Modeling cognitive processes of natural reading with transformer-based Language Models

Bruno Bianchi[1,2,*]
bbianchi@dc.uba.ar

Fermín Travi[1,2]
ftravi@dc.uba.ar

Juan E. Kamienkowski[1,2,3]
jkamienk@dc.uba.ar

[1]Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación
Buenos Aires, Argentina.

[2]Laboratorio de Inteligencia Artificial Aplicada
CONICET-Universidad de Buenos Aires.
Instituto de Ciencias de la Computación (ICC).
Buenos Aires, Argentina.

[3]Maestría en Explotación de Datos y
Descubrimiento del Conocimiento
Universidad de Buenos Aires
Buenos Aires, Argentina.

Abstract

Recent advances in Natural Language Processing (NLP) have led to the development of highly sophisticated language models for text generation. In parallel, neuroscience has increasingly employed these models to explore cognitive processes involved in language comprehension. Previous research has shown that models such as N-grams and LSTM networks can partially account for predictability effects in explaining eye movement behaviors, specifically Gaze Duration, during reading. In this study, we extend these findings by evaluating transformer-based models (GPT2, LLaMA-7B, and LLaMA2-7B) to further investigate this relationship. Our results indicate that these architectures outperform earlier models in explaining the variance in Gaze Durations recorded from Rioplantense Spanish readers. However, similar to previous studies, these models still fail to account for the entirety of the variance captured by human predictability. These findings suggest that, despite their advancements, state-of-the-art language models continue to predict language in ways that differ from human readers.

1 Introduction

Language is one of the defining characteristics of human beings. This unique capability allows us to communicate in complex ways to express thoughts [7]. Throughout history, various scientific fields—such as linguistics, psychology, and neuroscience—have studied language from different perspectives. In recent years, the field of Natural Language Processing (NLP) has gradually advanced toward algorithms and models capable of replicating human language with remarkable fidelity [18, 20, 21, 4]. Although the primary aim of this field is not to understand language itself but rather to simulate it computationally, state-of-the-art models provide us with valuable tools that may contribute to understanding language in the brain. These tools complement ongoing efforts in cognitive neuroscience, which aims to elucidate the neural mechanisms underlying language. Ultimately,

[*] https://liaa.dc.uba.ar/bruno-bianchi-en/

Latinx in AI @ NeurIPS-24 (LXAI 2024).

this creates a virtuous cycle where insights from both language and the brain's mechanisms drive advancements in algorithm development [26].

In this context, researchers have examined how various algorithms from the field of NLP can estimate the human word-predictability. This variable reflects how likely it is that a human reader can anticipate a specific word before reading it, and it is widely known to have an impact on how words are processed. Predictability is typically measured using a task called the cloze-task, where participants are asked to complete sentences with a single word [25]. Predictability (cloze-Pred henceforth) is then calculated as the frequency with which participants choose the word that originally continues the sentence.

Utilizing this variable to analyze various measurements of human cognition reveals its significant impact. For instance, studies measuring electroencephalography (EEG) signals during reading demonstrate that the brain's event-related potentials elicited by word processing (whether within sentences or in isolation) exhibit a marked correlation with the cloze-Pred of the presented words around 400 ms after word onset [16]. This effect, known as the N400 (named for its negative polarity and 400 ms latency), has been extensively investigated over the past four decades [15]. Broadly speaking, the N400 reflects the cognitive effort required for processing the word read. Nonetheless, despite substantial efforts dedicated to this research over recent decades, its underlying causes are still widely discussed.

In a related but often independent field of study, eye movements are analyzed to gain deeper insights into cognitive processes during reading. In this domain, eye movements serve as an indirect yet reliable measure of the timing of cognitive processes. It is proposed, for instance, that the duration for which the eyes fixate on a word (which can vary) correlates with the time required by the brain to process that word [11]. Consequently, a substantial body of research in neurolinguistics focuses on reading studies that record fixation durations [22, 13, 14, 5]. These durations are subsequently employed in statistical models to analyze the variables influencing them. Such analyses, conducted in various forms over recent decades, have elucidated that certain properties of words, such as their length or lexical frequency, significantly impact fixation duration. From these correlations, it is hypothesized that length and lexical frequency influence word processing. For example, the observation that more frequent words are fixated for shorter durations suggests they are easier to process, as individuals likely possess a mental lexicon where more frequently encountered words are more readily accessible [14].

Regarding cloze-Pred, it has been widely demonstrated that it negatively correlates with fixation time [14, 12]. In other words, more predictable words are fixated for shorter durations. This, in turn, has led to hypotheses about how predictions are generated in the brain and how these predictions facilitate the anticipation and acceleration of processing future stimuli, a concept that could extend to cognition beyond the realm of reading.

However, due to the costs associated with obtaining reliable measurements of cloze-Pred, studies often analyze the same texts repeatedly (for which measurements have already been made). Consequently, the development of a model capable of generating computational Predictabilities similar to human responses could represent a significant advancement in the study of language processing in the brain. Additionally, studies aimed at developing such a model would provide deeper insights into the underlying mechanisms of predictions generated by these models [3, 28, 26].

In recent decades, numerous studies have attempted to model cloze-Pred using computational models, referred to as computational predictabilities (comp-Pred). However, these efforts have achieved only partial success to date [19, 8, 3, 9, 1, 28, 23, 17]. Recently, there has been a growing interest in employing deep learning models for this purpose. Hofmann and colleagues [8, 9] trained N-grams, Topic Models (LDA), and Recurrent Neural Networks using datasets from Wikipedia texts and movie subtitles. Their analysis of variance in fixation times attributed to these probabilities led to the conclusion that computational models can better account for eye movements than cloze-Pred. Furthermore, Shain et al. [23] explored the relationship between eye movement metrics and predictions from various statistical language models (Ngrams, GPT2, GPT3, amoung others), revealing a logarithmic connection. Adopting a different methodology, Lopes Rego et al. [17] utilized cloze-Pred and comp-Pred generated by GPT2 and LLaMA to simulate eye movement metrics using the OB1-reader model [24]. They found that the simulated metrics produced with comp-Preds outperformed those generated with cloze-Pred in explaining human eye movements.

Another approach to analyze this was taken by Bianchi et al. [3] and Umfurer et al. [28]. In these studies they did not directly analyze the variance explained by each type of predictability (Cloze or computational), as this approach could overlook the possibility that different variables capture distinct portions of the variability in eye movement data. Notably, Bianchi et al. [3] observed that, unlike cloze-Pred, all the computational predictability measures (models used: N-grams, LSA, word2vec, FastText, AWD-LSTM) captured a portion of the variance originally explained by word lexical frequency. This effect was slightly reduced when using an LSTM-based language model trained on Spanish Wikipedia and fine-tuned on texts from a similar domain to the evaluation materials (narrative texts). Additionally, their findings also show that the residuals of the used statistical models(linear mixed models) contain variance explained by cloze-Pred. That is, the variance accounted for by the analyzed comp-Pred is not the same than the cloze-Pred.

We are currently in an era where transformer-based architectures dominate the field of NLP [29, 20, 6]. The present work aims to extend our previous efforts, first by utilizing transformer-based architectures, and second by refining the training corpus. Specifically, we aim to generate comp-Preds using a GPT2 model trained in Spanish, fine-tuned with two specialized corpora: one from the same literary domain as the evaluation texts, and the other from the same regional variant of Spanish as the participants (Rioplatense Spanish). Additionally, we are also testing predictions from more modern Large Language Models, like Llama and Llama2 in their 7B versions.

2 Methods

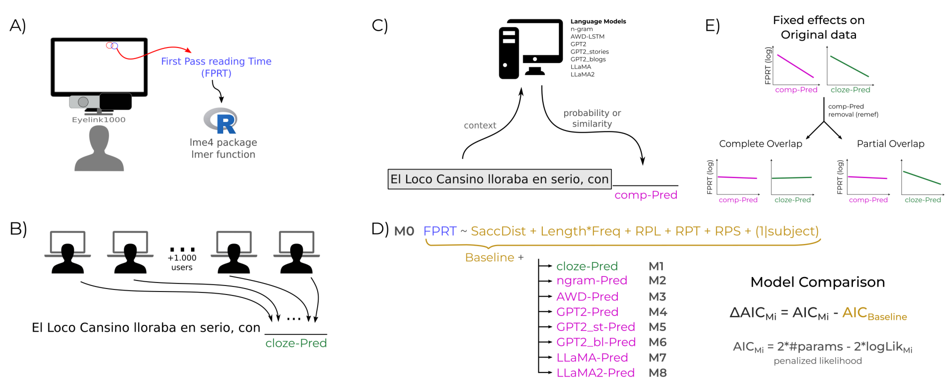

Figure 1: **Experimental designs:** **(A)** Eye movements were recorded in native Spanish readers reading 8 narrative stories. The eye movement measure (FPRT) was used as dependent variable in statistical models in R. **(B)** cloze-Pred for texts used in (A) was estimated from online responses by +1,000 participants in a cloze-task experiment. **(C)** comp-Pred for the same texts was estimated with several Language Models. **(D)** Statistical models ran in the present study. cloze- and comp-Preds were individually added to a baseline model. Model comparison was performed using AIC difference. **(E)** Diagram of expected results for cloze and comp-Pred effects after the Remef analysis.

2.1 Data: Texts, Cloze-Pred Estimation, and Eye Movements

In this study, we use the cloze-Pred and eye movement data previously published by Bianchi and colleagues [3]. Eye movements (specifically, fixation durations) were recorded from 36 participants while reading 8 narrative texts. From these recordings, First-Pass Reading Time (FPRT: the sum of all fixations on a word during its first encounter) was calculated for each word read by each participant (Figure 1A). This resulted in a total of $54,121$ data points, with an average of 1503 ± 618 per participant, 6765 ± 3226 per text, and 20 ± 35 fixations per word, covering $2,588$ unique words. FPRT will be used as the dependent variable in our statistical models. The corpus includes each of the 8 texts fully annotated with the following variables of interest: *Saccadic Distance:* the number of characters traversed by the eye before the current fixation; *Word Length:* the number of characters in the fixated word; *Frequency:* lexical frequency of the word based on LexEsp; *Rel pos:* relative positions in the line, text, or sentence; *Length:Freq:* interaction between Word Length and Frequency; *Pred:* Cloze and computational predictabilities (see below). For more details, see [3].

2.2 Human Predictability (cloze-Pred)

Data previously published by Bianchi and colleagues Bianchi et al. [3] includes cloze predictability (cloze-Pred) values for all words in the dataset. Predictions for each word were gathered through an online cloze experiment involving approximately 1,000 participants (Figure 1B). Each participant read at least one of the eight texts from the corpus, sequentially completing approximately one out of every thirty words. The cloze-Pred values were calculated based on an average of 13 responses per word, with a range of 8 to 37 responses.

2.3 Computational Predictability (comp-Pred)

To model cloze-Pred, we generated computational predictabilities (comp-Pred) using a Spanish version of GPT2 and two versions of LLaMA. These models were provided with the test texts to extract next-word probabilities corresponding to the completion of the original text (Figure 1C).

In this study, we introduced the GPT2 language model architecture and compared it to previous models used by Bianchi et al. [3] and Umfurer et al. [28]. The model used was trained by the *DeepEsp* consortium and is available through the *HuggingFace* repository.[2] This model was trained on 11.5GB of Spanish text, comprising Wikipedia (3.5GB) and books (8GB) from various domains (narrative, short stories, theater, poetry, essays, and popular science). This pre-trained model will be used for two independent fine-tunings:

1. For the first fine-tuning of this model, we used the corpus of narrative texts employed in previous studies [3]. Due to licensing restrictions, this corpus is not publicly accessible. The corpus contains 2,082 stories (600MB), with a wide range of literary genres and authors from different nationalities.
2. For the second fine-tuning texts sourced from Argentinian blogs were used (28MB). This corpus is being prepared for public release.

Both fine-tunings were performed on all parameters of the original model, using the code provided by HuggingFace for causal language model training.[3] Fine-tunings were performed using a Nvidia Titan RTX (VRAM 24Gb). The first fine-tuning took around 5 hours. The second one, less than an hour.

Finally, two models from the LLaMA [27] family were used. Specifically, we employed the LLaMA and LLaMA2 models, both in their 7-billion-parameter versions. The pretrained versions from the HuggingFace repository were used.[4] Contrary to GPT2, their pretraining corpus was multilingual.

2.4 Linear Mixed Models

We employed Linear Mixed Models (LMMs, *lme4* v.3.1-144, R v.3.6.3 [2]) to analyze the log-transformed First Pass Reading Time (FPRT) variable. This variable is proposed to capture effects related to the early processing of words [10]. Each of the fitted LMM returns a t-value for each co-variable, calculated by normalizing the estimated effect (slope) by its standard deviation (SD). Given the high number of data points at our disposal, we can consider infinite degrees of freedom. Thus, co-variables with $|t - value| > 2$ are considered to have a significant effect on the dependent variable. That is, under an assumed normal distribution, those estimated slopes that are more than 2 SD away from 0 have less than a 5% probability of being 0 (see [3] for more details). All the code necessary to run these statistical tests is openly available[5].

Our starting point is a baseline model (M0, Figure 1D) that includes variables typically used in the study of eye movements (saccadic distance; word length; frequency; rel pos; length:freq) [3, 13]. This model serves as a foundation for understanding the effects of both cloze and computational predictability (see [3, 28] for a discussion on these effects).

For comparing each resulting Linear Mixed Model (LMM) to the baseline model (M0), we use the Akaike Information Criterion (AIC). AIC evaluates the relative information loss of a model compared

[2]https://huggingface.co/DeepESP/gpt2-spanish

[3]https://github.com/huggingface/transformers/tree/main/examples/pytorch/language-modeling

[4]https://huggingface.co/meta-llama

[5]https://github.com/brunobian/Modeling-cognitive-processes-LMs-Neurips2024

to other models, with a penalty for adding co-variables; a model that loses less information (i.e., has a lower AIC) is considered to be of higher quality. It is important to note that AIC is a relative metric, not an absolute one, meaning it can only be used to compare nested models fitted to the exact same dataset. Therefore, we report the difference in AIC between each model and the baseline model (ΔAIC): the smaller this value, the better the model's performance.

Lastly, we reanalyzed all the models by adding cloze-Pred to its residuals. To achieve this, we extracted the residuals of each model by removing the estimated fixed effects. These residuals were then used in a new linear mixed model, with Cloze-Pred as the only fixed effect, keeping the same random effect structure. The aim of this analysis is to analyze the overlap between the variance explained by cloze-Pred and each comp-Pred. If a given comp-Pred can explain all of the variance of cloze-Pred (i.e. the computational model predicts like human readers), the effect of cloze-Pred on the residuals must be close to zero , (Figure 1E).

3 Results

Tables 1 and 2 presents the results (t-values) of the Linear Mixed Models (LMMs) conducted using the (log) First Pass Reading Time as the dependent variable against various combinations of co-variables. Models M0 through M3 (Table 1) correspond to models analyzed in previous work by Bianchi et al. [3] and Umfurer et al. [28]. M0 is considered the Baseline Model, which includes all co-variables unrelated to Predictability. All of them have been proved to have a significant effect on the FPRT, as shown by the t-values presented in the corresponding column.

Table 1: **Previous Linear Mixed Model Results:** Each column represents a different linear model. All models were applied to the same dataset. The results indicate the t-value obtained for each co-variable across the models. The row labeled "Cloze-Remef" refers to the t-value of the Cloze-Pred co-variable when used as the sole predictor in an LMM fitted on the residuals of the original model. All these models were previously presented by Bianchi et al. [3] and Umfurer et al. [28]. The row labeled "ΔAIC" refers to the difference between the AIC of each model and the AIC of M0.

Co-variable	**M0** Base	**M1** Cloze	**M2** Ngram	**M3** AWD epubs
Saccadic Dist.	44.44	44.63	43.92	44.31
Length (inv)	-18.15	-18.56	-19.10	-18.59
Frequency (log)	-10.83	-10.60	-1.93	-5.77
Rel pos line	4.14	3.96	4.33	4.12
Rel pos text	-3.93	-3.28	-3.87	-4.68
Rel pos sntc	-5.36	-4.85	-5.76	-4.97
Len:Freq	16.98	17.17	15.68	16.91
Pred (logit)	–	-16.23	-21.02	-18.23
ΔAIC	0	-254	-425	-274
Cloze-Remef	-16.14	0.00	-9.47	-7.37

M1 incorporates cloze-Pred, representing the variable we aim to approximate with comp-Pred. As expected, this co-variable shows a significant negative effect, indicating that more predictable words are fixated for shorter durations. Notably, the inclusion of cloze-Pred has minimal to no impact on the effects of the other variables, suggesting that it explains a portion of the FPRT variance that was not captured by any other co-variable in M0. Additionally, this model has a better fit to the data than M0, according to the difference of their AICs (ΔAIC $= -254$).

M2 represents the linear mixed model in which predictions from the best Ngram language model trained by Bianchi et al. [3] are added to the M0. As discussed by Umfurer et al. [28], this simple model, based solely on counting the frequency of word chains of length N, generates one of the most human-like predictabilities to date. In the corresponding column of the table, it can be observed that the effect of this comp-Pred is greater than that of cloze-Pred, and is also negative. However, the addition of this co-variable caused significant changes in other effects, primarily in the Frequency effect. This suggests that the Ngram predictions account for variance that was previously attributed to Lexical Frequency. This is likely because the Ngram model, which counts word occurrences, is highly correlated with word frequency. Regarding the overall M2 performance, the AIC difference with M0 shows that M2 outperforms M1. That is, the co-variables combinations of M2 better explain the variance of the dependent variable.

Finally, M3 represents the best result obtained in Umfurer et al. [28], where an AWD-LSTM network [18] was trained on Wikipedia and fine-tuned with a large corpus of narrative texts. Similar to the effect of Ngram comp-Pred, the AWD-LSTM comp-Pred effect is negative and larger than that of cloze-Pred. Additionally, it introduces significant, though smaller, changes to the other effects in the baseline model. In this case, the AIC analysis shows a slightly better fit to the data than M1, but worse than M2.

The "Cloze-Remef" row shows the t-value for cloze-Pred, obtained after fitting a new LMM on the residuals from each model, using only this co-variable. This approach allows us to analyze how much variance remaining in the residuals can be explained by the cloze-Pred. Trivially, in M0, the effect of cloze-Pred on the residuals is almost identical to its effect in M1. In contrast, in M1, there is no remaining variance in the residuals that can be explained by cloze-Pred. However, residuals of M2 and M3 contains some unexplained variance that can be accounted for by cloze-Pred, with the LSTM network's comp-Pred leaving the least unexplained variance. It is noteworthy how the overall performance of each LMM on the data, measured with AIC, does not reflect the similarity of the corresponding comp-Pred with cloze-Pred.

Models M4 through M8 (Table 2) correspond to novel models, where several versions of GPT2 and LLaMA Language Models are tested. M4 through M6 represent the statistical models corresponding to the predictabilities from the GPT2 variations. In M4, we introduced comp-Pred as computed with the pre-trained GPT2 model with no fine-tuning, downloaded directly from the HuggingFace repository. M5 and M6 introduce comp-Pred computed through the GPT2 model fine-tuned on short stories and Argentine blogs, respectively. Models M7 and M8 represent the statistical models corresponding to the predictabilities from LLaMA-7B and LLaMA2-7B models.

Table 2: **Novel Linear Mixed Model Results:** Each column represents a different linear model. All models were applied to the same dataset. The results indicate the t-value obtained for each co-variable across the models. The row labeled "Cloze-Remef" refers to the t-value of the Cloze-Pred co-variable when used as the sole predictor in an LMM fitted on the residuals of the original model. The row labeled "ΔAIC" refers to the difference between the AIC of each model and the AIC of the M0.

Co-variable	**M0** Base	**M4** GPT2	**M5** GPT2 stories	**M6** GPT2 blogs	**M7** Llama-7B	**M8** Llama2-7B
Saccadic Dist.	44.44	44.03	44.01	44.00	44.42	44.31
Length (inv)	-18.15	-20.04	-19.90	-19.90	-18.25	-19.93
Frequency (log)	-10.83	-4.98	-5.12	-5.01	-9.71	-6.42
Rel pos line	4.14	4.45	4.56	4.61	4.22	4.25
Rel pos text	-3.93	-4.56	-4.15	-4.25	-4.47	-4.76
Rel pos sntc	-5.36	-3.83	-3.81	-3.92	-4.47	-4.76
Len:Freq	16.98	15.59	15.58	15.61	17.20	16.43
Pred (logit)	–	-22.51	-22.52	-22.54	-9.17	-20.73
ΔAIC	0	-482	-482	-482	-81	-407
Cloze-Remef	-16.14	-6.18	-6.39	-6.58	-13.19	-6.12

All three models based on GPT2 (M4, M5, and M6) show similar t-values for comp-Pred (t-val$_{Pred,M4-M6}$ ~ -22) when compared to the Ngram-based model (M2) (t-val$_{Pred,M2}$ $= -21,02$). However, when analyzing the impact of these comp-Preds on the other co-variables (compared to M0), we observe smaller variations than in M2. This is particularly evident in the effect of lexical Frequency (t-val$_{Freq,M0}$ $= -10.83$). While in M2 lexical frequency loses significance (t-val$_{Freq,M2}$ $= -1.93$), in models M4 to M6 the t-values remain above the significance threshold (t-val$_{Freq,M4-M6}$ ~ -5.00). In this regard, the fine-tuned models seem to perform slightly better than the original model.

With respect to the residual variance analysis (Cloze-Remef), we observe that cloze-Pred is still able to capture enough variance to have a significant effect. Nevertheless, these remaining effects (t-val$_{Remef,M4-M6} \sim -6.40$) are smaller that the effects captured in the residuals of M2 (t-val$_{Remef,M2} = -9.47$) and M3 (t-val$_{Refemef,M3} = -7.37$).

When comparing the comp-Pred derived from LLaMA models (M7 and M8), clear differences emerge between the two versions. The first version (LLaMA-7B) exhibits a smaller comp-Pred effect (t-val$Pred, M7 = -9.17$), suggesting that its predictions are less aligned with cloze-Pred. Moreover, its influence on the Frequency effect is more subtle compared to the comp-Pred effects observed in previous models. Additionally, a significant amount of variance in this model's residuals can still be explained by cloze-Pred (t-val$remef, M7 = -13.19$), indicating incomplete overlap between these predictions.

In contrast, the second version (LLaMA2) shows a marked improvement. The comp-Pred effect is more comparable to previous models (t-val$Pred, M8 = -20.73$). However, this stronger effect comes with a notable reduction in the Frequency effect. Interestingly, LLaMA2, the most advanced model tested so far, leaves the least amount of unexplained variance for cloze-Pred (t-val$remef, M8 = -6.12$), demonstrating its superior ability to account for the variance in predictability.

When it comes to overall model performance, with the exception of M7, where the first version of LLaMA produced suboptimal results, all models exhibited comparable AIC values. Notably, the model incorporating LLaMA2-7B predictabilities, which were the most similar to cloze-Pred, demonstrated slightly lower performance compared to the models using GPT2-based predictabilities.

4 Discussion

Recent advances in Natural Language Processing (NLP) models have opened new avenues for their application in Cognitive Neuroscience, helping to shed light on the underlying processes of language comprehension. Previous studies have made significant strides in understanding how these models, particularly causal language models, can estimate the likelihood of a human reader knowing a word before encountering it (i.e., cloze Predictability, or cloze-Pred) as a co-variable in statistical models analyzing eye movements [3, 28, 9, 8, 17]. In this study, we extended this research by evaluating the performance of transformer-based architectures in this context. Importantly, this research examined eye movement data from native Spanish readers analyzing Spanish texts. Despite Spanish being one of the most spoken languages globally, research on reading processes in Spanish remains underrepresented in the psycholinguistic literature.

To this end, we analyzed the probabilities generated by the LLaMA-7B, LLaMA2-7B models, and three versions of the GPT2 model (a pre-trained version and two custom fine-tuned versions) as co-variables in the same linear mixed models used by Bianchi et al. [3]. These models use First Pass Reading Time (FPRT, or Gaze Duration) as the dependent variable, along with a variety of co-variables corresponding to each word. The different computational predictability (comp-Pred) values generated by these models were added one at a time to the statistical models, which were then compared with the baseline model without any predictability (M0). Additionally, we analyzed the residuals of each model to assess the remaining variance that could be explained by cloze-Pred.

Results obtained in the present study shows that the GPT2 performance was independent of whether the model was used as-is from the HuggingFace repository or whether it was fine-tuned with domain-specific texts or the dialect variant of Spanish spoken by the readers. Two points about the fine-tuning are important to highlight. First, the original GPT2 model's training data included narrative texts, so the fine-tuning may not have added much novel information to the model. Second, both fine-tunings were conducted with very small datasets (Rioplatense Spanish: 28MB, Narrative Stories: 600MB) compared to the original training corpus (11GB). Thus, it is possible that fine-tuning had little to no impact on the learned weights.

Regarding the use of LLaMA family models, both the first and second versions in their 7-billion-parameter (7B) variants, the results show a significant difference between the two versions. Notably, the first version (analyzed in M7), released in 2023, performs worse than the GPT2 model, which was published four years earlier. This is evident not only in the smaller effect size of the corresponding comp-Pred but also in the residual variance that can be explained by cloze-Pred. However, when analyzing the second version of the LLaMA model (analyzed in M8), slightly better results are observed compared to GPT2, both in terms of residual variance and reduced interference with the Lexical Frequency effect. This leads to the hypothesis that more modern models are not only generating predictions that are somewhat closer to those of human readers, but they are also achieving this by becoming less dependent on the frequency of word occurrences in the lexicon.

The improvement in comp-Preds from transformer-based models, compared to previously used models (particularly AWD-LSTM), highlights the advantages of this architecture, which appears to capture richer linguistic information. Transformers benefit from processing entire input sequences simultaneously, allowing them to retain information from distant words without loss. Moreover, the increasing complexity of these models in recent years has further enhanced their predictive accuracy. This is largely driven by the significant growth in the number of internal parameters. For instance, while the AWD-LSTM model we used contains approximately 10^7 parameters, both GPT2 and LLaMA models have 10^9 parameters, representing a two-order-of-magnitude difference. This increase in the number of parameters, combined with recent optimizations to the transformer architecture, allows for enhanced extraction of intrinsic textual information, leading to improved language modeling and prediction capabilities. These improvements are evident both in the models' commercial applications as AI assistants and in our empirical findings. Our results demonstrate that transformer-based models generate predictions that rely less on lexical frequency compared to previous models.

There are two main benefits to our analysis with respect to previous works [8, 9, 23, 17]: (1) the interference generated by comp-Pred in the other co-variables of the models, and (2) the capacity of cloze-Pred to capture a portion of the residual variance. These analyses are essential for gaining a deeper understanding of computational predictions. Examining how these predictions account for variance in Fixation Duration in isolation from other effects could lead to overly optimistic conclusions. For instance, if the focus is solely on the observation that the effects of GPT2 or LLaMA2 are greater than that of cloze-Pred, or even that the AIC of the fitted LMMs are better, it might be concluded that predictions from state-of-the-art language models outperform those from human readers. However, our analysis reveals that this improvement comes at the expense of the Frequency effect (which has been extensively studied in the field of neurolinguistics) and leaving part of the cloze-Pred variance unexplained. In future work, it would be interesting to extend this analysis to other languages and datasets to determine whether the effects found in our Spanish corpus are replicated in the data analyzed in the related literature.

In summary, this work contributes to the existing literature on predictions during reading, enhancing our understanding of both human predictability and computational predictabilities. Over the next few years, these two avenues must proceed hand in hand to foster synergy between the fields of cognitive neuroscience and artificial intelligence. This collaborative approach will enable us to combine resources and efforts to better understand the internal processes of both the human brain and state-of-the-art AI models, particularly in NLP. Furthermore, as advancements in AI continue to unfold, integrating insights from cognitive neuroscience could lead to the development of more sophisticated and human-like models. Ultimately, this intersection of disciplines holds the promise of unlocking new pathways for enhancing our comprehension of language processing, potentially paving the way for applications that not only improve AI performance but also offer deeper insights into human cognition.

5 Limitations

As previously mentioned, the corpora used for fine-tuning the GPT2 models are much smaller than the corpus used for pre-training. This may have resulted in the fine-tuning process not producing substantial changes in the pre-trained model's weights. However, it would be interesting to explore in greater depth whether it is possible to enrich these models with more information, both from the specific domain (narrative texts) and from the variant of Spanish spoken by the readers (Rioplatense Spanish). In the near future, we plan to expand the Rioplatense Spanish corpus to conduct a more thorough analysis of this type of fine-tuning. Additionally, we aim to further explore and diversify the analyses related to using these model outputs to improve our understanding of cognitive processes.

It is of great importance, both for the field of cognitive neuroscience and for artificial intelligence, that studies of this nature be conducted in as many languages and dialects as possible. Only in this way can we understand the general characteristics of our brain, while also ensuring equitable access to technological resources, such as today's generative NLP models.

References

[1] Abdullah Can Algan. Prediction of words in turkish sentences by lstm-based language modeling. Master's thesis, Middle East Technical University, 2021.

[2] Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, et al. lme4: Linear mixed-effects models using eigen and s4. *R package version*, 1(7):1–23, 2014.

[3] Bruno Bianchi, Gastón Bengolea Monzón, Luciana Ferrer, Diego Fernández Slezak, Diego E Shalom, and Juan E Kamienkowski. Human and computer estimations of predictability of words in written language. *Scientific reports*, 10(1):1–11, 2020.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Charles Clifton Jr, Fernanda Ferreira, John M Henderson, Albrecht W Inhoff, Simon P Liversedge, Erik D Reichle, and Elizabeth R Schotter. Eye movements in reading and information processing: Keith rayner's 40 year legacy. *Journal of Memory and Language*, 86:1–19, 2016.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. The faculty of language: what is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579, 2002.

[8] Markus J Hofmann, Chris Biemann, and Steffen Remus. Benchmarking n-grams, topic models and recurrent neural networks by cloze completions, eegs and eye movements. In *Cognitive approach to natural language processing*, pages 197–215. Elsevier, 2017. doi: https://doi.org/10.1016/B978-1-78548-253-3.50010-X.

[9] Markus J Hofmann, Steffen Remus, Chris Biemann, Ralph Radach, and Lars Kuchinke. Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, 4, 2021.

[10] Albrecht Werner Inhoff. Two stages of word processing during eye fixations in the reading of prose. *Journal of verbal learning and verbal behavior*, 23(5):612–624, 1984.

[11] Marcel A Just and Patricia A Carpenter. A theory of reading: from eye fixations to comprehension. *Psychological review*, 87(4):329, 1980.

[12] Juan E Kamienkowski, M Julia Carbajal, Bruno Bianchi, Mariano Sigman, and Diego E Shalom. Cumulative repetition effects across multiple readings of a word: Evidence from eye movements. *Discourse Processes*, 55(3):256–271, 2018.

[13] Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. Length, frequency, and predictability effects of words on eye movements in reading. *European journal of cognitive psychology*, 16(1-2):262–284, 2004.

[14] Reinhold Kliegl, Antje Nuthmann, and Ralf Engbert. Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *Journal of experimental psychology: General*, 135(1):12, 2006.

[15] Marta Kutas and Kara D Federmeier. Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, 62(1): 621–647, 2011.

[16] Marta Kutas and Steven A Hillyard. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205, 1980.

[17] Adrielli Tina Lopes Rego, Joshua Snell, and Martijn Meeter. Language models outperform cloze predictability in a cognitive model of reading. *PLOS Computational Biology*, 20(9): e1012117, 2024.

[18] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*, 2017. doi: https://arxiv.org/abs/1708.02182v1.

[19] James KY Ong and Reinhold Kliegl. Conditional co-occurrence probability acts like frequency in predicting fixation durations. *Journal of Eye Movement Research*, 2(1), 2008. doi: https://doi.org/10.16910/jemr.2.1.3.

[20] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *preprint*, 2018.

[21] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[22] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998.

[23] Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121, 2024.

[24] Joshua Snell, Sam van Leipsig, Jonathan Grainger, and Martijn Meeter. Ob1-reader: A model of word recognition and eye movements in text reading. *Psychological review*, 125(6):969, 2018.

[25] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433, 1953.

[26] Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, 32, 2019.

[27] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[28] Alfredo Umfurer, Juan E Kamienkowski, and Bruno Bianchi. Using lstm-based language models and human eye movements metrics to understand next-word predictions. In *XXII Simposio Argentino de Inteligencia artificial (ASSAI 2021)-JAIIO 50 (Modalidad virtual)*, 2021.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction accurately reflect the paper's contributions and scope by clearly outlining the objectives of comparing state-of-the-art language models in predicting eye fixation durations during reading. Additionally, the integration of approaches from artificial intelligence and cognitive neuroscience is consistently emphasized throughout the paper, reinforcing the relevance of the findings to both fields.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated Limitations section that addresses the constraints of the authors' work. For instance, it highlights the preference for using larger training datasets for fine-tuning to enhance model performance.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient information for reproducing the main experimental results that support its claims and conclusions. The eye movement data and cloze-Pred values are publicly available from the original authors, while the GPT2 and LLaMA models used are open access. Although the fine-tuning corpora are not published, the fine-tuned models did not significantly alter the results. Furthermore, some of this data will be published shortly.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All the data used in this study comes from openly available datasets that have been published for public use, ensuring proper credit to the original creators and compliance with their licensing terms.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The Methods section provides all the necessary information regarding the fine-tuning of GPT2. The LLaMA versions used in this study were not fine-tuned.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Yes, the Methods section includes comprehensive information about the statistical analysis, including a clear explanation of t-value interpretation and references for further discussions on the topic. This ensures that statistical significance is reported appropriately.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The Methods section includes detailed information about the computer resources utilized for fine-tuning the GPT2 model, ensuring that readers have sufficient details to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and conducted our research in accordance with all its guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This paper discusses the need to conduct such analyses in underrepresented languages, at least in the literature, such as Spanish. We believe that this could enhance our understanding of our culture and lead to the development of better NLP models.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Most of the data used in this work has already been published. The data we haven't released due to rights restrictions consists of widely disseminated texts that do not contain sensitive material that could be misused.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All the data used in this study comes from openly available datasets that have been published for public use, ensuring proper credit to the original creators and compliance with their licensing terms.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All the code utilized in this study is openly published on GitHub, ensuring thorough documentation is available alongside the assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The data used from these sources were not obtained specifically for this study but rather sourced from publicly available datasets.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The data used from these sources were not obtained specifically for this study but rather sourced from publicly available datasets.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.