence and arousal. In human ratings there is an inverted U-shape relationship, because words with low and high valence have higher arousal ratings than words with medium valence ratings. Figure 4 shows that this is the case for the multi-word expression estimates as well. In particular, negative words have a high arousal (see Warriner et al. for a similar pattern in people).

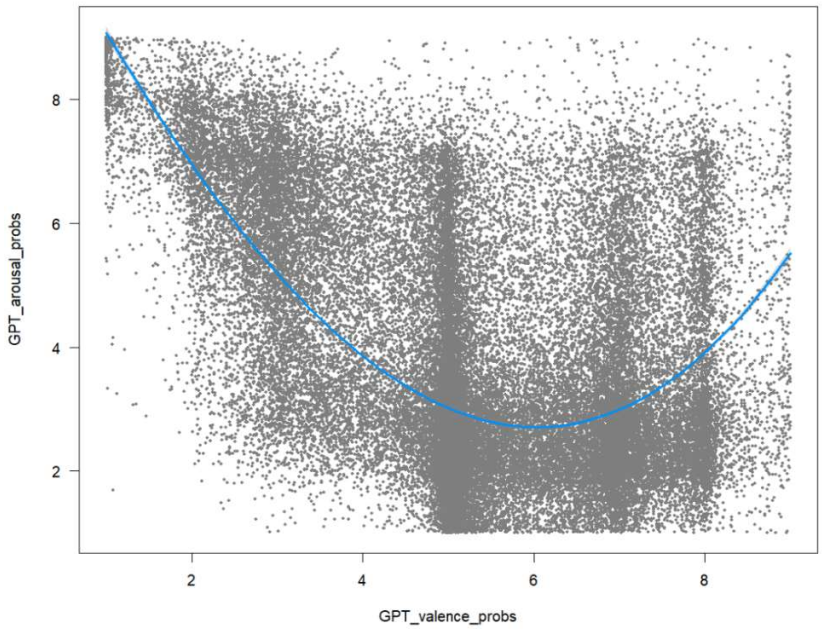

Figure 4: Correlation between the GPT arousal estimates and the valence estimates for the multi-word expressions.

At the same time, there are exceptions that help us evaluate the quality of the GPT estimates. The 18 expressions with valence estimates between 4 and 6 and arousal estimates above 8.8 were all related to sexuality (sex bomb, sexual pleasure, Latin lover, sexual relationship, sex talk, sexual relation, love affair, …). They were arousing but not necessarily in a positive or negative way, also because there is a gender difference for these words, with men rating such words higher on valence than women (Warriner et al., 2013).

The seven negative expressions with valence estimates below 1.6 and arousal estimates lower than 4 were all infrequent expressions (hare lip, coon bear) and contained outdated slurs (to be found at osf).

The seven positive expressions with valence estimates above 8.999 and arousal estimates below 4 were related to positive experiences that are relaxing rather than arousing (best friend, summer vacation, perfect in every way, vacation time, Merry Christmas and a Happy New Year, summer holiday, Heaven on earth).

Figure 5 shows the distributions of the GPT estimates obtained. They resemble the word distributions shown in Figures 2 and 3. The valence estimates are heavily centered on the mean and most expressions are at the low arousal end. Using ranks of the probability-based estimates can tease the values further apart (if that is desired).

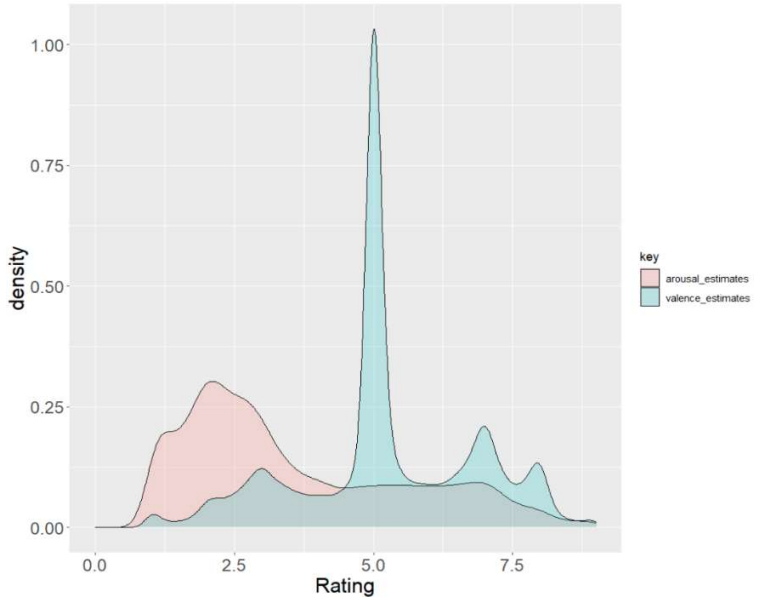

Figure 5: Distribution of the arousal and valence GPT-4o estimates for the multi-word expressions of Muraki et al. (2023).

Study 4: Validation of the GPT-4o valence estimates for multi-word expression

Although the GPT estimates of valence and arousal for multi-word expressions seem sensible and consistent with the high correlations found in Studies 1 and 2, we cannot be sure that the estimates are sound without comparing a representative sample of them with newly collected human data.[7] Such validation would also address the concern that the high correlates between LLM-estimates and human ratings found in Studies 1-2 are due to leakage and data contamination (Trott, 2024). These terms refer to the fact that all human data we used for validation were online when GPT-4o was trained and in principle could have been part of the training set of GPT-4o. If the human ratings were part of the training materials, the good performance of GPT-4o observed in Studies 1-2 may be limited to the words present in the datasets and may not generalize (well) to new stimuli for which human data are not yet available. Given that no valence and arousal ratings have been collected yet for multi-word expressions, it could be that the GPT-4 estimates of them are considerably worse than what we observed in Studies 1-2.

To test this possibility, we collected new human valence ratings for 96 multi-word expressions. The selection of the expressions was important, as we did not want it to be influenced by experimenter bias (Forster, 2000; Kuperman, 2015). Given the multitude of expressions to select from, choices by researchers may be influenced by the pattern they hope to find. Specifically, we had to be careful not to choose expressions with low GPT estimates that also looked negative to us, and expressions with high

[7] The authors thank the reviewers for pointing this out to them.

GPT estimates that looked positive to us as well. Then all we would show would be that we as humans are able to pick out good stimuli for other people.

Instead, stimulus selection had to be fully automatic and reproducible (see code at the osf site). First, we limited the multi-word expressions to those likely to be known to participants. This was done by only selecting expressions with GPT-based familiarity ratings larger than 6 on a 7-point scale (Brysbaert et al., 2024). Second, we separated the expressions into eight bins depending on their valence estimate: <2, 2-3, …, 8+. From each bin a random selection of 12 expressions was sampled. This ensured an even distribution of words across the entire range of GPT valence estimates.

The 96 expressions were presented to paid participants via Prolific (Peer et al., 2022p.). They were given instructions based on prompt 2 (see osf for the specific instructions used) and asked to rate the valence of each expression on a 7-point scale from 1 (very negative) to 7 (very positive). A 7-point scale was used instead of a 9-point scale because there is good evidence that people do not give more detailed information about words on a 9-point scale than on a 7-point scale or even a 5-point scale (Albrecht, 2024; Kusmaryono et al., 2022; Laming, 2004).

The selection criteria given to Prolific were: US resident, monolingual English, between 23 and 65 years old, more than 90% approval rate. The survey took about 10 minutes, for which participants received $3. Participants were given access in groups of 4-6 participants to control for gender ratio and reliability of ratings (without looking at correlation with the GPT estimates, which was assessed only after all data, including arousal data, were collected). Because more women than men participated, the study was restricted to men when we had ratings from 8 women, making a total of 16 participants (8 women and 8 men).

To check whether participants had to be excluded because of careless responding, we looked at the participant-rest correlations and the outcome of a factor analysis (Revelle, 2023). These indicated that no participants had to be excluded and that all participants loaded on a single factor. The correlations between participants and the rest of the group ranged from .41 to .90. Reliability of the valence ratings was omega total = .97.

The correlation between GPT estimates and mean participant ratings was r(94) = .95 (see Figure 6). The high correlation is partly due to the fact that the stimuli were evenly distributed across the entire range of GPT estimates (range maximization) but is still impressive and higher than we expected. The correlation removes the concern that GPT estimates are only good for stimuli that could have been part of the training set.

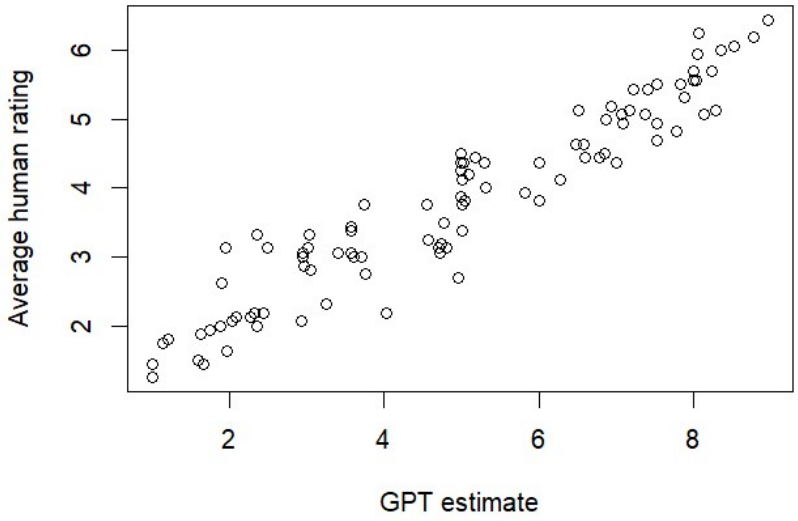

Figure 6: Correlation between the GPT-4o valence estimates and newly collected participant ratings for 96 randomly selected stimuli covering the entire distribution of GPT-4o estimates.

Another noteworthy aspect of the validation study is that there are fewer expressions with extreme ratings in the human data than in GPT-4o (see also Figures 1-3). This is probably because human ratings are an average of several people. Few people give extreme ratings and the number seems to decrease as the Likert scale has more options (this may be a reason why a 9-point scale does not give much more information in people than a 7-point scale). While there are few human ratings below 2 and above 6 in Figure 6 (and Figure 7), GPT is more radical and differentiates at the extremes. Future research will have to show whether these distinctions are useful for research.

Study 5: Validation of the GPT-4o arousal estimates for multi-word expression

The same procedures were used to select 96 items differing in GPT estimates of arousal and present them to women and men via Prolific with prompt 3. Because the participant-rest correlations were lower, more participants were tested, and the study was restricted to men after 12 women had taken part. In total, ratings were collected from 23 people (12 women, 11 men).

Participant-rest correlations and factor analysis did not force us to exclude any participant.[8] The correlations between participants and the rest of the group ranged from .28 to .81. All participants fitted

[8] We have noticed before that our attrition rate is lower than that of other authors. The only reason we can think of is that we try to make the study an enjoyable experience for participants (explain the reason for the study, pay

within a one-factor model, suggesting the absence of separate groups of raters. Reliability of the arousal ratings was omega total = .96. The correlation between GPT estimates and mean participant ratings was r(94) = .92 (see Figure 7).

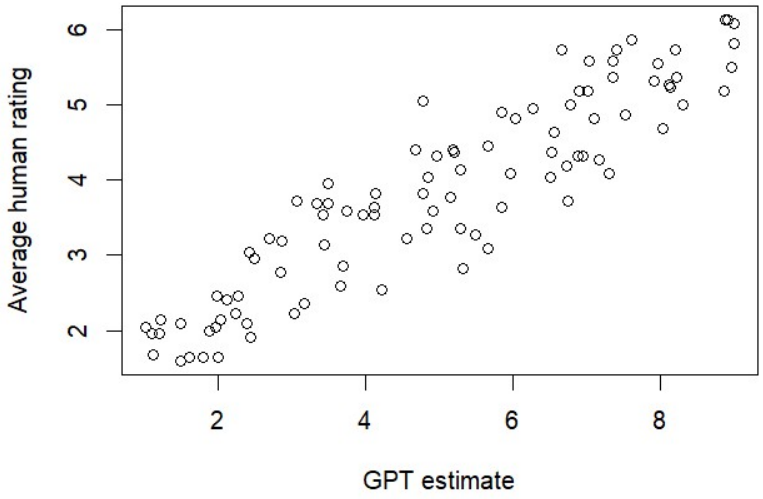

Figure 7: Correlation between the GPT-4o arousal estimates and newly collected participant ratings for 96 randomly selected stimuli covering the entire range of GPT estimates.

Discussion and availability

The purpose of this article was to see if good estimates of concreteness, valence and arousal for multi-word expressions can be obtained with large language models. Such estimates cannot easily be obtained with other artificial intelligence tools because combinations of words often have different meanings than the sum of the individual words and because multi-word expressions can take many forms given that words in them can be inflected, derived or omitted depending on the context in which the expression occurs.

We addressed the problem in a systematic way. First (in Study 1), we investigated whether it was possible to obtain concreteness estimates for multi-word expressions that correlated well with the human ratings collected by Muraki et al. (2023). We found that GPT-4o gave good estimates and that including a few examples of words at each end of the continuum in the instructions slightly improved the

participants well, give a progress bar) and provide clear information about the fact that we will manually check the quality of performance and will not pay participants if there are empirical reasons for concern. More information on good practices in online studies can be found in Rodd (2024).

results.[9] Correlation with the human ratings was r = .8. At the same time, the distribution of AI estimates differed from that of the human ratings (Figure 1). We tried out the effects of a few prompt changes and the results remained largely the same (correlations changed less than .04), suggesting that simple prompts similar to those used for human ratings will provide good results in GPT-4o.

Next (in Study 2), we obtained the same promising results when we tried to estimate the Warriner et al. (2013) word ratings of valence and arousal. The GPT-4o estimates were as good as or better than the estimates obtained with other artificial intelligence tools (Tables 2 and 3).

Finally (in Studies 3-5), we collected and validated valence and arousal estimates for multi-word expressions. Although we did not have a large-scale human dataset as a criterion, all the evaluations we tried showed that the estimates behave as expected. We see no arguments why they would be inferior to the concreteness estimates. The validation data from Studies 4-5 also mitigate concerns that GPT-4 estimates are only valid for stimuli that were part of the model's training material (Trott, 2024). They generalize to stimuli that have not yet been tested.

Showing that GPT-4o provides good estimates of concreteness, valence, and arousal for words and multi-word expression is interesting, because it allows researchers to obtain values for the stimulus materials they are interested in, also materials not covered here.[10] The estimates need not be limited to English, given that GPT-4o is available for many languages. At present, the outcome in these languages is not as good as in English (Martinez et al., in press), but this is likely to improve in the near future.

To ensure that everyone has access to the validated estimates and to reduce the environmental costs if everyone must perform the same analysis over and over again, we make our concreteness, valence, and arousal estimates available in easy-to-use Excel files. There is a file for 63,680 multi-word expressions, and a file for 126,397 words. The latter was obtained by combining the list of Brysbaert et al. (2014) with the lists of Gao et al. (2023), Scott et al. (2019), and Hollis et al. (2017). This master list includes some faulty entries (mainly from Hollis et al.), but should provide estimates for nearly all words researchers are interested in.

For each variable, the lists contain four columns with:

- The dominant estimate returned by GPT-4o (on a 5-point scale for concreteness[11] and on a 9-point scale for valence and arousal)
- The more precise estimate based on the probabilities of the ratings.
- The relative rank of the stimulus, going from 0 to 1 (i.e., the rank of the stimulus based on the probabilities estimate, divided by the total number of stimuli in the list).

[9] Researchers who disagree with these choices, are of course free to try out other prompts and compare them to the data we obtained. For instance, we kept close to the human ratings and the existing literature by collecting concreteness estimates on a 1-5 scale but valence and arousal estimates on a 1-9 scale. Authors may have reasons to prefer estimates on the same scale.

[10] We tested whether the estimates differ as a function of the order of words given to the model. We compared the multi-word concreteness ratings for a random list and a list ordered according to the Muraki et al. (2023) ratings going from abstract to concrete. The GPT estimates were the same.

[11] The GPT estimates correlate r = .89 with the word concreteness ratings of Brysbaert et al. (N = 34,246), which compares well to Hollis et al. (r = .83) and Thompson & Lupyan (r = .86).

- The rank of the stimulus, going from 1 to 100 (by rounding up the relative ranks), which may be handier for stimulus selection.

All listings are available at https://osf.io/k5a4x/. They can be freely used for research and education, but not for commercial purposes (creative commons license CC BY-NC-SA).

Declarations

- Funding: This research was supported by the FUN4DATE (PID2022-136684OB-C21/C22) and ENTRUDIT (TED2021-130118B-I00) projects funded by the Spanish Agencia Estatal de Investigacion (AEI) 10.13039/501100011033. We also profited from the OpenAI research access program, which provided access to GPT-4o on a non-commercial basis. The Prolific studies were paid by the Basisfinanciering provided to research-active members of staff at Ghent University.

- Conflicts of interest: The authors ran the studies independently and do not expect any financial gain from them.
- Ethics approval: All studies followed the General Ethical Protocol of the Faculty of Psychology and Educational Sciences at Ghent University (no harm to participants, informed consent, the right to stop at any moment, consent to distribute and use the data in anonymized form for research purposes, …). Therefore, they need no explicit approval from the Faculty.
- Consent to participate: The participants in Studies 4 and 5 were given information about the nature of the task, its estimated duration and fee, and the way in which the data would be vetted and used. They gave explicit consent to participate and agreed that the data can be used and shared for research purposes after anonymization.
- Consent for publication: All authors consent.
- Availability of data and materials: All data and materials are available at https://osf.io/k5a4x/.
- Code availability: The R code used for the analyses is available at https://osf.io/k5a4x/.
- Authors' contributions: All authors have contributed to the ideas tested in the paper (and others that did not make the end report). Running the tests was done by the authors from Madrid. The Prolific data were gathered in Ghent. Ghent is also ultimately responsible for the correctness of the statistical analyses and writing.

References

Albrecht, M. (2024). Welke Likertschaal meet accurater? Een vergelijking van een schaal met 5 en 7 antwoordalternatieven voor het schatten van woordfrequentie [Which Likert scale measures more accurately? A comparison of a scale with 5 and 7 response alternatives for estimating word frequency]. Master thesis University Ghent. Available on October 26, 2024 at https://lib.ugent.be/nl/catalog/rug01:003213290.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. Journal of Memory and Language, 62(1), 67–82.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. Applied Linguistics, 25(3), 371-405.

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems, 226, 107134.

Bonin, P., Méot, A., & Bugaiska, A. (2013). Norms and comprehension times for 305 French idiomatic expressions. Behavior Research Methods, 45, 1259-1271.

Brysbaert, M., Martínez, G., & Reviriego, P. (2024). Moving beyond word frequency based on tally counting: AI-generated familiarity estimates of words and phrases are a better index of language knowledge. Available on October 26, 2024 at https://osf.io/preprints/psyarxiv/kgevy.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. Behavior Research Methods, 46, 904-911.

Buechel, S., Rücker, S., & Hahn, U. (2020). Learning and evaluating emotion lexicons for 91 languages. arXiv preprint arXiv:2005.05672.

Citron, F. M., Cacciari, C., Kucharski, M., Beck, L., Conrad, M., & Jacobs, A. M. (2016). When emotions are expressed figuratively: Psycholinguistic and Affective Norms of 619 Idioms for German (PANIG). Behavior Research Methods, 48, 91-111.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? Applied Linguistics, 29(1), 72-89.

Costa, B. F. G., Lombardi, A. G., & González, A. A. O. (2022). Descriptive norms for 1.082 Chilean-Spanish idiomatic expressions. Revista Signos. Estudios de Lingüística, 55(110), 1057-1076.

Dashtipour, K., Gogate, M., Gelbukh, A., & Hussain, A. (2022). Extending persian sentiment lexicon with idiomatic expressions for sentiment analysis. Social Network Analysis and Mining, 12, 1-13.

de Zubicaray, G. I., & Hinojosa, J. A. (2024). Statistical Relationships Between Phonological Form, Emotional Valence and Arousal of Spanish Words. Journal of Cognition, 7(1).

Díez-Álamo, A. M., Díez, E., Wojcik, D. Z., Alonso, M. A., & Fernandez, A. (2019). Sensory experience ratings for 5,500 Spanish words. Behavior Research Methods, 51, 1205-1215.

Ferré, P., Sánchez-Carmona, A. J., Haro, J., Calvillo-Torres, R., Albert, J., & Hinojosa, J. A. (2024). How does emotional content influence visual word recognition? A meta-analysis of valence effects. Psychonomic Bulletin & Review. https://doi.org/10.3758/s13423-024-02555-8

Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. Memory & Cognition, 28(7), 1109–1115.

Gao, C., Shinkareva, S. V., & Desai, R. H. (2023). SCOPE: the South Carolina psycholinguistic metabase. Behavior Research Methods, 55(6), 2853-2884.

Gao, C., Shinkareva, S. V., & Peelen, M. V. (2022). Affective valence of words differentially affects visual and auditory word recognition. Journal of Experimental Psychology: General, 151(9), 2144-2159.

Gavilán, J. M., Haro, J., Hinojosa, J. A., Fraga, I., & Ferré, P. (2021). Psycholinguistic and affective norms for 1,252 Spanish idiomatic expressions. Plos One, 16(7), e0254484.

Haagsma, H., Bos, J., & Nissim, M. (2020). MAGPIE: A large corpus of potentially idiomatic expressions. In 12th Language Resources and Evaluation Conference: LREC 2020 (pp. 279-287). European Language Resources Association (ELRA).

Hills, J., & Anadkat, S. (2023). Using logprobs. Available on October 18, 2024 at https://cookbook.openai.com/examples/using_logprobs.

Hinojosa, J. A., Guasch, M., Montoro, P. R., Albert, J., Fraga, I., & Ferré, P. (2023). The bright side of words: Norms for 9000 Spanish words in seven discrete positive emotions. Behavior Research Methods. Advance publication at https://doi.org/10.3758/s13428-023-02229-8.

Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. Quarterly Journal of Experimental Psychology, 70(8), 1603-1619.

Hsu, W. (2020). The Most Frequent Opaque Idioms in English News. PASAA: Journal of Language Teaching and Learning in Thailand, 60, 23-59.

Hubers, F., Cucchiarini, C., Strik, H., & Dijkstra, T. (2019). Normative data of Dutch idiomatic expressions: Subjective judgments you can bank on. Frontiers in Psychology, 10, 1075.

Ivanova, A. A., Sathe, A., Lipkin, B., Fedorenko, E., & Andreas, J. (2024). Log probability scores provide a closer match to human plausibility judgments than prompt-based evaluations. In South NLP Symposium. Available at https://southnlp.github.io/southnlp2024/papers/southnlp2024-poster-47.pdf.

Kuperman, V. (2015). Virtual experiments in megastudies: A case study of language and emotion. Quarterly Journal of Experimental Psychology, 68(8), 1693-1710.

Kusmaryono, I., Wijayanti, D., & Maharani, H. R. (2022). Number of response options, reliability, validity, and potential bias in the use of the likert scale education and social science research: A literature review. International Journal of Educational Methodology, 8(4), 625-637. https://doi.org/10.12973/ijem.8.4.625

Kyröläinen, A. J., Keuleers, E., Mandera, P., Brysbaert, M., & Kuperman, V. (2021). Affect across adulthood: Evidence from English, Dutch, and Spanish. Journal of Experimental Psychology: General, 150(4), 792-812.

Lada, A., Paquier, P., Dosi, I., Manouilidou, C., Sprenger, S., & Keulen, S. (2024). Four hundred Greek idiomatic expressions: Ratings for subjective frequency, ambiguity, and decomposability. Behavior Research Methods, 1-15. Available at https://doi.org/10.3758/s13428-024-02450-z

Laming, D. (2004). Marking university examinations: some lessons from psychophysics. Psychology Learning & Teaching, 3(2), 89-96.

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables?. Quarterly Journal of Experimental Psychology, 68(8), 1623-1642.

Martínez, G., Conde, J., Merino-Gómez, E., Bermúdez-Margaretto, B., Hernández, J. A., Reviriego, P., & Brysbaert, M. (in press). The continued usefulness of vocabulary tests for evaluating large language models. Plos One.

Martínez, G., Conde, J., Reviriego, P., Merino-Gómez, E., Hernández, J. A., & Lombardi, F. (2023). How many words does GPT know? The answer is ChatWords. arXiv preprint arXiv:2309.16777.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013a. Efficient estimation of word representations in vector space. ICLR. https://arxiv.org/abs/1301.3781.

Mohammad, S. (2018, July). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers) (pp. 174-184).

Muraki, E. J., Abdalla, S., Brysbaert, M., & Pexman, P. M. (2023). Concreteness ratings for 62,000 English multi-word expressions. Behavior Research Methods, 55(5), 2522-2531.

Nordmann, E., & Jambazova, A. A. (2017). Normative data for idiomatic expressions. Behavior Research Methods, 49, 198-215.

Open AI. (2023) GPT-4. Assessed at https://openai.com/index/gpt-4/ on July 29, 2024.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. Behavior Research Methods, 54, 1643-1662.

Plisiecki, H., Sobieszek, A. (2024) Extrapolation of affective norms using transformer-based neural networks and its application to experimental stimuli selection. Behavior Research Methods 56, 4716–4731. https://doi.org/10.3758/s13428-023-02212-3

Proos, M., & Aigro, M. (2023). Concreteness ratings for 36,000 Estonian words. Behavior Research Methods. Advance publication at https://doi.org/10.3758/s13428-023-02257-4.

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. Quarterly Journal of Experimental Psychology, 68(8), 1584-1598.

Revelle, W. (2023). Psych: Procedures for psychological, psychometric, and personality research, Version 2.3.9. Retrieved from https://cran.r-project.org/web/packages/psych/index.html

Rodd, J. M. (2024). Moving experimental psychology online: How to obtain high quality data when we can't see our participants. Journal of Memory and Language, 134, 104472.

Saxena, P., & Paul, S. (2020). Epie dataset: A corpus for possible idiomatic expressions. In Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23 (pp. 87-94). Springer International Publishing.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. Behavior Research Methods, 51, 1258-1270.

Senaldi, M. S., Titone, D. A., & Johns, B. T. (2022). Determining the importance of frequency and contextual diversity in the lexical organization of multi-word expressions. Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 76, 87–98.

Solovyev, V., Islamov, M., & Bayrasheva, V. (2022, November). Dictionary with the evaluation of positivity/negativity degree of the Russian words. In International Conference on Speech and Computer (pp. 651-664). Cham: Springer International Publishing.

Sprenger, S. A., Beck, S. D., & Weber, A. (2024). What Fires Together, Wires Together: The Effect of Idiomatic Co-Occurrence on Lexical Networks. Languages, 9(3), 105. https://doi.org/10.3390/languages9030105

Thompson, B., & Lupyan, G. (2018, July). Automatic estimation of lexical concreteness in 77 languages. In The 40th annual conference of the cognitive science society (cogsci 2018) (pp. 1122-1127). Cognitive Science Society.

Trott, S. (2024). Can large language models help augment English psycholinguistic datasets? Behavior Research Methods, 56, 6082-6100.

Wang, T., & Xu, X. (2023). The good, the bad, and the ambivalent: Extrapolating affective values for 38,000+ Chinese words via a computational model. Behavior Research Methods. Advance publication at https://doi.org/10.3758/s13428-023-02274-3.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731-5780.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods, 45, 1191-1207.

Westbury, C., Keith, J., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2015). Avoid violence, rioting, and outrage; approach celebration, delight, and strength: Using large text corpora to compute valence, arousal, and the basic emotions. Quarterly Journal of Experimental Psychology, 68(8), 1599-1622.

Yi, W., & Zhong, Y. (2024). The processing advantage of multi-word sequences: A meta-analysis. Studies in Second Language Acquisition, 46(2), 427-452.