Word predictability in Portuguese: Cloze norming study vs. LLMs

Author: Jane Aristia[1]
j.aristia@gmail.com

[1]Modal'X, Université Paris Nanterre, Nanterre, France

ORCID ID : 0000-0002-7668-7043

Abstract

With the rise of large language models (LLM), there has been deemed a possible alternative to human participants in many scientific domains, including linguistic studies, the cloze study. Cloze probability is used to inform researchers as to how predictable a word is within a certain sentential context. It is a common tool in linguistic studies to understand language production and processing. Previous studies (e.g., Jacobs et al., 2022; Lopes Rego et al., 2024) have compared LLM performance with traditional cloze studies and their results are promising. Nonetheless, these studies were done in English. Hence, we would like to know LLM performance in the Portuguese language. Here, we compared results from a traditional cloze study with two LLM, such as: Grevásio (Santos et al., 2024) and Tucano (Corrêa et al., 2024) then performed a correlation analysis to investigate their performance. The results show a moderate and weak correlation between the cloze probability from human participants and the LLMs. These results highlight the gap between human performance and LLM, specifically in cloze probability.

1. Introduction

The growing interest in understanding language prediction is accompanied by the increased need for cloze norming study as it gives us information about the predictability of a word within a sentence. Traditionally, this norming study requires participants to complete a sentence (e.g., "The students are studying hard because tomorrow they will have an ___ "). It measures the predictability of a target word based on the sentential context. The cloze probability is obtained by calculating the proportion of participants' word responses for each sentence. However, with the advancement of the large language model (LLM), we observe the shift towards LLM as an alternative. Some researchers argued that LLM is comparable to human response and can be used to understand human cognition (Hu et al., 2022, 2024). However, this idea is still debatable, as those who oppose it argue that LLMs use different mechanisms and it is premature to use them to understand human cognition (Katzir, 2023; Leivada et al., 2024).

Despite these debates, researchers still try to use LLM to understand linguistic processing. For instance, it has been used for linguistic tasks such as cloze norming study (Jacobs et al., 2022; Lopes Rego et al., 2024). In a traditional cloze norming study we need to collect human participant data, which takes time and sometimes we also need to provide financial compensation for participants. In short, this traditional method is time-consuming and costly. On the other hand, LLM enables us to obtain data efficiently and faster. LLM allows us to skip this data collection step because the cloze probability is calculated through the tokenization of sub-words of each word in a sentence. Therefore, LLM could be deemed as a promising alternative to the traditional cloze norming study.

To make sure that LLM could provide a good word probability, there have been studies that compare cloze norming studies with LLM (e.g., Jacobs et al., 2022; Lopes Rego et al., 2024) and the results showed that they are indeed comparable. For instance, Lopes

Rego et al. (2024), compared the traditional cloze norming study with cloze results from LLMs such as GPT-2 and Llama to investigate the reading models such as OB-1 reader (Snell et al., 2018). Their study showed that LLMs' results are more accurate in predicting human eye movement towards the anticipated words than the traditional cloze study. These studies seem to suggest that it is indeed promising to use LLM as an alternative to traditional cloze study. Nonetheless, we need to be cautious with the possibility of overfitting from LLM.

Looking at how LLM is comparable to traditional cloze study, we are interested to know if this is also applied in European Portuguese because most of the previous studies were done in English. So, this study aims to see if cloze probability from LLMs is comparable to Portuguese human cloze probability. For this purpose, we used cloze norming datasets from Aristia et al. (in preparation) and for the LLMs we used two Portuguese text generation models such as Grevásio (Santos et al., 2024) and Tucano (Corrêa et al., 2024). These were the most recent openly accessible text generation models we could find.

2. Methods

2.1 Cloze norming study

The data were taken from cloze norming studies in Aristia et al. (in preparation) wherein they were adapted from the sentence pool in Frade et al (2021). Aristia et al. (in preparation) conducted two cloze studies; the first study was to get the probability of the article in each sentence and the second to get the probability of the noun. In this present study, we only used 117 sentences that were used in their experiment and were obtained from the second cloze study. The average age of the participants is 27.69 years old (age range: 20-50 years old). There were 45 female participants from 125 participants.

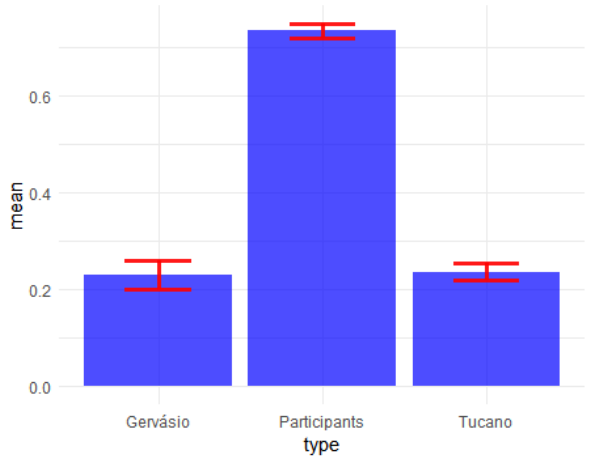
2.2 Cloze probability with LLM

For the LLMs, we used Grevásio (Santos et al., 2024) and Tucano (Corrêa et al., 2024). Grevásio is an open-source trained decoder model from the LLaMA family. This model was trained with a supervised fine-tuning method wherein the dataset was labelled. The dataset was from GLUE and SuperGLUE that were machine translated into European Portuguese. The other model that we used here is Tucano, a Transformer-based model that is pre-trained in Portuguese. Nonetheless, here we used the supervised fine-tuning model of Tucano, which is called Tucano-2b4-Instruct. It is trained with several datasets such as the Portuguese version of 1 million GPT4, Orca word math problems (Mitra et al., 2024) in Portuguese, and the Aira dataset (Corrêa, 2024). Nonetheless, Tucano, unlike Grevásio, was trained not only in European Portuguese but also in Brazilian Portuguese.

Further, to obtain the target word's probability from the LLM, we adapted the code from GPT-2-for-Psycholinguistic-Applications developed by Samer Nour Eddine which allows us to use LLM for the Portuguese language. The sentences were parsed into sub-word tokens and each word position was marked. The cloze probability of each word was obtained by calculating the conditional probabilities of each word's sub-word token.

Figure 1

Average Cloze probability from the Cloze norming study, Grevásio PT and Tucano BR



Note. The bar reflects the standard error (SE).

3. Results

3.1 Cloze probability

3.1.1 Cloze norming study with participants

The average cloze probability of the sentences was .73 (range= .41 - 1, SD = .17), depicted in Figure 1.

3.1.2 Cloze obtained through LLM

The average target word probability (see Figure 1) from Grevásio was .23 (range = 7.79E-08 - .99, SD = .32), and Tucano was .24 (range = 8.55E-05 - .85, SD = .19).

3.2 Data similarity analysis

Figure 1 illustrates the distribution of the data. To verify the similarity between the distributions of data obtained from human participants and LLM distance analysis using Jensen-Shannon method (Cha, 2007) was performed. To run this analysis the *'philentropy'* package (Drost, 2018) in R (2000) was used. The results showed that the LLM data of both Grevásio (p<.001) and Tucano (p<.001) was significantly different from the traditional cloze study using human participants.

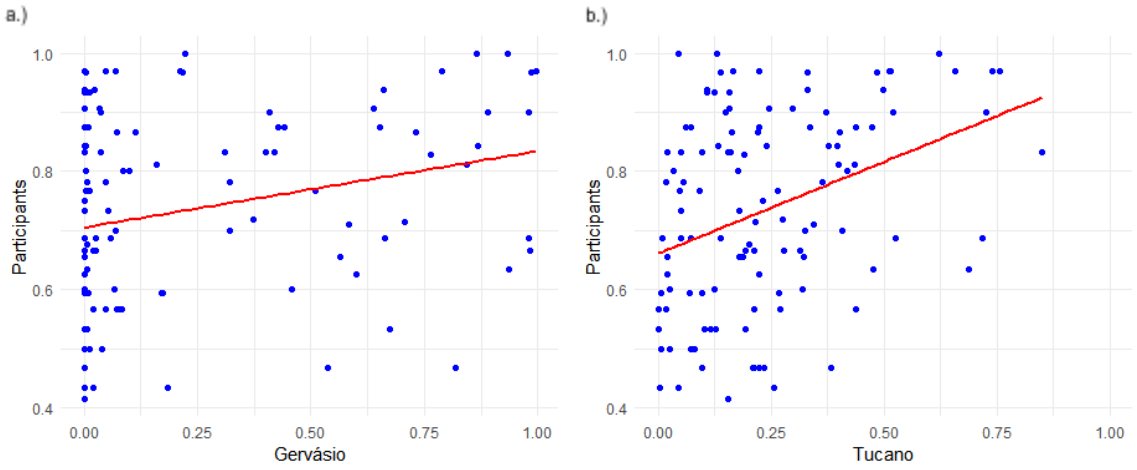
3.3 Correlation analysis

To investigate if there are correlations between traditional cloze study and LLMs, we conducted Spearman correlation (Wissler, 1905) in R. It is a non-parametric test that looks at a monotonic relationship, which is less restrictive than assuming linearity in the data. It also determines the strength and direction of this observed correlation. Through this analysis, we found that Tucano performed better than Grevásio when their results were compared with the traditional cloze study. We observed significant and weak positive correlation (See Figure 2) between the word probability from traditional cloze study with Grevásio, $r(115) = .25$, p =.007, 95%CI [.10, .45]; and, moderate significant correlation with Tucano, $r(115) = .36$, p

< .001, 95%CI [.19, .51]. These confidence intervals (CI) were obtained through the bootstrapping technique, which is a method that can estimate the uncertainty in the data through random resampling, without assuming normal distribution in the data. We can see in Figure 3 that the correlation values for both Grevásio and Tucano fall within the CI range. Although their values were overlapping, Tucano performed slightly better than Grevásio as the lowerband of the CI was higher for Tucano.

Figure 2

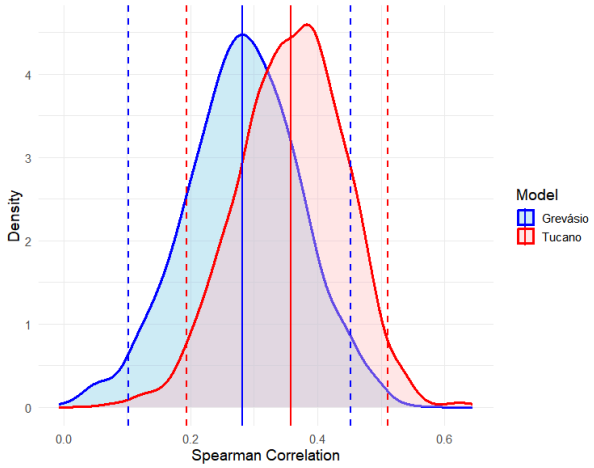
Plots depicting correlation between word probability from cloze study.



Note.a.) Correlation between participants' data and Grevásio; b.) Correlation between participants' data and Tucano

Figure 3

Bootstrapping distribution of Spearman correlation analysis



Note. The dashed line depicts the 95% confidence interval (CI) from each model, blue for Grevásio and red for Tucano, while the solid line depicts the mean correlation of each model. For Grevásio, the CI is between .10 and .45. For Grevásio, the CI for Spearman is between .19 and .51.

4. Discussion

The purpose of this study is to see if LLMs can be used as an alternative to traditional Cloze studies. Visually, in Figure 1, we can see that there are huge differences between the distribution of cloze probability obtained by human participants versus those using LLMs. This observation is confirmed by the similarity analysis wherein the cloze probabilities obtained through LLM are significantly different from the human participants data. Despite that, the statistical results show there are significant correlations between them. Nonetheless, it needs to be noted that they are not strongly correlated as seen in Figure 2. This indicates that for Portuguese language, the LLMs are not yet ready to be used as an alternative to traditional cloze study with human participants.

Differences in the mean probability between the cloze probability from human participants and LLM, as seen in Figure 1, and weak to moderate correlations in Figure 2, are in line with the findings of Jacobs et al. (2024). They conducted four experiments comparing the cloze study from Peele et al. (2020) with three LLMs: GPT-2 (Radford et al., 2019), RoBERTa (Liu et al., 2019), and Pythia (Biderman et al., 2023). In the first experiment, they compared probabilities from the cloze study with the model probabilities and they found no linearity in the correlation. In the second experiment, they used the rank of probable responses from both the cloze study and LLM, and again the results showed weak correlation between them. In the third experiment, they aimed to assess if the model training affects the fitting to the human data. In the fourth experiment, they conducted a clustering analysis to evaluate the semantic production of humans and LLM. In short, their first two studies focused on showing differences between human cloze probability and LLM.

LLM calculates the probability between words in a sentence differently, that is why we could observe differences between data from the human participant data and LLM in the present study. Günther and Cassani (2025) argued that LLM uses probability between tokens to predict the upcoming words. Therefore, in a sentence completion task like a cloze study, LLM tends to use a frequent word that may be less grammatical, or a less probable word that may be more grammatical (Katzir, 2023). For instance, Katzir (2023) showed that to continue a sentence such as: *"The little duck that met the horses with the blue spots yesterday __",* the GPT model preferred to use *'are'* rather than *'destroys'* as a continuation of this sentence. On the other hand, human participants are able to make grammatical sentences with a less probable word because they can make use of the sentential context. In the same vein, Cai et al. (2023) study also showed that LLMs do not really rely on context to resolve syntactic ambiguity in a sentence as it relies more on sub-word tokens. Another possible reason for

these weak correlations is perhaps due to the way the LLMs are fine-tuned, as they tend to perform well on the task related to their fine-tuning focus (Denning et al., 2025).

Nevertheless, the weak to moderate correlation between human participants' cloze probability and LLMs could also be interpreted to indicate that there is a possibility to use them as complementary data with human participants instead of an alternative. Caution not to use LLM as an alternative also comes from a recent *Event-Related-Potentials* (ERP) study (Arkhipova et al., 2025). ERP is a common method in psycho/neurolinguistics studies that investigate linguistic information that is used by the brain during language comprehension.

Apart from that, in the future, it is not impossible to use LLM as an alternative to human data. To achieve this, improvements are needed. For instance, additional data training and fine-tuning in Portuguese text is needed, specifically with cloze task. Also, Portuguese variance need to be taken into account, for instance there are differences in the word choice between Brazilian Portuguese and European Portuguese. Thus, taking this into account will increase the accuracy of LLM.

5. Conclusion

All in all, this study is an early attempt to compare traditional cloze study using human participants with Portuguese LLM. Portuguese LLM has not yet achieved the same level of performance in traditional cloze tests as in English. Nevertheless, it did show a similar trend as seen in the correlation results, wherein we observed there is an increase of LLM probability for words with higher cloze probability and particularly more stable and promising results from Tucano than Grevásio. Hence, it is favorable to recommend LLM as complementary data with human participants' data rather than as a replacement and it needs to be kept in mind that LLM computes words' probability differently than human. To enhance human-like performance of LLM, further studies and improvements are required.

6. Competing interests

The author declares that there are no competing interests.

7. Data accessibility statement

The data, codes, and materials that support this study are available in https://osf.io/f8jrv/ .

8. Declaration of use of AI

The author declares that ChatGPT was used for information search, grammar checking and sentence rephrasing to make it clearer under the supervision of the author. It was also used to assist the coding part.

9. Ethics and consent

Data from cloze study that involved human participants was part of an EEG study (Aristia *et al.*, in preparation) that has been approved by the ethic committee of *Faculdade Psicologia, Universidade de Lisboa*.

10. Author's role

Jane Aristia : conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, writing.

11. Acknowledgement

The cloze data from this study was part of an EEG study (Aristia *et al.*, in preparation) that was conducted at Prof AP's lab, VoicES lab, Universidade de Lisboa, in collaboration with SF, LAPSO-ISCTE. The author also thanked SNE for his advice on the comparison

between human cloze and LLM; and, PZ who proofread the manuscript and check the grammar of this manuscript.

Link to preprint

https://doi.org/10.31234/osf.io/e8fzj_v2

References

Aristia, J., Frade, S., & Pinheiro, A. *(in preparation)*. Prediction by production in spoken sentence processing.

Arkhipova, Y., Lopopolo, A., Vasishth, S., & Rabovsky, M. (2025). When Meaning Matters Most: Rethinking Cloze Probability in N400 Research. *bioRxiv*, 2025-04.

Cai, Z. G., Duan, X., Haslett, D. A., Wang, S., & Pickering, M. J. (2023). Do large language models resemble humans in language use?. *arXiv preprint arXiv:2303.08014*.

Cha, S. H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City*, *1*(2), 1.

Corrêa, N. K. (2024). Dynamic normativity: Necessary and sufficient conditions for value alignment. *arXiv preprint arXiv:2406.11039*.

Corrêa, N. K., Sen, A., Falk, S., & Fatimah, S. (2024). Tucano: Advancing Neural Text Generation for Portuguese. *arXiv preprint arXiv:2411.07854*.

Denning, J. M., Snefjella, B., & Blank, I. A. (2025). Do Large Language Models know who did what to whom?. *arXiv preprint arXiv:2504.16884*.

Drost, H. G. (2018). Philentropy: information theory and distance quantification with R. *Journal of Open Source Software*, *3*(26), 765.

Frade, S., Pinheiro, A. P., Santi, A., & Raposo, A. (2022). Is second best good enough? An EEG study on the effects of word expectancy in sentence comprehension. *Language, Cognition and Neuroscience*, *37*(2), 209-223.

Günther, F., Cassani, G., & Günther, F. (2025). Large Language Models in psycholinguistic studies.

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2022). A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*.

Hu, J., Mahowald, K., Lupyan, G., Ivanova, A., & Levy, R. (2024). Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, *121*(36), e2400917121.

Jacobs, C. L., Hubbard, R. J., & Federmeier, K. D. (2022, February). Masked language models directly encode linguistic uncertainty. In *Proceedings of the Society for Computation in Linguistics 2022* (pp. 225-228).

Jacobs, C. L., Grobol, L., & Tsang, A. (2024). Large-scale cloze evaluation reveals that token prediction tasks are neither lexically nor semantically aligned. *arXiv preprint arXiv:2410.12057*.

Katzir, R. (2023). Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. *Biolinguistics*, *17*, 1-12.

Leivada, E., Dentella, V., & Günther, F. (2024). Evaluating the language abilities of Large Language Models vs. humans: Three caveats. *Biolinguistics*, *18*, 1-12.

Lopes Rego, A. T., Snell, J., & Meeter, M. (2024). Language models outperform cloze predictability in a cognitive model of reading. *PLOS Computational Biology*, *20*(9), e1012117.

Mitra, A., Khanpour, H., Rosset, C., & Awadallah, A. (2024). Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*.

Peelle, J. E., Miller, R. L., Rogers, C. S., Spehar, B., Sommers, M. S., & Van Engen, K. J. (2020). Completion norms for 3085 English sentence contexts. *Behavior Research Methods*, *52*, 1795-1799.

R Core Team. (2000). R language definition. *Vienna, Austria: R foundation for statistical computing*, *3*(1), 116.

Santos, R., Silva, J., Gomes, L., Rodrigues, J., & Branco, A. (2024). Advancing Generative AI for Portuguese with Open Decoder Gervásio PT. *arXiv preprint arXiv:2402.18766*.

Snell, J., van Leipsig, S., Grainger, J., & Meeter, M. (2018). OB1-reader: A model of word recognition and eye movements in text reading. *Psychological review*, *125*(6), 969.

Wissler, C. (1905). The Spearman correlation formula. *Science*, *22*(558), 309-311.