# Can Small Large Language Models Estimate the Hedonic Valence of Words?

Paulo C. Vaz de Barros
*Engenharia de Computação e Informação*
*Universidade Federal do Rio de Janeiro*
Rio de Janeiro, Brazil
paulcvaz.20231@poli.ufrj.br

Gabriel C. Motta Ribeiro
*Programa de Engenharia Biomédica/COPPE*
*Universidade Federal do Rio de Janeiro*
Rio de Janeiro, Brazil
gabrielcasulari@peb.ufrj.br

Frederico C. Jandre
*Programa de Engenharia Biomédica/COPPE*
*Universidade Federal do Rio de Janeiro*
Rio de Janeiro, Brazil
jandre@peb.ufrj.br

***Abstract*—In recent years, small Large Language Models (sLLMs) have increasingly been recognized for their utility due to their cost-effective performance in generating high-quality responses without relying on cloud APIs that may pose privacy concerns. This study conducts a comparative ablation analysis to evaluate the hedonic valence rating capabilities of three distinct models: Llama3.2, Phi-4, and nomic-embed-text-v1.5. The investigation involves assigning valence ratings on a 9-point scale to 140 words sampled from an extensive human-rated dataset, a subset of which was used in a previous study. The chatbot models, Llama3.2 and Phi-4, were employed via ollama using prompts specifically engineered to solicit emojis and numerical valence ratings. Nomic embeddings were used in a linear regression between the word embeddings and the human ratings. Statistical analysis revealed significant correlations between the models' outputs and human ratings (p-value $\leq 0.001$). Despite limitations, these results underscore the potential of sLLMs and embedding models in enhancing sentiment analysis. This study shows how sLLMs can effectively approximate word valence and may help in linguistic research.**

***Index Terms*—Small Large Language Models, Hedonic Valence Rating, Sentiment Analysis, Word Embeddings, Linear Regression, Chatbot Models, Model Evaluation and Performance, Natural Language Processing, Llama3.2, Phi-4, nomic-embed-text-v1.5, Machine Learning in Linguistics**

## I. Introduction

In recent years, Large Language Models (LLMs) have undergone significant improvements in computational efficiency. Models with just a few billion parameters are capable, in some tasks, of providing responses with a quality close to that of much larger models [1]. Furthermore, previous studies have shown that language models can be used as simulacra of human evaluations in text [2], even responding to prompt modifications that include a modifier indicating the group to which the simulated human belongs [3]. Studies such as [4] have shown the ability of GPT-4 to recognize and rate emotions from images, while [5] has shown how GPT-4 and GPT-3.5 perform in rating emotions in text containing emojis.

In these papers, the capacity of LLMs in sentiment analysis is investigated. A notable limitation, however, is their exclusive reliance on models accessed through cloud APIs. Local evaluation offers a compelling alternative for users who wish to avoid third-party APIs, perhaps due to privacy considerations or the desire to maintain sensitive data securely within a local environment. Small LLMs (sLLMs) present a viable solution as they are capable of local execution without extensive computational infrastructure such as multiple GPUs and substantial RAM.

Embedding models are specialized tools that translate the human-understandable data (text, images, audio) into high-dimensional numerical tensors, which LLMs can process. Unlike simple numerical encodings such as word counting or one-hot encoding, embeddings are dense and rich in information. This density empowers these tensors to encode deep semantic meaning, allowing words representing similar ideas or concepts to be positioned numerically "close" to one another in the embeddings space [6].

In this study, we investigate the ability of sLLMs and embedding models that use transformers architecture [7] in sentiment analysis, specifically in the classification of words according to their affective content. The evaluation uses the valence domain of the Self-Assessment Manikin (SAM) scale. The SAM scale, developed by Margaret Bradley and Peter Lang, is a non-verbal pictorial questionnaire designed to directly measure a person's feelings in response to a stimulus [8]. It simplifies affect measurement across three domains: valence, arousal, and dominance.

The valence represents a continuous scale of pleasantness or unpleasantness. Using this specific domain, we can effectively assess how well sLLMs and embedding models classify the degree of positive or negative emotional content in words.

This paper seeks to answer the following questions: Given that sLLMs inherit capabilities from larger ones, can they simulate human evaluation of valence of words effectively? Are embedding models capable of encoding concepts like "positive" and "negative"? If so, can their embeddings be used to classify the hedonic valence of words?

The objective is to compare the performance of three language models, two sLLMs and one embedding model, in acting as human simulacra for the evaluation of valences of words.

This study was partially funded by the Brazilian agencies CNPq, CAPES.

## II. METHODS

### A. Models

This study employed two conversational sLLMs and one embedding model: Llama3.2[1] (Meta, USA) with 3 billion parameters, Phi-4[2] (Microsoft Research, USA) with 14 billion parameters, and nomic-embed-text-v1.5[3] [9] (Nomic AI, USA) with 137 million parameters. All sLLMs were executed locally using the Ollama framework and imported through the terminal using the following commands: `ollama pull llama3.2` and `ollama pull phi4`. The embedding model (Nomic) was used via Transformers Library available in Hugging Face.

These three models were selected due to their top ranking in usage (among the highest number of pulls) on both Ollama and Hugging Face. Furthermore, their development by well-established big tech, known for their consistent production of high-quality products, further justified their inclusion.

### B. Prompt

The base prompt was written according to that used in a previous study [2], and provided to the chatbot models (Llama3.2 and Phi-4). The returned emojis were not considered in the analysis. The base prompt is as follows:

*How negative or positive is this word on a 1-9 scale? Answer only with a facial emoji and a number, with 1 being 'very negative' and 9 'very positive'. Here is the word: < word >*

### C. Datasets

The dataset analyzed in this study is publicly available online and was developed from ratings collected via the Amazon Mechanical Turk crowdsourcing [10]. Participants rated each word on a 9-point scale across three distinct emotional dimensions, although here only the hedonic valence was used. In the present study, two subsets of this dataset (totaling 13,915 words) were used: a regression set consisting of 2,782 words, specifically employed for estimating the coefficients of the regression model used with the embedding model; and a final set of 140 words, distinct from those used in regression, also drawn from the dataset and the same used in the reference study [2]. The dataset was ordered by the valence column and the regression set was generated by traversing this ordered dataset in steps of five words and randomly selecting this, or the directly preceding, or succeeding word. Such procedure aimed to preserve the distribution of the original dataset. The computed valence scores of the words from the final set were then compared to human-rated values.

This specific sampling strategy for the regression set (intervals of 5 words) was determined through an iterative empirical process aimed at balancing accuracy and avoiding overfitting. Preliminary tests with larger intervals sometimes led to a less representative training set, potentially increasing the risk of the regression model overfitting, while smaller intervals provided insufficient data diversity, potentially leading to underfitting.

[1]https://ollama.com/library/llama3.2:3b

[2]https://ollama.com/library/phi4:14b

[3]https://huggingface.co/nomic-ai/nomic-embed-text-v1.5

### D. Experimental Procedures

All data were processed using Python notebooks, on a system equipped with an Intel i9 processor, 128 GB RAM, and an NVIDIA RTX 4090 GPU with 24 GB VRAM. The responses from each model were saved in files that includes the word itself, the value returned by the model, and the average value assigned by a group of human raters.

To determine the hedonic valence of words using Nomic, this experiment employed ordinary least squares (OLS) linear regression between the embedding tensors (independent variables), extracted using Nomic, and the human ratings from the 2,782-word regression set (dependent variable). The resulting vector of coefficients is then applied to the embeddings of the 140 words in the final set in order to estimate their valence.

The process of adjusting the chatbot options also involved modifying some of the models' input parameters to increase the experiment's reproducibility. Tab. I shows the parameters used for the requests with the final set of words.

TABLE I: Parameters used for ollama.chat() request.

| Parameter | Setting |
|---|---|
| temperature | 0 |
| top_p | 1 |
| top_k | 1 |
| repeat_penalty | 1 |
| seed | 42 |

The configuration parameters for the sLLMs, as presented in Tab. I, resulted in a consistent numerical outputs for identical prompts, with no variability observed with given seed due to the temperature setting being fixed at zero. A Top_p value of 1 ensures that all tokens are considered when selecting the next word; given that the temperature is set to zero, this parameter likely had no practical impact on the output. Top_k set to 1 restricts the model to choosing only the single token with the highest probability, leading to the most deterministic outcome possible.

These parameters should enable the replication of these results to be achieved across different environments, provided the same dataset, subset and exact model versions are utilized.

### E. Statistical analysis

Violin plots were generated for the valence of the whole dataset, as well as for the 140-word sample for each model.

In the data analysis we considered the sLLMs and the linear regression of the embedding model as instruments to measure the valence of words, comparing their measurements to a gold-standard instrument, i.e. the volunteers from the original study [10]. Considering only the final set, the words were evaluated by the three models. The models' valences were compared to the human valences to assess how well the machine-generated ratings represented such reference ratings, by fitting a linear regression, similar to the analysis in [2].

The choice of Pearson correlation and linear regression to evaluate model performance is driven by the numerical nature

of the reference data. Although the rating scale is ordinal, the final valence value assigned to a word is an average of multiple ratings, which gives it an interval nature.

The regression parameters, p-values and Pearson correlation coefficients were calculated with the *linregress* function from the Scipy library. P-values below 0.05 were considered significant. The mean absolute error of the outputs was calculated for each model.

To compare the models' outputs, a repeated-measures ANOVA (ANOVA-RM) was conducted.

## III. Results

The data presented in this study were generated on May 19, 2025, using the models in their latest version in this day. A sample of 5 words and their ratings, given by each model, can be visualized in Tab. II.

TABLE II: Example of hedonic valence assigned to five words by different language models and average of a group of humans (H.V.).

| Word | H. V | Phi-4 | Llama3.2 | Nomic | ChatGPT |
|---|---|---|---|---|---|
| attacker | 2.23 | 2 | 7 | 2.89 | 4.0 |
| adversity | 4.26 | 3 | 6 | 4.25 | 5.0 |
| fraction | 5.21 | 5 | 5 | 5.13 | 5.0 |
| ask | 5.95 | 6 | 5 | 6.20 | 5.0 |
| relaxing | 8.19 | 8 | 8 | 6.24 | 9.0 |

For the valence scores generated by the sLLMs' output, both models produced only integer ratings from 1 to 9, as specified in prompt (Sec. II-B). In contrast, the embedding model yielded non-integer values within the range of 1.78 to 7.31.

For each model, the mean absolute error (MAE) and the p-value derived from a linear regression between the machine-generated scores and the human-rated values were computed (Tab. III). The results were as follows: Phi-4 (MAE $= 0.98$, $p < 0.001$), Llama3.2 (MAE $= 1.40$, $p < 0.001$), and Nomic (MAE $= 0.73$, $p < 0.001$). The Pearson correlation coefficients further indicate that Phi-4 outputs exhibit the strongest correlation with human ratings ($R = 0.81$), followed by the outputs from the embedding model ($R = 0.71$), despite having fewer parameters than the other models. Outputs from Llama3.2 showed the weakest correlation ($R = 0.39$).

A slope (a) closer to 1 and an intercept (b) closer to 0 generally indicate better model performance (Tab. III).

TABLE III: Results of linear regression between the machine-generated and human valence ratings.

| Model | Intercept | Slope | R-value | MAE |
|---|---|---|---|---|
| Phi-4 | -0.71 | 1.20 | 0.81 | 0.98 |
| Llama3.2 | 3.87 | 0.45 | 0.39 | 1.40 |
| Nomic | 2.34 | 0.54 | 0.71 | 0.73 |
| ChatGPT | 0.55 | 1.01 | 0.72 | 0.58 |

The R-value indicates the proportion of the variance in the dependent variable (hedonic valences given by the models) that is predictable from the independent variable (Reference Value by [10]). A higher R-value (closer to 1) suggests a better fit of the model to the data. R-value is defined as

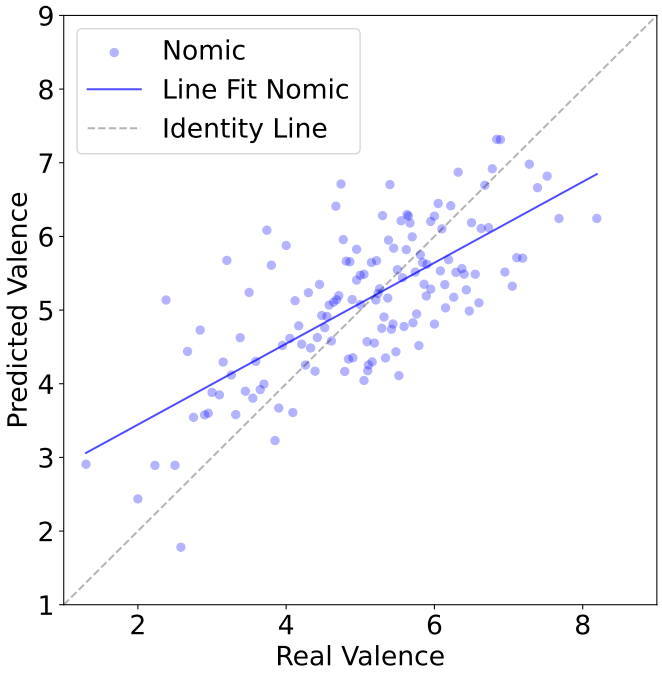


Fig. 1: Comparison between values generated through the embedding model to the final set and the reference value for this set.

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

The valence values generated by the embedding model (Fig. 1), Phi-4 (Fig. 2) and ChatGPT (Fig. 4) showed a large correlation (R-value $> 0.7$) and align well with the target values. Llama3.2 in contrast, scored a lower R-value = 0.39 suggesting a weak linear relationship probably caused by the higher occurrence of words with predicted valence equal to 5 (Fig. 3).

## IV. Discussion

This small ablation study showed that:

1) In sentiment analysis tasks utilizing sLLMs, a greater number of parameters does not necessarily equate to superior performance (compare the Pearson correlation between Llama3.2 with 3 billion Parameters and Nomic with 137 Million Parameters);
2) sLLMs can indeed evaluate the hedonic valence of words, perhaps, for each case there is a better suited model for the task;
3) Phi-4 surpasses ChatGPT (LLM used with openAI API, from the reference study [2]) in correlation in the task proposed, see Tab. III.

As observed in the violin plot in Fig. 5, in the universe dataset there's a concentration of data points in the mid-range of valence scores (around 5), with fewer data points at the extremes ($x > 7$ and $x < 3$). This non-uniform distribution has a

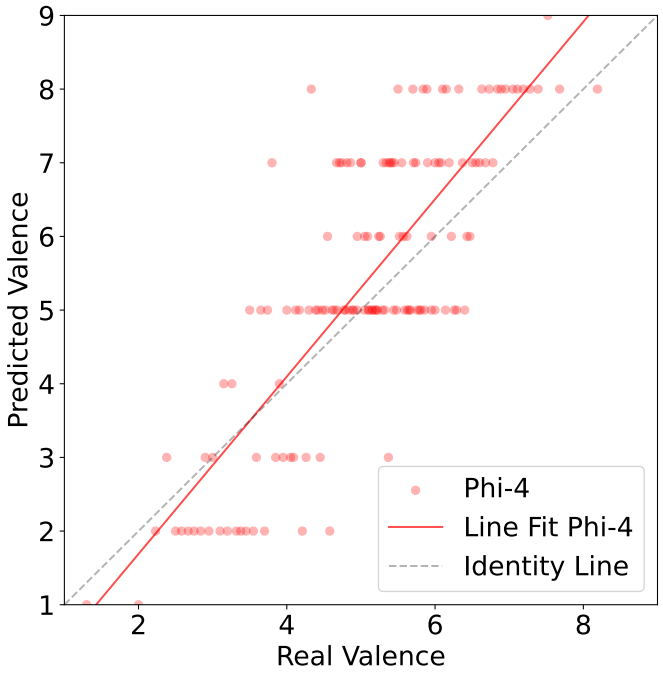


Fig. 2: Comparison between values generated by Phi-4 to the final set and the reference value for this set.

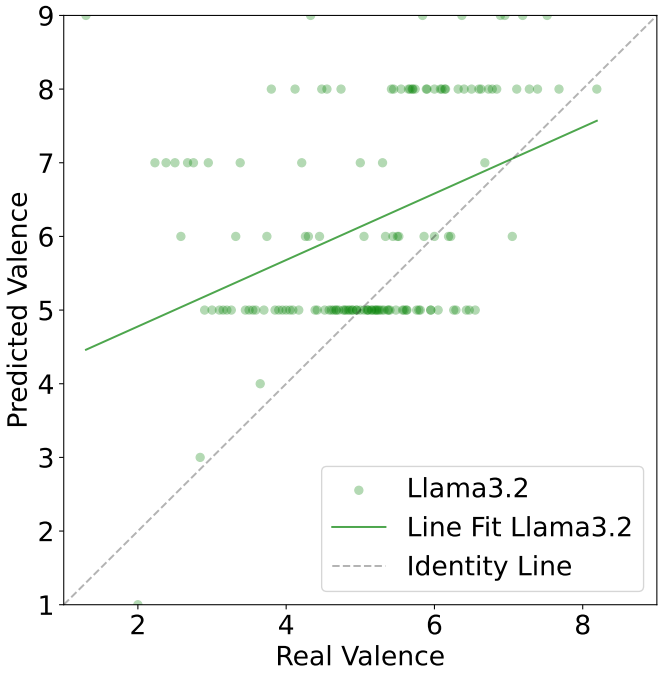


Fig. 3: Comparison between values generated by Llama3.2 to the final set and the reference value for this set.

visible effect, bringing the valence predictions of models using a sample of the dataset towards the middle, a phenomenon visually represented in the dispersion graph of each model (Fig 1, 2, 3).

The embedding model may be affected by the nonuniformity of data, since the training set inherently represents a subset of the universe. Data is denser in the middle of the histogram, potentially leading to reduced accuracy at the extremes (Fig. 1, 6).

Employing this embedding model for the task of evaluating the hedonic valence of words presents a dual nature of benefits and drawbacks. A primary limitation stems from its reliance on a pre-existing labeled dataset to train the regression layer. This dependency inherently limits its usefulness in situations where labeled data is unavailable or has a non-uniform distribution. Conversely, this embedding model offers considerable advantages in terms of computational efficiency and methodological simplicity. Once the embeddings are computed, mapping them to target values can be achieved through relatively simple methods such as OLS linear regression. These techniques are not only computationally inexpensive but also offer a high degree of interpretability. This trade-off between resource efficiency and adaptability to specific tasks renders embedding-based approaches particularly appealing for applications constrained by limited computational resources or where model transparency is crucial.

Furthermore, it is important to consider the potential impact of the embedding space itself on the final task performance. The quality and representativeness of the pre-trained embeddings directly influence the effectiveness of the downstream regression model. If the embedding space does not adequately capture the nuances and relationships relevant to the target task, even a well-trained regression layer may struggle to produce accurate predictions. Exploring the characteristics of the chosen embedding model, such as its training data and architecture, and how well it aligns with the specific requirements of the evaluation task, could provide further insights into the observed performance and potential limitations.

Although Llama3.2 did not exhibit strong performance in the specific task of valence prediction, it remains a promising option for building lightweight conversational agents. In the present study, Llama3.2 completed all 140 inference requests in approximately 20 seconds, demonstrating its potential for real-time applications where rapid response is critical despite modest predictive accuracy in specialized tasks.

Turning back to the evaluation of the Phi-4 model, its performance in predicting hedonic valence achieved a Pearson correlation coefficient of R = 0.81. This result surpasses that of ChatGPT, which reached an R value of 0.72 under similar evaluation conditions. Importantly, Phi-4 attained this higher degree of correlation despite operating entirely offline, without reliance on cloud-based services or proprietary APIs. As an open-source model, Phi-4 outputs not only showed greater predictive accuracy but also highlighted the viability of local, transparent systems for affective computing tasks. Its performance reinforces the growing potential of small open-source language models in contexts that require data sovereignty, reproducibility, or deployment in constrained environments.

The comparative results presented here not only highlight the superior performance of Phi-4 in the specific task of

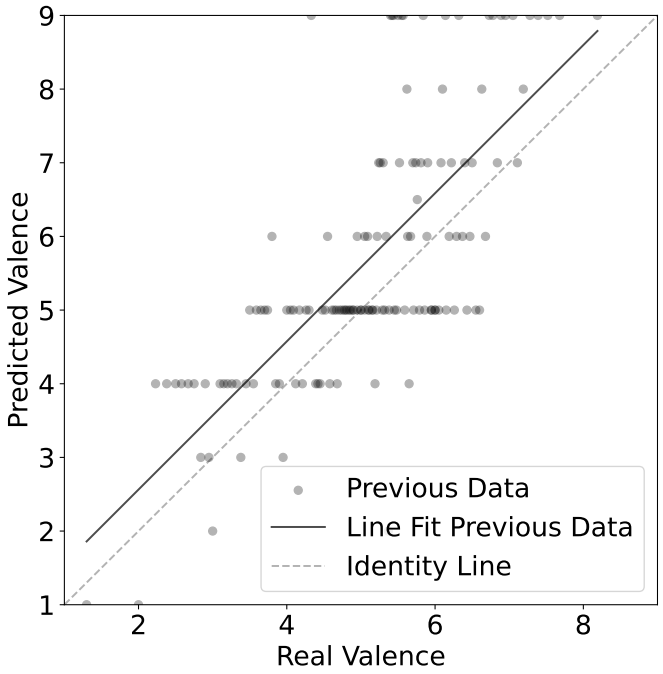


Fig. 4: Comparison between values generated by ChatGPT in [2] to the final set and the reference value for this set.

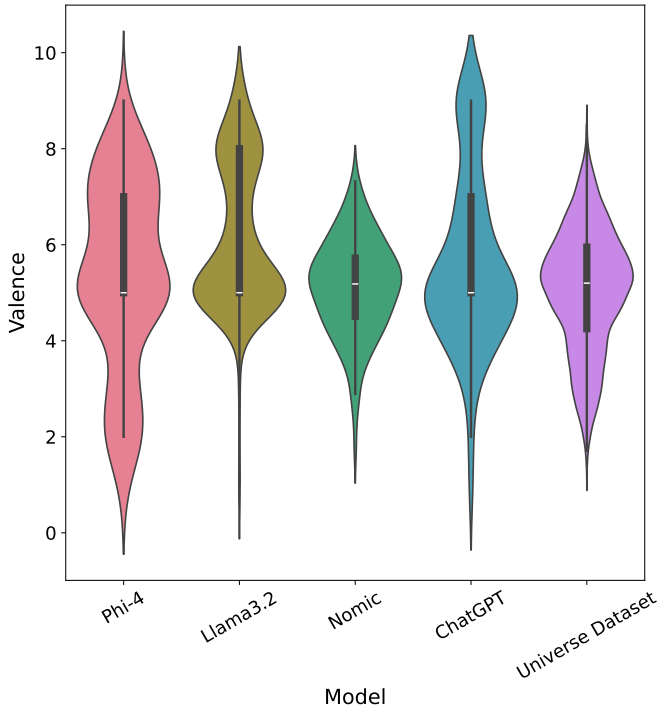


Fig. 5: Violin plot to the valences of each set (generated by the models and given by the universe dataset)

valence prediction but also reflect the rapid advancements in small-scale language models. The apparent improvement in correlation, even when using an offline, open-source model like Phi-4 with only 14 billion parameters, underscores the evolution of lightweight LLMs in recent years, pointing to a trend in which smaller models are becoming increasingly capable of performing complex semantic tasks with high accuracy and low operational overhead.

The results of the ANOVA-RM between the models in this study and ChatGPT from [2] revealed a statistically significant difference in performance ($F(3, 417) = 23.10$, $p < 0.001$). This indicates that there is a significant variation in the measured variable (valence scores) across the models being compared. Furthermore, the analysis suggests that these significant differences are primarily driven by the valence scores of the Llama3.2, as its T-Values in the Tukey Test were the highest in comparison to other pairs of models (Fig. 7).

Despite the embedding-based model (Nomic) achieving a higher correlation with the human ratings (Pearson's R $= 0.71$) compared to the Llama3.2 ($R = 0.39$), it is important to consider the nature of its training process. The embeddings were derived from textual representations trained on large corpora of language [9], and their evaluation in this study involved a supervised learning approach (OLS linear regression) to map the embedding-derived valence scores to human ratings. In contrast, Llama3.2, a large language model with approximately 3 billion parameters, generated valence outputs directly from prompts without further fine-tuning or prompt engineering for this specific model.

### A. *Limitations*

The performance of all models could potentially be improved through a dedicated prompt engineering process that may guide its output towards more accurate, relevant, and desired responses [11]. The sensitivity to changes in the prompt was not tested, but certainly is an aspect to investigate.

Similarly, the use of other LLMs was not tested and may lead to results different from those presented in this document.

While the human-rated dataset serves as the main standard for evaluating sentiment analysis models in this study, it is crucial to acknowledge the impact of the inherent subjectivity and variability present in human judgment. This dataset, for instance, provides mean valence ratings categorized by participant groups. A closer inspection reveals notable disparities for certain words. For example, the word "drought" received a mean valence of 4.71 from group 'Young' but a significantly lower 1.67 from group 'Old', with a difference of approximately 3 points on the 9-point scale. Similarly, "gym" was rated 7.18 by group 'Young' and 4.0 by group 'Old', highlighting substantial disagreement.

These instances underscore that "human consensus" is often an averaged representation across diverse individual experiences and perspectives, and that even a gold standard can contain internal variance. This variability in human ratings presents a subtle but significant challenge for models aiming to perfectly replicate human perception, as they are trained to approximate a mean that may not fully capture the full spectrum of human affective responses.

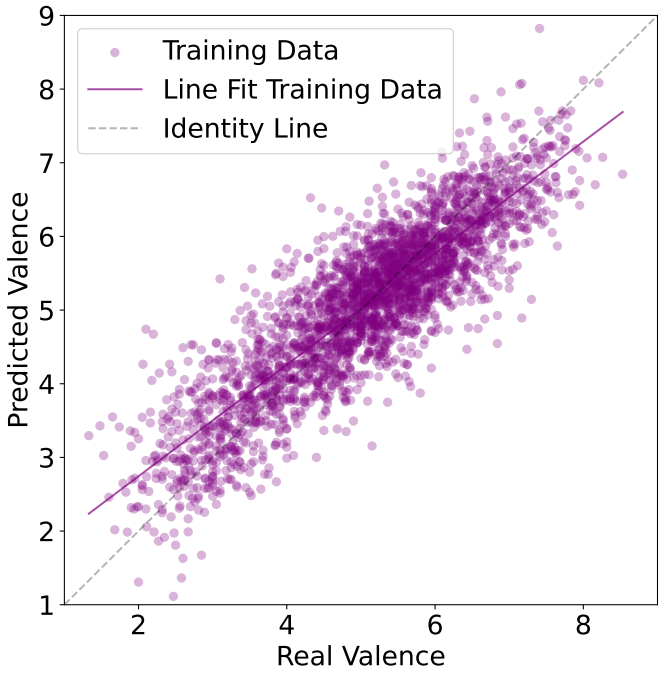


Fig. 6: Valence values generated by the embeddings model against the human-rated reference values for the words used in training. It represents an idealized scenario, assuming perfect overlap between training and test data.

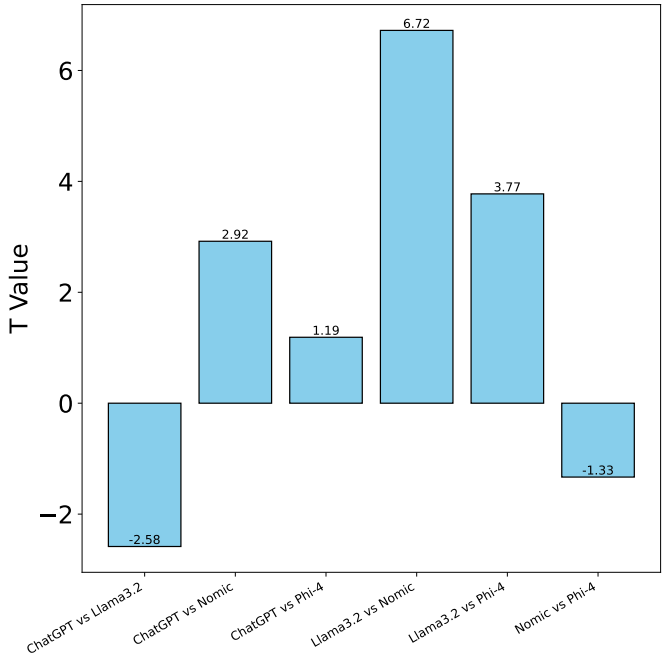


Fig. 7: Tukey comparison test between the valence values generated by the models.

## V. Conclusion

This small study answers the question posed, providing evidence that it may be feasible to use small large language models and embedding models without further fine-tuning, with short prompts and simple techniques, in the task of estimating the hedonic valences of words locally. Some sLLMs may be better-suited for the task, such as Nomic in comparison to Llama3.2. The ratings given by the models in the present study were found to be close to those of human groups. This finding is consistent with the results of [2], [4], and [5], which also reported close to human-level performance against human evaluations.

Finally, this study also shows that there is a linear relation, even if not perfect, between embeddings and hedonic valence of words. Thus, future works in the area may explore the use of other SAM Domains (dominance and arousal) with complete sentences, usage of images to predict their valence and try to simulate different groups like young and old.

## Acknowledgments

This paper was produced using the following tools:

- Gemini AI for grammar revision and phrasing suggestions.

P.C.V thanks Raphael Henrique from UFRJ for the fruitful conversations during the development of the study about the theme.

## References

[1] F. Wang, Z. Zhang, X. Zhang, Z. Wu, T. Mo, Q. Lu, W. Wang, R. Li, J. Xu, X. Tang *et al.*, "A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness," arXiv preprint arXiv:2411.03350, 2024.

[2] F. C. Jandre, G. C. Motta-Ribeiro, and J. V. A. da Silva, "Could large language models estimate valence of words? A small ablation study," in *Proceedings of CBIC*, 2023.

[3] B. M. Silva, J. V. Assumpção-Silva, W. Ricardo-Teixeira, H. Brener, M. Joffily, G. C. Motta-Ribeiro, and F. C. Jandre, "Valence Ratings of Lemmas Generated by a Large Language Model Simulating Multiple Human Responses," in *CBEB*, 2024.

[4] H. Alrasheed, A. Alghihab, A. Pentland, and S. Alghowinem, "Evaluating the capacity of large language models to interpret emotions in images," *PLoS One*, vol. 20, no. 6, p. e0324127, 2024.

[5] Y. Zhou, P. Xu, X. Wang, X. Lu, G. Gao, and W. Ai, "Emojis Decoded: Leveraging ChatGPT for Enhanced Understanding in Social Media Communications," *ICWSM*, vol. 19, pp. 2302–2316, 2025.

[6] V. Boykis, "What are embeddings," Zenodo, 2023. [Online]. Available: https://doi.org/10.5281/zenodo.8015029

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[8] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[9] Z. Nussbaum, J. X. Morris, B. Duderstadt, and A. Mulyar, "Nomic Embed: Training a Reproducible Long Context Text Embedder," 2024.

[10] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behavior Research Methods*, vol. 45, pp. 1191–1207, 2013.

[11] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT," arXiv preprint arXiv:2302.11382, 2023.