The truth is no diaper: Human and AI-generated associations to emotional words

Špela Vintar[*,†] **& Jan Jona Javoršek**[†]
[*]University of Ljubljana, Slovenia
[†] Jožef Stefan Institute, Ljubljana, Slovenia
{spela.vintar, jona.javorsek}@ijs.si

Abstract

Human word associations are a well-known method of gaining insight into the internal mental lexicon, but the responses spontaneously offered by human participants to word cues are not always predictable as they may be influenced by personal experience, emotions or individual cognitive styles. The ability to form associative links between seemingly unrelated concepts can be the driving mechanisms of creativity. We perform a comparison of the associative behaviour of humans compared to large language models. More specifically, we explore associations to emotionally loaded words and try to determine whether large language models generate associations in a similar way to humans. We find that the overlap between humans and LLMs is moderate, but also that the associations of LLMs tend to amplify the underlying emotional load of the stimulus, and that they tend to be more predictable and less creative than human ones.

Introduction

The free association task is a simple method of eliciting spontaneous connections between words from individuals. Given a cue word, such as *fish*, the participant is asked to respond with one or several words that come to mind, such as *trout*, *sea bass*, *lake*, *fishing* or *fresh*. The technique was first used in psychology by Galton (1880) and later by Freud to examine his patients' thoughts, in particular hesitations and emotional reactions to certain stimuli (Freud 1913). By mid-20th century, the association task was considered a prominent way of accessing mental representation and memory, and was used extensively in cognitive psychology and cognitive science (Deese 1959).

More recently, a number of research studies have been dedicated to comparing internal semantic representations (including word associations) to distributional semantics models and, in the past few years, large language models (Mandera, Keuleers, and Brysbaert 2017; Nematzadeh, Meylan, and Griffiths 2017; Günther, Rinaldi, and Marelli 2019). While first generation LLMs would generate a lot of erratic associations where the link between the cue and response was obscure at best (Brglez, Vintar, and Žagar 2024), state-of-the-art models seemingly excel at mimicking almost any linguistic behaviour. Associations lie at the core of many creative endeavours, and - as recent studies show - can significantly improve tasks like humour generation (Tikhonov and Shtykovskiy 2024). But challenges remain, as their performance significantly drops in non-major languages (Xuan et al. 2025), and their ability to form associations in a human-like manner remains understudied.

We focus on associations to emotionally loaded words, with two research questions in mind: 1. Do LLM-generated associations resemble human ones in terms of creativity, type and overlap, and 2. If emotionally loaded words elicit particular types of sentiments in human responses, can we observe the same distribution of positive and negative sentiments in the generated responses? Our experiments are performed for Slovenian, a morphologically rich language from the family of Slavic languages which currently has just over 2 million speakers.

Related work

Before we review other attempts to compare human and computer-generated associations, it is important to present works which categorize the processes in human associative behaviour. A comprehensive explanation is given in Clark (1970), who distinguishes amongst association eliciting experiments where participants are under major, moderate or no time pressure. If the person is allowed some time, they "react with rich images, memories, or exotic verbal associations, and these give way to idiosyncratic, often personally revealing responses." Under time pressure, associations become more superficial, predictable and closely related to the stimulus. Clark then continues to present two broad categories of responses: *paradigmatic*, where the response falls into the same syntactic category as the stimulus, and *syntagmatic* where it does not; both mechanisms being governed by implicit linguistic rules such as minimum-contrast, feature-deletion and -addition or idiom-completion. A further claim is that paradigmatic associations are much more frequent than syntagmatic ones.

Fitzpatrick (2007) challenges the belief that human associative behaviour is homogeneous, and her experiments show large variability among adult native speakers in the categories of responses. She introduces a more fine-grained categorization (which we describe in the section Approach) and shows that associative mechanisms reveal an intricate interplay between cues, responses and individual respondents' associative styles.

In recent years, several authors have explored the ability of vector space models to represent conceptual organization. Mandera, Keuleers, and Brysbaert (2017) performed a detailed evaluation of correlations between human semantic spaces and corpus-based vector representations, whereby for the former they use semantic priming, semantic relatedness judgements and word associations. They find that static neural models outperform traditional count models, and that the window size used in training plays a significant role in the performance of the models.

In an experiment by Nematzadeh, Meylan, and Griffiths (2017) human word associations were compared to nearest neighbours suggested by word2Vec and GloVe, and they show that overall correlation is low and that static word embeddings fail to capture certain critical aspects of human associations. A similar conclusion was proposed by Günther, Rinaldi, and Marelli (2019), who emphasize that "word meanings are acquired through experience". For this reason, models trained directly on introspective data generally outperform corpus-trained ones (De Deyne, Perfors, and Navarro 2016).

A more recent detailed discussion of the complexity of human associative behaviour and neural modelling is provided by Richie, Aka, and Bhatia (2024) who also train their GloVe model on English SWOW (De Deyne et al. 2018) and achieve good prediction results using a variety of asymmetric measures.

Finally, (Lavorati et al. 2024) generate the entire SWOW dataset with Mistral and present some properties of LLM-generated associations against human ones. While no detailed comparison of response categories was performed, the LLM generated almost 3 times fewer unique responses than humans, showing that the diversity and variability of human responses is much higher than for those generated by the current models.

Our own experiments extend the work cited above in that we focus on emotional words and their polarity, we compare the novelty and creativity of human vs. LLM-generated associations and compute average overlap between them.

Datasets and methods

We use the recently constructed SWOW-SL dataset as the source of human association norms. SWOW-SL[1] was compiled as the Slovenian branch of the Small World of Words project[2] (De Deyne et al. 2018), a site which currently collects word associations for 19 languages of the world including Slovenian. The SWOW-SL v1.0 consists of human responses to 1,000 cue words, totalling over 60,000 associations (20,186 unique responses) from 1396 participants. Since our focus is on emotionally loaded words, our second important resource is the Slovenian Emotion Dimension and Emotion Association Lexicon SloEmoLex 1.0 (Brglez et al. 2024), an extension of the LiLaH CroSloEng Emotion lexicon created by Ljubešić et al. (2020). The lexicon contains Valence, Dominance and Arousal scores for almost 20,000 Slovenian words, of which around 14,000 are also assigned binary variables for Positive and Negative sentiment along with a discrete model of emotion covering anger, anticipation, disgust, fear, joy, sadness, surprise and trust.

Using the score for Valence which is generally considered as a measure of emotional load we select top 20 and lowest 20 Slovenian words, whereby the words must also be included as cues in the SWOW-SL association database. These words represent samples of positively and negatively loaded parts of the vocabulary, and include:

Positive: *resnica* (truth), *odlično* (excellent), *zmagovalec* (winner), *srečen* (happy), *sposoben* (skillful), *zadovoljen* (content), *zabava* (fun), *mir* (peace), *zlat* (gold), *svoboda* (freedom), *zdrav* (healthy), *mama* (mom), *mati* (mother), *dobiček* (profit), *praznik* (holiday), *sprejet* (accepted), *prijazen* (kind), *sestra* (sister), *ženski* (female), *dogovor* (agreement)

Negative: *napadalec* (attacker), *padec* (fall), *povzročiti* (inflict), *nevarnost* (danger), *poraz* (defeat), *strel* (shot), *bolnik* (patient), *nevaren* (dangerous), *žrtev* (victim), *bolečina* (pain), *smrt* (death), *pomanjkanje* (poverty), *pritožba* (complaint), *poškodovan* (hurt), *nasprotnik* (opponent), *slabo* (bad), *nasilje* (violence), *zavrniti* (reject), *izdati* (betray), *umreti* (die)

Approach

After selecting the emotional cues, we retrieve associations from the SWOW-SL database, and generate associations using 3 contemporary LLMs:

- Llama-3.3: a text-only 70B instruction-tuned model by Meta,
- GaMS-9B: a new improved and larger model of the GaMS (Generative Model for Slovene) family, based on Google's Gemma 2 family and continually pretrained on Slovene, English and some portion of Croatian, Serbian and Bosnian corpora, developed by CJVT[3],
- Claude 3.7 Sonnet: a commercial reasoning model developed by Anthropic and an enhanced version of Sonnet 3.5, number of parameters not public.

The prompt to generate association is zero-shot: "What does the word X remind you of?", with the specific instruction to all models to provide only 3 associations to each cue. Once the models provided their responses we perform several sets of analyses: firstly, we examine the association inventories and look for unexpected, novel or creative responses; secondly, we compute overlap between all 4 versions (human vs llama/gams/claude); and thirdly, we label associations with Valence scores and examine the average emotional polarity of each setting.

Creative responses. For the purposes of this manual analysis, we loosely rely on the categories of responses as defined by Fitzpatrick (2007):

[1] `http://hdl.handle.net/11356/1980`

[2] `https://smallworldofwords.org`

[3] Centre for Language Resources and Technologies at the University of Ljubljana.

Cue [**Gloss**]	**Response** [**Gloss**]
resnica [truth]	*ni plenica* [no diaper]
▶ 'the truth is no diaper', words rhyme	
resnica [truth]	*skazica* [turns.out]
▶ 'truth turns out': rhyming allusion to a common expression *pravica – skazica* (what is just is rewarded in the end)	
odlično [excellent]	*oreščki* [nuts]
▶ refering to *Odlično* nuts and seeds brand in Slovenia	
odlično [excellent]	*lari fari* [blah blah]
srečen [happy]	*vrvica* [string]
▶ probably a reference to an old Slovenian movie *Sreča na vrvici* [Happiness on a leash] about a boy and his dog	
srečen [happy]	*lonec* [pot]
▶ probably a reference to the rhyme *srečen konec, počen lonec* [happy end – broken pot]	
zmagovalec [winner]	*the winner takes it all (abba)*
▶ reference to an Abba song	
sposoben [capable]	*sanjati* [of.dreaming]
mir [peace]	*megla* [fog]
mir [peace]	*golobica* [dove]
▶ reference to the dove as a symbol of peace	
mama [mom]	*stara mama k štrudl peče* [granny baking strudel]
mama [mom]	*rdeča* [red]
mama [mom]	*neprespana* [sleepless]
mama [mom]	*mari* [Mari]
▶ a female name	
mama [mom]	*luč* [light]
strel [shot]	*sarajevo* [Sarajevo]
▶ probably a reference to the Bosnian capital under siege	
nasilje [violence]	*črna* [black]
smrt [death]	*ja* [yes]
smrt [death]	*deževna trata* [rainy meadow]
▶ possibly personal reference	
smrt [death]	*nena ga več vidiš* [Nena can you see it yet]
▶ a personal comment (in Croatian)	
smrt [death]	*komot* [easily]
smrt [death]	*zakaj je banana rumena* [why are bananas yellow]
▶ an absurdist comment	
smrt [death]	*svet je mejhn* [the world is small]
Llama	
mir [peace]	*miran* ['Miran'/peaceful]
▶ a name, or interference with Croatian for 'peaceful'	
svoboda [freedom]	*avstralija* [Australia]
smrt [death]	*pepelka* [Cinderella]
zavrniti [reject]	*opoziv* [?]
▶ a non-existing word resembling 'opposition' and 'appeal'	
claude	
poraz [defeat]	*nezmaga* [non-victory]
▶ a made-up word	
žrtev [victim]	*trpin* [sufferer]
▶ a rare word	
GaMS	
zlat [golden]	*goldman* ['goldman']
▶ possibly interference from English	
ženski [female]	*spolov* ['of.gender']

Table 1: Examples of unexpected, novel or creative responses from human and LLM contributions

- **Meaning-based association**: *x* means the same as *y*, *x* and *y* come from the same lexical set, *x* and *y* have some other conceptual link
- **Position-based association**: *y* follows or precedes *x* directly or with words between them
- **Form-based association**: *y* is *x* plus or minus affix, *y* looks or sounds similar to *x*
- **Erratic**: *y* has no decipherable link to *x* or no response given

The fourth category is the one potentially containing the most creative responses, although other categories may also involve mechanisms which evoke unexpected responses in humans based on rhyme, memories or personal idiosyncrasies. We inspect human and generated associations looking for responses which go beyond typical conceptual or position-based links between cue and response.

Human. A manual analysis of human associations shows the great variability, individuality and creativity in the responses, where the relations between cues and responses range from wordplay, intertextual or intercultural references up to completely obscure associations where the connection evades explanation. Some examples with attempted gloss translations and explanations are listed in Table 1.

LLMs. As opposed to the richness of the human associative space, LLMs produce associations which are far less original and may only occasionally depart from the expected meaning- or position-based responses, so the number of examples in Table 1 is rather small. In addition, LLM-generated associations typically follow the part-of-speech category of the cue or remain within the pool of morphologically related words (eg. *nevaren* [dangerous]: *tvegan* [risky], *hazarden* [hazardous], *nevarnost* [danger], *varovanje* [protection]) as opposed to human responses which are truly associative in that they are in part motivated by the underlying emotion (*nevaren* [dangerous]: *moški* [man], *pes* [dog], *človek* [(hu)man]). The claim that paradigmatic associations are much more frequent than syntagmatic ones (Clark 1970) is thus even truer for LLMs.

Overlap

We compute average overlap by intersecting sets of responses for each cue in each setting (Table 2). The model corresponding most closely to human associations is the presumably largest, Claude 3.7 Sonnet, followed by Llama and GaMS respectively. As for correspondences between models, the highest match is between Claude and the Slovenian GaMS, which perhaps means that additional target language data improves performance on the association task; an assumption otherwise not proven by the Human-GaMS overlap.

	Human	Llama	GaMS
Llama	18.33	/	12.50
GaMS	14.17	12.50	/
Claude	23.33	14.17	17.50

Table 2: Overlap between human and models' associations

Sentiment analysis

In the subsequent analysis of the sentiment of associations we label each response with its Valence score (between 0 and 1, 0 signifying extremely negative, 0.50 neutral and 1 extremely positive) and the binary Positive and Negative scores. We analyse the 20 positive and 20 negative cues separately, by computing average scores per cue and each of the two groups.

	Valence	Positive	Negative
Human	0.73	0.53	0.11
Llama	0.78	0.41	0.05
Claude	0.78	0.72	0.08
GaMS	0.71	0.53	0.15

Table 3: Average sentiment scores for positive cues

	Valence	Positive	Negative
Human	0.37	0.07	0.47
Llama	0.29	0.04	0.58
Claude	0.29	0.03	0.58
GaMS	0.25	0.07	0.58

Table 4: Average sentiment scores for negative cues

Tables 3 and 4 show the average sentiment scores of responses to emotional words for positive and negetive cues. As explained above, valence scores higher than 0.50 indicate positive sentiment, hence all models generate predominantly responses with a positive sentiment to positive cues and vice versa, with Llama and Claude even surpassing humans in both extremes. The columns Positive and Negative record the average number of responses in those two categories, whereby these values are binary in SloEmoLex (a word can be either positive or not), and not all words have this value assigned. Results are most conclusive for negative cues, where it seems that all models generate even more negative associations than humans. (See Figure 1.)

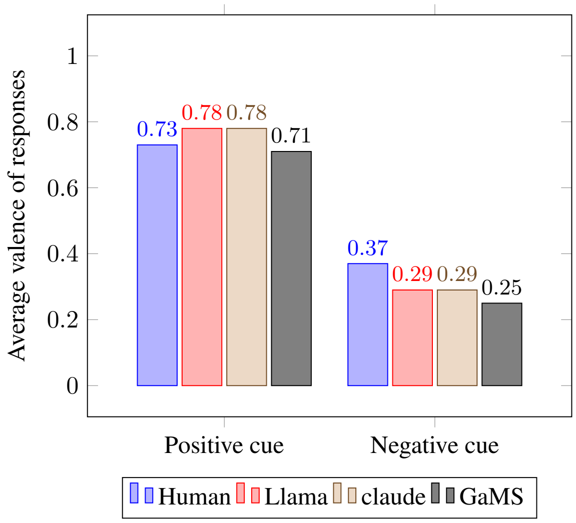

Figure 1: Valence for positive and negative cues

Discussion and conclusions

The experiments described show that LLMs are able to generate association-like responses to a zero-shot prompt, even in a small language like Slovene. A fair portion of these responses overlapped with human association norms, with the largest model (Claude 3.7 Sonnet) achieving the highest overlap. It should be noted here that seemingly small numbers do not necessarily mean poor results, as human-human overlap typically does not exceed 40 %.

In line with our own previous work (Brglez, Vintar, and Žagar 2024), LLM-generated associations generally follow the paradigmatic pattern more often than the syntagmatic one, frequently listing synonyms or near synonyms, and rarely departing from the part-of-speech of the cue. While the present study did not focus specifically on the comparison of response categories, a manual inspection of the results indicates slight variations between models in their typical behaviour. It would appear that LLMs come in different flavours, or exhibit distinct and consistent personal traits which affect - among other things - their linguistic behaviour. This should not surprise us as recent studies within the newly emerging field of machine psychology indeed explore personality traits of LLMs (Tommaso et al. 2024; He and Liu 2025).

When human and LLM responses are compared in terms of creativity or unexpectedness, humans produce a much richer and more varied mix of responses. We are aware of the fact that the creativity of LLMs could be activated using a different prompt, possibly including a variety of human responses or explicitly requesting creativity. We deliberately refrained from prompt engineering to retain the similarity between the association task for humans and LLMs.

The sentiment analysis tentatively shows that the models tend to follow the sentiment of the cue and only rarely depart from it in a different direction, rather the responses reinforce and often amplify the original sentiment. This finding is in line with the observation of Lavorati et al. (2024) about LLM bias in generated associations: if human responses to *man* and *woman* contained traces of gender stereotypes (*woman: sex, beauty; man: strong*, Mistral's responses were overtly stereotypical (*woman: makeup, hair, fashion, beauty; man: job, career, computer, work*). One reason for the exaggeration of the original sentiment might be that the models almost never generate the antonym as response, while this is a strong association mechanism in humans. An antonym typically has the reverse sentiment than the cue, so this might explain why the average Valence scores for human responses lie more towards the middle of the range as compared to LLMs.

Our findings contradict the observations of Guo et al. (2023) who claim that ChatGPT's responses exhibit less bias and more objectivity than human answers to a collection of questions. This may be due to the fact that question answering, the task that instruction-tuned LLMs have been specifically trained for, triggers different behaviour than the association task where the model is prompted to explore its cognitive and experiential space (*"What does the word X remind you of?"*).

Even if the (lack of) diversity and variability in LLM-generated responses may say little about their (lack of) creativity, we believe that the tendency to produce outputs which follow the grammatical category, sentiment and semantic domain of the cue captures but a fraction of human associative behaviour, and that departures from the expected – which used to be quite common in 1st generation LLMs – now seem to be distinctive traits of human associations.

In future experiments it would make sense to perform a more detailed and fine-grained categorization of responses in order to see how models differ in their association patterns, a phenomenon we have observed but not yet evaluated. Also, the sample of 40 cues is too small to allow for any generalization of the above observations.

Author Contributions

Špela Vintar was in charge of planning the study, conducted parts of the analysis and wrote a significant part of the manuscript. Jan Jona Javoršek participated in the analysis, evaluation and interpretation of the results, and contributed to all versions of the manuscript.

Acknowledgments

This contribution was partly funded by ARIS grants P6-0215 Slovene Language - Basic, Contrastive, and Applied Studies, GC-0002 Large Language Models for Digital Humanities and I0-0005 The JSI infrastructure program.

References

[2024] Brglez, M.; Caporusso, J.; Hoogland, D.; Koloski, B.; Pollak, S.; and Purver, M. 2024. Slovenian Emotion Dimension and Emotion Association Lexicon SloEmoLex 1.0. http://hdl.handle.net/11356/1875.

[2024] Brglez, M.; Vintar, Š.; and Žagar, A. 2024. How Human-Like Are Word Associations in Generative Models? An Experiment in Slovene. In Zock, M.; Chersoni, E.; Hsu, Y.-Y.; and de Deyne, S., eds., *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024*, 42–48. Torino, Italia: ELRA and ICCL.

[1970] Clark, H. H. 1970. Word associations and linguistic theory. *New horizons in linguistics* 1:271–286.

[2018] De Deyne, S.; Navarro, D.; Perfors, A.; Brysbaert, M.; and Storms, G. 2018. The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods* 51.

[2016] De Deyne, S.; Perfors, A.; and Navarro, D. J. 2016. Predicting human similarity judgments with distributional models: The value of word associations. In Matsumoto, Y., and Prasad, R., eds., *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1861–1870. Osaka, Japan: COLING.

[1959] Deese, J. 1959. Influence of Inter-Item Associative Strength upon Immediate Free Recall. *Psychological Reports* 5(3):305–312.

[2007] Fitzpatrick, T. 2007. Word association patterns: Unpacking the assumptions. *International Journal of Applied Linguistics* 17(3):319–331.

[1913] Freud, S. 1913. On beginning the treatment (further recommendations on the technique of psycho-analysis I). *Standard edition* 12:121–144.

[1880] Galton, F. 1880. I.—Statistics of Mental Imagery. *Mind* os-V(19):301–318.

[2019] Günther, F.; Rinaldi, L.; and Marelli, M. 2019. Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science* 14:1006–1033.

[2023] Guo, B.; Zhang, X.; Wang, Z.; Jiang, M.; Nie, J.; Ding, Y.; Yue, J.; and Wu, Y. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. arXiv:2301.07597.

[2025] He, J., and Liu, J. 2025. Investigating the Impact of LLM Personality on Cognitive Bias Manifestation in Automated Decision-Making Tasks. arXiv:2502.14219.

[2024] Lavorati, C.; Abramski, K.; Rossetti, G.; and Stella, M. 2024. LLM-Generated Word Association Norms. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*. IOS Press. 3–12.

[2020] Ljubešić, N.; Markov, I.; Fišer, D.; and Daelemans, W. 2020. The LiLaH Emotion Lexicon of Croatian, Dutch and Slovene. In Nissim, M.; Patti, V.; Plank, B.; and Durmus, E., eds., *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, 153–157. Barcelona, Spain (Online): Association for Computational Linguistics.

[2017] Mandera, P.; Keuleers, E.; and Brysbaert, M. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language* 92:57–78.

[2017] Nematzadeh, A.; Meylan, S. C.; and Griffiths, T. L. 2017. Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 39.

[2024] Richie, R.; Aka, A.; and Bhatia, S. 2024. Free Association in a Neural Network. *Online First Publication*. doi:10.1037/rev0000396.

[2024] Tikhonov, A., and Shtykovskiy, P. 2024. Humor Mechanics: Advancing Humor Generation with Multistep Reasoning. In *Proceedings of the International Conference on Computational Creativity 2024*.

[2024] Tommaso, T.; Hegazy, M.; Lemay, D.; Abukalam, M.; Rish, I.; and Dumas, G. 2024. LLMs and Personalities: Inconsistencies Across Scales. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

[2025] Xuan, W.; Yang, R.; Qi, H.; Zeng, Q.; Xiao, Y.; Xing, Y.; Wang, J.; Li, H.; Li, X.; Yu, K.; Liu, N.; Chen, Q.; Teodoro, D.; Marrese-Taylor, E.; Lu, S.; Iwasawa, Y.; Matsuo, Y.; and Li, I. 2025. MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation. arXiv:2503.10497.