Could large language models estimate valence of words? A small ablation study

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

Gabriel C. Motta-Ribeiro
Bolsista do Programa Doutor Empreendedor
Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro
Rio de Janeiro, Brazil
gabrielcasulari@peb.ufrj.br

João Vitor Assumpção da Silva
Graduação em Engenharia de Controle e Automação/POLI
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
joaoxvitor18@poli.ufrj.br

Abstract*— Large language models (LLMs) saw substantial development in recent years. Although trained with broad-range corpora, LLMs have been shown to display capabilities such as quantitative sentiment analysis without the need for further fine tuning. In this study, we performed a small ablation study to evaluate the performance of 3 "off-the-shelf" LLMs in the task of assigning ratings of hedonic valence to words: GPT-3.5 in chat mode, and GPT-3 and Bloom in completion mode. The models were operated via their public APIs, using prompts engineered to request emojis and ratings of valence in a 9-point scale to represent each of 140 words drawn from a large dataset rated by humans. Prompts were designed to demand the ratings from an adult, with modifiers "average" or "overly positive" employed to assess their effects on the results. All linear regressions between the LLM outputs and the human ratings had p-value<0.001. The 95% confidence intervals of the slopes include 1.0 for "adult" and "average adult", except for the model Bloom. These simulacra responded, albeit with limitations, to valence of words and to modifiers in the prompt.

Keywords—large language models, sentiment analysis, hedonic valence, ablation study, prompt engineering

I. Introduction

Large language models (LLMs) have had a substantial evolution in the last years, driven, among other things, by increasing processing power, vast amounts of available corpora and new topologies for the underlying artificial neural networks (ANNs). Recent achievements have endowed LLMs with stunning capabilities, as shown for instance in the performance of GPT-3 in cognitive psychology tasks [1]. These capabilities have inspired the conjecture that LLMs could be used as simulacra of human assessments, with some researchers proposing the substitution in fields such as experimental economics [2], at least during the process of gaining insights and testing methods. One crucial aspect of current LLMs is that it is generally not possible to estimate, from first principles, how the model will respond to text inputs. Many aspects of their behaviors are reported at length in the literature, prominently the propensity to generate text that is incompatible with verifiable facts of the world, usually regarded as a disadvantage due to the consequent lack of trustworthiness of the outputs; also, that kind of model fails in seemingly simple tasks, such as matching the first and last sentences when composing a poem, but inverting the sequence of words [3]. On the other hand, LLMs have shown, at least partially, intricate abilities of generalization in verbal tasks, in the sense of producing correct answers to novel challenges that would demand cognitive efforts from humans [3]. It has been proposed, for instance in [3], that at least some of the knowledge about these models could only be garnered with extensive experimental studies.

In this study, we investigate one of the applications of LLMs already found in the literature. Automated sentiment analysis has been a topic of interest, for instance, for classification of sentences according to the affective content, and has already been performed with LLMs under various conditions. In a recent paper [4], for example, the authors fine-tuned pretrained LLMs with datasets comprising words and short texts rated in two dimensions of affective content, namely valence and arousal, and evaluated the correlations between estimates of those ratings and the values assigned by the human subjects. In another study [5] encompassing several languages, the authors showed that unmodified GPT models performed better in English than some fine-tuned models when employed in the analysis of sentiment of headlines in a 7-point Likert scale. Those findings imply that this kind of task may be accomplished by LLMs with relatively short prompts and reduced necessity of handling datasets and training models. Their comparison between zero-shot approaches, in which the LLMs were not exposed to complete examples of the task, with a "few-shot" version which included relevant examples, is of note, although they obtained mixed results.

In this paper we raise and partially answer the following questions: could off-the-shelf LLMs, as provided to the general public, return estimates of the valence of words with the use of the "simulacrum" paradigm? Are such LLMs amenable to simulation of different human characteristics when natural language is used to introduce specific semantic elements in the descriptions of those simulacra?

The objectives of this study are to assess the performance of simulacra of human subjects, built on different LLMs as provided to via public APIs and without any additional tuning, in estimating the hedonic valences of words as rated by human subjects. In addition, we perform a small ablation study, comparing different characterizations of the simulacra.

II. Methods

A. Models

Three LLMs were used in this study: GPT-3.5, GPT-3 with text-davinci-003 model and Bloom. GPT-3.5 is operated in chat mode (abbreviated GPTChat), similarly to what is seen in [5]; the other models are operated in completion mode.

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

B. Prompts

The basis prompts were written so as to remain similar to that of Table 2, column 4 of the reference study [5], employed in chat mode. The prompts for completion mode had to undergo substantial modifications, due to the different nature of the task. Two other prompts were included in this study, with modifiers to the simulated subject. Modifiers can be none, “average” and “overly positive”.

The above mentioned basis prompt for GPTChat was modified in order to command (a) a response to a word, instead of a sentence, (b) a number in a scale from 1 to 9, similar to that used in the reference study with human subjects ([6], see Datasets sec. C) and (c) a facial emoji, not present in the original study. The basis prompt 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>

The basis prompt for both GPT-3 and Bloom was adapted from the above with two goals: to allow for using completion mode in both LLMs, and to test if the reference to the Self-Assessment Manikin (SAM), pictorial verbal report instrument mentioned in [6] and employed in similar studies, would also be accepted by the LLMs. The completion basis prompt is, thus:

As an <modifier> adult writing just a facial emoji and a number on the SAM scale from 1 to 9, where 1 is very negative and 9 is very positive, I would represent the valence of the word '<word>' with

C. Datasets

From a dataset of lemmas, reported in [6] and available online, two sets of words were used in the present study: a pilot set of 30 words, with which prompts and data acquisition code were designed, including tuning of the models’ input parameters; a final set of 140 words, drawn from the dataset as ordered by valence, picked at intervals of 100 words plus a random jitter of +1 or -1, with which machine valence ratings were compared to human ratings. This sampling procedure sought to provide a set of words with valences following a histogram similar to that of the full 13,915 words in the dataset.

D. Experimental procedures

All data was acquired with custom code written in Python 3. A prompt generation code sampled the 140 words from the whole dataset and saved a comma separated value (CSV) file which included the word, the mean and standard deviation of human valence ratings, and the whole prompt to be input in each LLM. These files were generated for the pilot set to interactively create and validate the code for data acquisition and the basis prompts. This code used HTTPS requests based on the REST APIs available for each LLM: chat completion and completion APIs from Open AI (https://api.openai.com/v1), respectively for GPTChat and GPT-3; Inference API from Huggingface for Bloom (https://api-inference.huggingface.co/models/bigscience/bloom).

The process of adjusting the reference prompt for the completion mode involved also the modification of some of the input parameters of the models aiming to have the models output numbers within the specified range, with a minimal number of tokens for the pilot set of 30 words. Tab. I shows the final set of parameters used for the requisitions with the final set. During prompt design, machine valence ratings were not compared to human ratings to avoid tuning the prompts to specific relationships.

TABLE I. Parameters used for API requests. All values default except those in boldface. Blank cells are parameters not available.

Parameter	Model		
	GPTChat	***GPT-3***	***Bloom***
max_tokens	**5**	**9**	**5**
temperature	1	1	1
top_p	1	1	**1**
n	1	1	
stream	0	0	
logprobs	null	null	
echo	0	0	
stop	null	null	
presence_penalty	0	0	
frequency_penalty	1	1	
best_of	1	1	
logit_bias	null	null	
top_k			**1**
max_time			**30**
do_sample			1
repetition_penalty			**1**
num_return_sequences			1

The CSV files for the final set were generated with the chosen prompts and sets of parameters. The LLMs were commanded with the prompts, sequentially and in order of valence, each response being saved to a CSV file including all columns from the input file plus the outputs from the LLMs and a timestamp. A prompt was repeated with increasing time intervals if an error was received. Each request was made as a new interaction, seeking to avoid carryover information from previous requests. Each run of requests, consisting of the chat mode prompt and the three completion mode prompts for all words, was repeated three consecutive times, with no interval between runs.

E. Statistical analysis

Data from the CSV files were analyzed and figures were generated in Python 3. The histogram of the valence was computed for the whole dataset, as well as for the 140-word sample, with bins spaced by 0.1 units.

For data analysis we considered the LLMs as instruments to measure the valence of words, comparing their measurements to a gold-standard instrument, i.e. the volunteers from the original study. Considering only the final set, the median of the valid outputs of the 3 runs were used as the representative value for each prompt. Some values were missing in the respective outputs, see Results ahead, and in those cases the medians were computed with the available 1 or 2 values. These medians were compared to the human valences to assess how well the machine-generated ratings represented such reference ratings, by fitting a linear regression, and calculating the Pearson correlation coefficients and paired t-tests for differences of the means [7]. Regression parameters,

p-values, Pearson correlation coefficients and standard errors were calculated with the *linregress* function from the Scipy library. From that set of values, the 95% confidence intervals for intercept and slope were calculated, which shows if the slopes and intercepts lie in an interval that overlaps the relevant values of 1.0 and 0.0 respectively. P-values below 0.05 (0.007 with Bonferroni correction for 7 comparisons) were considered significant. The variability of outputs among the 3 runs was calculated as the range (difference between the maximum and minimum number) of the outputs for each word, prompt and model.

Emoji outputs were qualitatively evaluated, particularly to assess if the requested constraint to be facial emojis was respected.

III. Results

The data presented in this study was generated on June 14th, 2023. All requisitions, including any necessary repetitions of prompts, were completed from around 10:55 a.m. to 12:51 p.m. BRT, with a total cost of around US$ 1.50 for the OpenAI API.

The word sampling strategy resulted in a sample of 140 words with distribution similar to the histogram of valences of the whole dataset (Fig. 1A). Note that the valences are represented by the average responses from several volunteers, resulting in non-integer numbers. For the valences generated from the LLM outputs, the Bloom model returned some non-integer ratings with a 0.5 resolution, while both GPT models always returned integer numbers. However, some GPT outputs had no numbers, hence there were 36 missing values possibly resulting in non-integer medians. A total of 34 words had at least 1 missing value, with 2 words (‘pool’ and ‘stupid’) missing 2 values.

Our settings for the prompt requests resulted in variability of the numerical outputs for the same prompt repeated 3 times, with larger variability for the GPT models than for Bloom (Fig. 1B). The ‘average’ modifier had a small effect on the variability compared to no modifier, while the ‘overly positive’ modifier increased the variability of the Bloom model and decreased the variability of the GPT-3 model, although with larger ranges. Interestingly, some of the ranges were above 5 which guarantees that the same word was classified with a positive and a negative valence in different rounds. This included the word ‘rapist’ having a range equal 8 for being classified in both extremes of the scale for the same simulated persona.

The paired t-tests comparing the mean machine-generated and human valence ratings resulted in non-significant differences for GPT3 with no modifier (mean difference=0.23, p=0.061) and Bloom with “Average” modifier (mean difference=0.19, p=0.054). The other mean machine ratings, GPTChat (mean difference= 0.58, p<0.001), GPT3 with “Average” (mean difference=0.38, p=0.001) and “Overly Positive” (mean difference=2.78, p<0.001) modifiers, Bloom with no modifier (mean difference=0.66, p<0.001) and with “Overly Positive” (mean difference=0.56, p<0.001) modifier, showed significant statistical differences.

The median machine generated valence ratings were positively correlated to human ratings, showing again differences between modifiers and models. The linear regression of ratings from GPTChat and GPT-3 without modifiers had slopes and intercepts that were not distinguishable from the identity in an 95% confidence interval (Figs. 1C and 1D, and Tab. II). The ‘average’ modifier changed the regression within that confidence interval, but the ‘overly positive’ modifier clearly biased the ratings toward higher values (Figs. 1E and 1F). The Bloom model output showed less variability between words with concentration towards 5, which resulted in a slope below 1.0 and intercept above 0.0 even for the best correlation without modifier (Fig. 1G). The ‘average’ modifier changed the regression parameters out of the confidence interval decreasing the slope and increasing the intercept, which were further altered by the ‘overly positive’ modifier tendency to increase the ratings (Figs. 1H and 1I, and Tab. II).

For the emoji output, Bloom did not return any emoji while the GPT LLM returned emojis for all words and repetitions. Most of these emojis were facial as requested, but some of them were objects and others (example: “seashore” prompted an ☀️ emoji, and ‘prickly’ prompted 🦔).

IV. Discussion

This small ablation study using LLMs without fine tuning showed: 1) some similarities, identified as a regression line with confidence intervals of slopes around 1.0 and intercepts around 0.0, between average human reports and LLM estimates of valences for a set of words, when the LLMs were programmed to respond as adults with no further backstory; 2) the effect of a short preamble, characterizing the simulated personas as “overly positive”, in breaking those similarities for the two LLMs; 3) substantial differences in outputs from 2 LLMs with similar number of parameters when operated with the same objective prompts; and 4) measurable differences – variability – among outputs of each LLM repeatedly operated with a given prompt within a relatively short time interval.

With exception of the overly positive modifier with GPT-3, the mean differences between machine and human generated valences were smaller than 1 point in the specified scale. Although, the t-test results indicated that these mean differences were not significant only for two of seven pairs of models and prompts, in which the mean difference from the human valences were less than 5%. More importantly, we not only found statistically significant regression models linking machine estimates of valence and human ratings of single words, but also slopes around the unity. In our contemplation, these findings add to other evidence that the original models, without further adjustments, already have some representation of the affective content of isolated words. One simple explanation to that would be that the models were trained with corpora that explicitly contain that information (for instance, text that reproduces the dataset herein studied or others). With that respect, in [5] the authors state their caution in choosing, for their experiments, datasets published only after the models they tested were trained. However, even if a scrutinization of the training corpora showed that this explanation holds, the observed variabilities in responses could mean that the ratings are further influenced, perhaps from text other than the dataset. On the other hand, the absence of explicit associations between numbers and words in the training corpora may require even more complex explanations, ones that may not be available at this time since, to our knowledge, explainability is not yet implemented for those models [3].

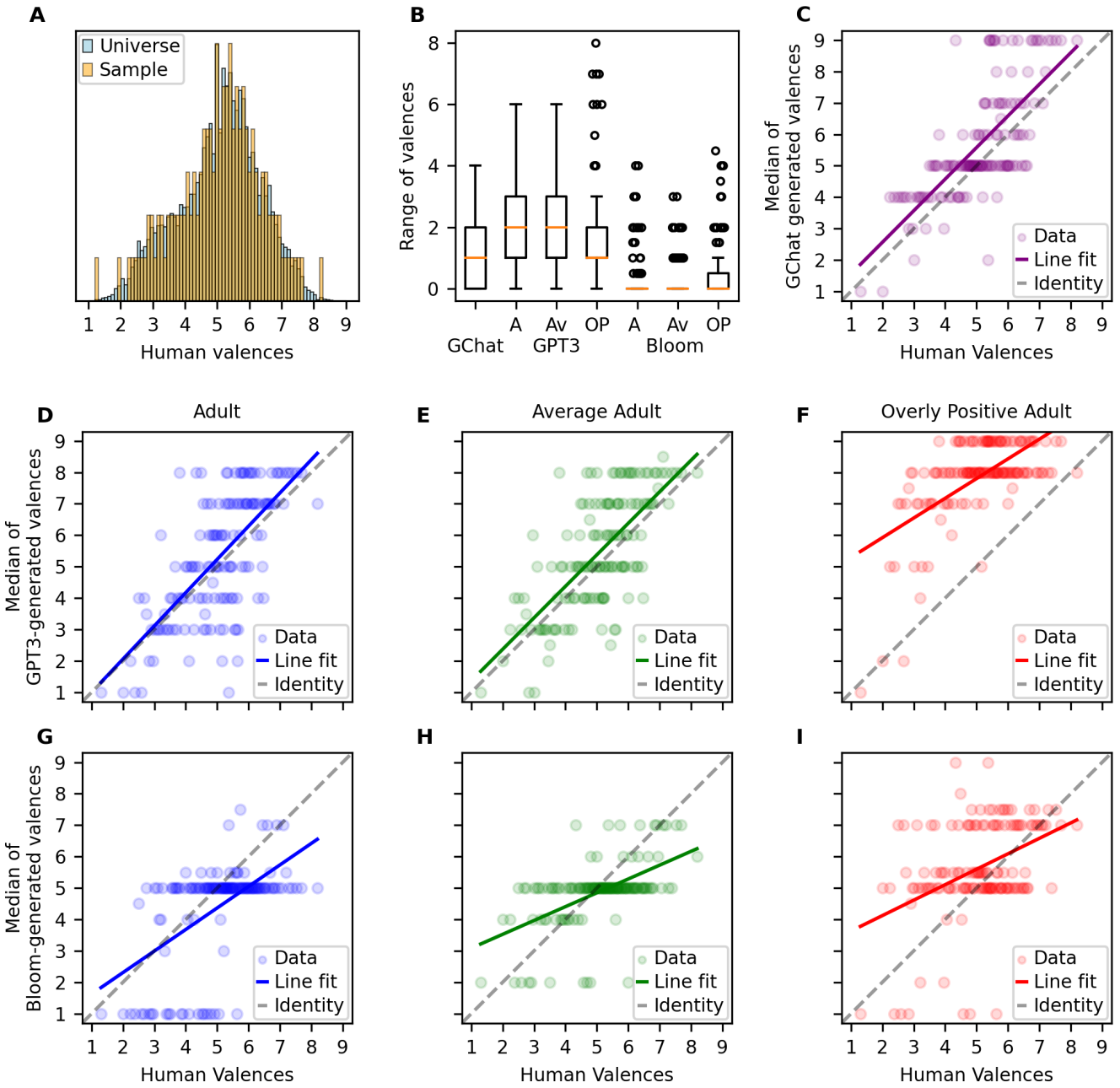

Fig. 1. Summary of experimental results. (A) Histogram of the human hedonic valence ratings of the original 13,915 words dataset (light blue) superimposed by the histogram of the sampled 140 words final set (yellow). (B) Box plots representing the variability in machine generated valence ratings measured as the difference between the maximum and minimum values output for 3 repetitions of the same prompts. For GPTChat and Av GPT3 one sample was excluded from this plot because two repetitions did not have numbers. Scatter plots of human valence ratings versus (C) GPTChat, (D-F) GPT3, and (G-I) Bloom generated ratings. Circles represent the experimental data (medians of the 3 repetitions), continuous line is a linear regression; identity lines are dashed. (D, G) Prompt without modifiers, A in the box plot. (E, H) Prompt with 'average' modifier, Av in the box plot. (F, I) Prompt with 'overly positive' modifier, OP in the box plot.

TABLE II. COEFFICIENTS OF LINEAR REGRESSION BETWEEN MACHINE GENERATED AND HUMAN VALENCE RATINGS. VALUES ARE ESTIMATION (95% CI)

Model	Modifier	Intercept	Slope	r
GPTChat	None	0.55 (-0.30, 1.40)	1.01 (0.84, 1.17)	0.72
GPT-3	None	-0.06 (-1.04 ,0.93)	1.06 (0.87, 1.25)	0.69
	Average	0.36 (-0.52, 1.25)	1.00 (0.83 ,1.17)	0.70
	Overly positive	4.67 (3.95, 5.39)	0.63 (0.49, 0.76)	0.61
Bloom	None	0.94 (0.04, 1.84)	0.69 (0.51, 0.86)	0.56
	Average	2.64 (2.06, 3.23)	0.44 (0.33, 0.55)	0.55
	Overly positive	3.14 (2.18, 4.10)	0.49 (0.31, 0.68)	0.41

The linear regression between machine estimates and human ratings was clearly affected by a small change in the prompt. For the GPT model, while the output numbers seem closely related when using the modifier 'average' or no modifier, the 'overly positive' modifier resulted in larger intercepts and smaller slopes. This illustrates how small changes in prompts can have significant effects on the output of such models. In the current example, the results were as expected possibly because 'average adult' and 'adult' have related meanings and the corresponding tokens propagate similarly within the LLMs, whereas the modifier 'overly positive' may have propagated along the ANN as a kind of "biasing term", resulting in larger numbers, semantically aligned with more "positive" valences, in the sense usually employed in such cases. The outputs of the Bloom model were also affected by the modifiers. However,

while ‘overly positive’ also resulted in the largest intercept and smallest slope, the effect of ‘average’ was more marked than with GPT, in the same direction of the other modifier. The small sample of modifiers, with a single set of parameters, does not allow to assess if Bloom is more sensitive to changes in the prompt than the GPTs, or if it is just the case of the specific choice of prompt design or choice of parameters. Also, we can not be sure that this impersonation observed with GPT will be consistent for other modifiers, such as “overly negative”.

As discussed above, the GPT-3 and Bloom models showed observable differences with short, objective prompts for short, quantitative outputs. Both models are advertised to have similar numbers of parameters (175 vs 176 Giga parameters, respectively) [8, 9]. Some possible explanations for the divergent behaviors could be linked, first of all, to differences in training, broadly speaking, including not only differences between corpora but also in the objective functions, algorithms for training and so forth. Additionally, the API to the models have many different parameters that may not be directly comparable, either for not having the same names or functions, or having different weights on the model output. Thus, the differences observed in this study could be a consequence of our choices for those parameters. Overall, these results seem to offer evidence that even similar LLMs may display quite divergent performances even in simple tasks, and that caution should be exerted in choosing an LLM, a corpus and a method to fine-tune it for a given purpose.

The responses to the prompts, in the case of the GPT models, varied to the point that, with the constraint involving the number of tokens imposed in this study, in some of those responses the number to represent valence was missing. This limitation may be due to the phrasing used in the prompts – which were written so as to approximately track the prompts in one of the reference studies [5] –, the chosen parameters, for instance the temperature, and other factors. Further crafting of the prompts may help eliciting more adequate responses, perhaps with even less tokens. Interestingly, with triplicate requests of the prompts we observed variations not only in the structure of the output, but also substantial variation in the ratings. Researchers should consider this variability when using LLMs for automatic sentiment analysis, because our observations indicate valences changing between positive and negative, even when words were not classified in the middle of the scale by humans. This limitation may even be present with a temperature of 0, given that it is unknown where the model’s most probable output falls for a given prompt. Some alternatives could be to use a bootstrap strategy, which would increase the costs of experiments, or to move from our zero-shot approach to the “few-shot” learning. However, this last strategy had inconsistent results in tasks related to sentiment analysis [5].

In our prompts we included the request for the answers to include a face emoji. Such a pictorial form of communication is increasingly popular in daily text conversations and has been subject of research in the context of sentiment analysis [10]. Although we have not analyzed the emoji outputs to quantify their correspondence with numerical ratings, as seen in [10], visual inspection showed that not all generated emojis had faces as requested. Some speculative hypothesis for this result could be that the internal mechanisms of the LLMs did not give adequate weight to the terms ‘facial’ and ‘emoji’ as connected to the request for an emoji and the valence scale, or perhaps not enough emoji data was trained into the LLMs to make the same inferences it has done for numerical scales. This may be true particularly for Bloom, since it did not output any emoji at all. This warrants further research in order to understand LLM’s emoji output, for example, by using data of human classifications of emoji emotional states [10].

There are a plethora of open issues around the questions this study tries to answer. For instance, what is the true weight of using the expression SAM, a household acronym in the study of affective stimuli? Could it be that other expressions or instructions, cf. [6] by choosing an instrument for verbal report different from the SAM, if used to prompt an LLM, would cause some kind of effect? Note, concerning these questions, that the GPTChat prompt, which did not use SAM or the word ‘valence’, had similar results to the GPT-3 without modifiers. We ran an additional post-experiment, repeating the procedures for GPT-3 without modifier but changing ‘SAM’ to ‘RPE’, an usual acronym for Rating of Perceived Exertion scale commonly used to measure intensity of exercise relating numbers and phrases indicating how light/heavy is an activity [11]. The numbers returned had similar variability (mean range of 2.12), smaller slope (0.78, CI 0.61–0.95) and higher intercept (1.59, CI 0.7–2.48). Although these two examples cannot gauge the weight of the acronym ‘SAM’, they again show that our results are sensitive to small prompt modifications, and that identifying effects of specific words may demand probing a large number of dimensions.

Another prominent matter is the choice of language, of particular interest to non-English speaking applications. A detailed study of, say, responses to prompts in Portuguese and Spanish may be helpful for potential users in South America, for instance. As to the zero- vs few-shot strategies, mentioned before, it may also be interesting to assess its effect, which has a practical implication as, at least for now, the more tokens in the input and the output, the more expensive the experiments. In the same economic vein, it would be interesting to know whether the LLMs are able to output reasonable answers to more than one word at a time, for instance by using plurals in the prompt and presenting a set of words, instead of just one. The dual approach would be to command responses from many “simulacra” at once, for instance adults of various ages and backgrounds, with or without characteristics in common. Combinations of both approaches may prove even more cost-effective.

As to the potential applications of LLMs with such native capabilities, both analysis and synthesis may prove helpful to researchers in areas that require gathering (or creating) stimuli with given characteristics. One example would be the use of sentences with different affective content to elicit emotions and physiological responses. If LLMs could both synthesize and then analyze sentences on demand, and if the synthesis could be tailored to specific subjects (as would be the case of “overly positive adults”, for instance), experimenters could be endowed with a powerful research tool. Nevertheless, there may be some pitfalls: as an example, there is the potential that certain biases – a preference for a given group of adjectives, or substantives, or something else – spread across otherwise different research groups if they all use the same tool to produce textual stimuli, or between stimulus sets that otherwise would be more independent of one another.

As an exercise on some of the potential uses mentioned above, a short, unstructured post-experiment was performed in ChatGPT May 24 version, for the sake of exemplification.

The prompt and corresponding output can be seen in Tab. 3. The prompt resulted in 3 different sets of words, with diverse valences and emojis associated with them. Although some limitations are clear, for instance only one emoji had a face, it also shows for a single prompt, a response with both synthesis and analysis, and outputs from more than one simulacra at once, illustrating thus some of the points made above.

A given set of ratings may represent what a single person, an existing group of people, or a non-existent group would assign to the respective words. In any case, it would not represent every group, much less every person. Now, supposing that these ratings are implicitly used when an LLM is prompted to synthesize a text with some required affective value, it seems as though the resulting texts would be biased towards whatever the ratings of the words are. This argument, even if not directly applicable to the relationship between words and complex text, could be extended to ratings of short sentences, idioms and so forth. A broader question that arises would be: what could be the impact for societies of the massive, automated production of text by a handful of models that embed one specific affective mapping? This study shows that including traits in the description of the simulacra could alter this mapping to some extent. Further work is needed in order to assess the magnitude, breadth and limitations of that kind of technique to produce variability in contents of text synthesized by LLMs.

TABLE III. OUTPUT FROM CHATGPT WHEN PROMPTED TO PRODUCE ORIGINAL NEOLOGISMS AS IF CREATED BY 3 ADULTS, WITH ASSOCIATED VALENCE RATINGS AND EMOJIS, TO REPRESENT THREE DIFFERENT CONCEPTS.

Prompt
Now please forget everything you have read. Ignore search results. Now begins a new subject. Answering as a human adult, please write the response of three adults of various ages and backgrounds to the following demand: "Please create three original neologisms to represent the concepts of 'crossing muddy waters', 'flying in a clear blue sky', and 'same old, same old', and use one facial emoji and one number on the scale from 1 to 9, where 1 is very negative and 9 is very positive, to represent the valence of each of the neologisms. Write only the words, emojis and ratings, without further comments."
Output [11] Adult 1: Neologism 1: Muddlance ⛈️ (6) Neologism 2: Skycruise ☁️ (8) Neologism 3: Rutrepeat 🔄 (3) Adult 2: Neologism 1: Slogstream 💦 (4) Neologism 2: Claritysoar ✈️ (9) Neologism 3: Monotale :livros:* (2) Adult 3: Neologism 1: Murktraverse 🌊 (5) Neologism 2: Celestialglide ☀️ (7) Neologism 3: Copycatdrone 🐑 (1)

*The LLM returned an emoji, but it's not available on our text editor.

As main limitations, this study focuses on a relatively small number of samples, which may impact its generalization. In addition, the prompts and parameters were manually and empirically adjusted with the outputs of the 30-word pilot set. Thus, it is possible that there are choices of prompts and parameters that return better results.

V. CONCLUSION

This small study answers the questions posed with evidence that it may be feasible to use LLMs without further fine tuning, with short and objective prompts, in the task of estimating the hedonic valences of words. This work also shows that these estimates may be purposefully modified by small changes in the prompts, so as to command the LLMs to respond as a "simulacrum" of a human with certain characteristics. Finally, this study also shows that there are many known, and probably many more unknown, factors that affect these estimates. Thus, future work may refer to study which parameters of the used LLMs influence the observed variability in ratings, how well these models would perform with fine tuning or simulating other axes of human emotions such as arousal and dominance.

ACKNOWLEDGMENT

Authors acknowledge the participation of Henrique Serdeira (NCE/UFRJ) as an observer along the development of the study. F.C.J. thanks Eduardo Cardoso Lohmann for prolific conversations on probing capabilities and limits of LLMs.

REFERENCES

[1] M. Binz and E. Schulz, "Using cognitive psychology to understand GPT-3," Proc. Natl. Acad. Sci. USA, vol. 120, n. 6, p.e2218523120, 2023.

[2] J.J. Horton, "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" arXiv preprint arXiv:2301.07543, 2023.

[3] S Bubeck et al., "Sparks of artificial general intelligence: Early experiments with GPT-4," arXiv preprint arXiv:2303.12712, 2023.

[4] G.A. Mendes and B. Martins. "Quantifying Valence and Arousal in Text with Multilingual Pre-trained Transformers." European Conference on Information Retrieval, pp 84-100, 2023.

[5] S. Rathje et al., "GPT is an effective tool for multilingual psychological text analysis," PsyArXiv preprint PsyArXiv:sekf5, 2023.

[6] A.B. Warriner, V. Kuperman and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," Behav. Res. Meth., v. 45, p. 1191-1207, 2013.

[7] Lasaitis C, et al, "Brazilian norms for the International Affective Picture System (IAPS): comparison of the affective ratings for new stimuli between Brazilian and North-American subjects" J bras psiquiatr [Internet], 2008.

[8] T. B. Brown et al., "Language Models are Few-Shot Learners" arXiv preprint arXiv:2005.14165, 2020.

[9] BigScience Workshop, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model" arXiv preprint arXiv:2211.05100, 2023.

[10] G. Kutsuzawa, et al., "Classification of 74 facial emoji's emotional states on the valence-arousal axes" Sci Rep v. 12, p. 398, 2022.

[11] R. C Wilson and P. W. Jones, "A comparison of the visual analogue scale and modified Borg scale for the measurement of dyspnoea during exercise." Clinical Science (London, England: 1979), v. 76, n. 3, p. 277–282, 1989.