Benchmarking Language Model Surprisal for Eye-Tracking Predictions in Brazilian Portuguese

Diego Alves
Saarland University
Saarbrücken, Germany
diego.alves@uni-saarland.de

Abstract

This study evaluates the effectiveness of surprisal estimates from six publicly available large language models (LLMs) in predicting reading times in Brazilian Portuguese (BP), using eye-tracking data from the RastrOS corpus. We analyze three key reading time measures: first fixation duration, gaze duration, and total fixation time. Our results demonstrate that surprisal significantly predicts all three measures, with a consistently linear effect observed across all models and the strongest effect for total fixation duration. We also find that larger model size does not necessarily provide better surprisal estimates. Additionally, entropy reduction derived from Cloze norms adds minimal predictive value beyond surprisal, and only for first fixation duration. These findings replicate known surprisal effects in BP and provide novel insights into how different models and linguistic predictors influence reading time predictions.

1 Introduction

In recent years, the use of large language models (LLMs) has emerged as a productive approach in cognitive science and psycholinguistics to better understand human language processing (Hale (2001); Armeni et al. (2017); Wilcox et al. (2020)). These models provide estimates of word predictability, which can be formalised through information-theoretic measures as surprisal.

Surprisal quantifies the unexpectedness of a word given its preceding context. It has been shown that this measure correlates with reading time metrics obtained through eye-tracking experiments (Smith and Levy (2013); Hofmann et al. (2022); Demberg and Keller (2008)). These findings support theories claiming that human language comprehension is governed, at least to some extent, by the efficient processing of probabilistic information.

Despite extensive research on surprisal linked to cognitive processing, the focus has largely been on English, leaving cross-linguistic applicability underexplored. Wilcox et al. (2023a) showed how well language model surprisal can predict reading times in eleven languages, providing important information regarding cross-linguistic variability in the cognitive processing of language. However, Brazilian Portuguese (BP) was notably absent from this analysis, leaving a gap in our understanding of the role of surprisal in the processing of this language.

To address this gap, the present study focuses on BP, employing the RastrOS corpus, a large-scale eye-tracking dataset collected from students in higher education in Brazil, which also includes carefully constructed norms of predictability of words (Leal et al., 2022).

The aim of this study is to evaluate how surprisal values derived from a variety of publicly available LLMs predict three key eye-tracking reading time measures: first fixation duration, gaze duration, and total fixation time in Brazilian Portuguese. Furthermore, we investigate the role of entropy reduction as a contributing factor in modelling reading times. We also assess the linearity of the relationship between surprisal and reading times, determining whether linear models sufficiently capture this mapping or whether more complex patterns are present.

Our work not only provides missing data for Brazilian Portuguese but also identifies the most effective publicly available LLMs for modelling human reading behaviour in this language. Moreover, it offers a baseline for researchers aiming to use surprisal to analyse linguistic phenomena in Brazilian Portuguese following information-theoretic principles such as the Uniform Information Density (UID) hypothesis (Jaeger and Levy, 2006).

The remainder of the paper is organised as fol-

Proceedings of First International Workshop on Gaze Data and NLP associated with RANLP 2025,
pages 7–17, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-104-2-002

lows. Section 2 reviews related work on using LLMs to model reading times. Section 3 presents the dataset, describes the large language models tested, and explains our evaluation methods. Section 4 then presents the results. We conclude with a summary of our findings and directions for future work in Section 5, followed by a discussion of the study's limitations in Section 6.

2 Related Work

Wilcox et al. (2023a) examined surprisal's relationship to reading times in eleven languages across five language families. Using monolingual and multilingual transformer-based language models (trained on the Wiki40B dataset, Guo et al. (2020), and mGPT, Shliazhko et al. (2024)), they showed that both surprisal and contextual entropy predict reading times, and that the relationship between surprisal and reading time is linear.

This linear relationship was also supported by Xu et al. (2023), who analysed seven languages and found evidence of superlinear effects in some cases, with results highly dependent on the language model used to estimate surprisal.

Additionally, Wilcox et al. (2023b) tested the quality–power (QP) hypothesis, which posits that higher-quality language models (LMs) better predict human reading behaviour. By training LMs on 13 languages with varying amounts of training data (from 1 million to 1 billion tokens), they found that, in most cases, models trained on more tokens showed stronger predictive power for eye-tracking data, supporting the QP hypothesis within the tested range.

Lin and Schuler (2025) proposed a neural study to complement these observations regarding reading time. By evaluating surprisal estimates from 17 Transformer models across three language families using fMRI data, they showed that the positive relationship between model perplexity and predictive power also generalizes to neural measures.

However, regarding LLMs, Oh and Schuler (2023) demonstrated that despite having better perplexity, larger models predict human reading times less accurately. Specifically, they tend to underpredict reading times for named entities and overpredict for function words, suggesting that memorization in these models reduces their alignment with human processing.

This tendency is also observed by Liu et al. (2023) who examined the effect of temperature scaling on large language model surprisal estimates and their fit to English reading time data, showing that calibration improves with model size, and temperature scaling significantly enhances prediction.

Moreover, Nair and Resnik (2023) demonstrated that while surprisal theory explains how a word's predictability influences processing difficulty via probabilistic updating, it does not fully capture all aspects of incremental processing, such as effects from low-frequency words and garden-path disambiguation. To address these limitations, Wang et al. (2025) developed a model that integrates syntactic information with statistical surprisal estimated from LLMs, resulting in significantly higher correlations with human reading times than surprisal alone.

Therefore, although surprisal alone cannot fully account for cognitive language processing, it has a significant impact across many languages. Additionally, both the size of the language model and the amount and quality of training data affect the relationship between reading time and word predictability. Consequently, it is crucial to identify the best language model for each language (and language variety) before applying surprisal estimates in various research fields.

3 Methodology

3.1 Eye-Tracking Data

The RastrOS corpus was developed to support psycholinguistic research on Brazilian Portuguese (BP), particularly focusing on lexical predictability and sentence processing. It comprises two main components: predictability norms collected via a Cloze test and eye-tracking data gathered from reading tasks.

A total of 393 native BP speakers from six Brazilian universities participated in the Cloze test, primarily undergraduate students. Each participant completed Cloze tasks on five randomly selected paragraphs, balanced across three genres: journalistic (40%), literary (20%), and popular science (40%).

The Cloze corpus includes 50 paragraphs, comprising 120 sentences and 2,494 words (2,831 tokens). Source texts were drawn from the Lácio-Web corpus (Aluísio et al., 2004), public domain literature, and contemporary online texts.

Participant responses were compared to target words based on orthographic match, morphosyntactic class (PoS), and inflection, with semantic simi-

larity assessed via word embeddings. The dataset is annotated with PoS tags (using the Palavras parser; Bick 2000), word frequency (from Corpus Brasileiro (Sardinha, 2010) and BrWaC (Wagner Filho et al., 2018)), and includes surprisal and entropy reduction values derived from the Cloze test results.

The eye-tracking data of the RastrOS were collected from 37 undergraduate students and were recorded using the EyeLink 1000 eye-tracker at a sampling rate of 1000 Hz.

Participants read 120 sentences taken from the same 50-paragraph Cloze corpus, a total of 2,494 words total (2,831 tokens including punctuation). Each sentence is annotated with 36 eye-tracking metrics (e.g., first fixation duration, gaze duration, and total fixation time).

3.2 Large Language Models

For our analysis, we selected six publicly available large language models that vary in the number of parameters and the training data used:

1. Bloom-560m[1] (Workshop, 2022) - Multilingual model trained on 1.5 TB of pre-processed text, of which 11.1% is Portuguese. 559 million parameters distributed over 24 layers with 16 attention heads and 1024-dimensional hidden states.

2. Bloomz-7b1[2] (Muennighoff et al., 2022) - Same training corpus as bloom-560m but with 7 billion parameters over 30 layers with 32 attention heads, and 4096-dimensional hidden states. bloomz is a fine-tuned version of bloom, trained with multitask instructions to improve zero-shot performance.

3. Llama-2-7B-hf[3] – Pretrained on 2 trillion tokens from public sources, then fine-tuned with public instruction datasets and over one million human-annotated examples. It has 1024 hidden dimensions with 32 attention heads over 32 layers (Wang et al., 2023). Evaluation tests were performed only in English.

4. Llama-3-2-1B[4] – 1 billion parameter model, pretrained on up to 9 trillion tokens of data in 8 different languages (including Portuguese) from publicly available sources.

5. Llama-3-2-3B[5] – Same training data as llama-3-2-1B but with 3 billion parameters,

6. Mistral-7b[6] – Trained on a mix of web data and code, with 7 billion parameters. It has 32 layers, 32 attention heads, and a hidden size of 4096 dimensions. The model evaluation was conducted exclusively on English (Jiang et al., 2023).

With this selection, our aim is to provide a meaningful comparison between language models of different sizes and training objectives, including models fine-tuned for specific tasks (e.g., bloomz-7b1), and models primarily focused on English (e.g., llama-2-7B-hf and mistral-7B), despite being trained on multilingual data. Unfortunately, only the BLOOM models provide sufficient information about the proportion of Portuguese in their training data, although they do not specify which variety of Portuguese was used.

3.3 Evaluation Methods

To evaluate the performance of the different language models in predicting reading times, we adopted the methodology proposed by Wilcox et al. (2023a).

Thus, although the RastrOS corpus provides 36 different word-based measures of reading time, we focus on three commonly used metrics (Rayner, 1998):

1. First fixation duration - the duration of the first fixation on a word during its first pass. Annotated as IA_FIRST_FIXATION_DURATION in RastrOS.

2. Gaze duration - the sum of all first-pass fixations on a word. IA_FIRST_RUN_DWELL_TIME in RastrOS.

3. Total fixation duration - the sum of all fixations on a word during the trial. IA_DWELL_TIME in RastrOS.

First fixation reflects the initial processing of a word and is associated with early stages of lexical

[1] https://huggingface.co/bigscience/bloom-560m

[2] bigscience/bloomz-7b1

[3] https://huggingface.co/meta-llama/Llama-2-7b-hf

[4] https://huggingface.co/meta-llama/Llama-3.2-1B

[5] https://huggingface.co/meta-llama/Llama-3.2-3B

[6] https://huggingface.co/mistralai/Mistral-7B-v0.1

access. Gaze duration captures the time spent on a word during first-pass reading and is sensitive to lexical and syntactic processing. Total fixation time includes any regressions back to the word and reflects later stages of comprehension, such as reanalysis or integration difficulties (Rayner, 1998).

3.3.1 Surprisal

To compute word-level surprisal values, we used the surprisal Python library[7]. Sentences from RastrOS were first recomposed and loaded in their original order. Using the AutoHuggingFaceModel interface provided by the library, we instantiated each selected model and computed token-level surprisal values for each recomposed sentence. As a post-processing step, we merged the subword tokens produced by the language models to reconstruct the original words for analysis.

The regression models follow the structure proposed by Wilcox et al. (2023a), aiming to predict the reading time $y(w_t, w_{<t})$ of a word w_t given its preceding context $w_{<t}$. The predictor vector x_t includes not only information about the target word w_t, but also features from the two preceding words w_{t-1} and w_{t-2}, in order to account for potential spillover effects on reading time.

Each model includes baseline predictors such as word length and log unigram frequency (corresponding to Word_Length and Freq_brWaC_log in RastrOS) for the target word and the two preceding words. These features form the baseline structure of the predictor vector x_t at position t.

We used linear mixed-effects regression models, implemented via the lmer() function from the lme4 R package (Bates et al., 2015).

To measure how much surprisal improves model performance, we compare the baseline model (Appendix A, equation 1) to models that include surprisal values, specifically the surprisal of the target word and its two preceding words as estimated by the LLM (Appendix A, equation 2). The delta is defined as the difference in per-word log-likelihood between the surprisal-enhanced model and the baseline: a positive delta indicates that surprisal helps the model better explain that word's reading time. By aggregating these deltas across all words, we assess whether incorporating surprisal significantly improves prediction accuracy.

Additionally, all regression models are trained and evaluated using 10-fold cross-validation. To assess the significance of the observed differences (Δ) between target and baseline models, we use a paired permutation test. This non-parametric test evaluates whether Δ significantly differs from zero and whether different models differ from each other, without assuming any specific distribution of the test statistic. p-values are computed based on the empirical distribution of likelihood differences, estimated by averaging over permutations of the likelihood values.

For each reading time measure, we compared the Δ values obtained using surprisal estimates from the LLMs listed in Subsection 3.2.

3.3.2 Entropy Reduction

Wilcox et al. (2023a) tested the influence of contextual entropy as a predictor, comparing it to a baseline model that included the features from the baseline structure combined with surprisal values.

Rather than using contextual entropy, we employed entropy reduction values from the RastrOS corpus (Entropy_Reduction), derived from Cloze test results. Lowder et al. (2018) demonstrated that entropy reduction significantly predicts reading time. This is limitation of this approach when compared to entropy estimates generated by a language model trained on a large corpus. Nevertheless, we decided to use the available entropy reduction values provided by the corpus to have at least an estimation of the effect.

Thus, using baselines that include surprisal values, we compared models for each reading measure and LLM by adding entropy reduction as a predictor, considering the target token and the two preceding tokens for both surprisal and entropy reduction, with Δ and the statistical tests as described in 3.3.1. The model including entropy reduction is described in Appendix B.

3.3.3 Linearity

For the analysis of surprisal and entropy reduction as predictors, we used regression models that assume a linear relationship between surprisal and reading time as supported by previous studies (e.g., Smith and Levy (2013); Wilcox et al. (2020); and Shain et al. (2024)). However, as recent work has challenged this assumption, proposing superlinear (e.g., Meister et al. (2021) and Hoover et al. (2022)) or sublinear (Hoover et al., 2023) links, we decided, following Wilcox et al. (2023a) to test this by comparing the performance Δ of our linear models with models capable of capturing non-linear rela-

[7] https://pypi.org/project/surprisal/

tionships.

To analyse linearity, we used generalized additive models (GAMs), which flexibly capture potential non-linear effects.

If the GAM fits a visually linear pattern, this supports the hypothesis of a linear link. We modelled reading times based on Freq_brWaC_log, Word_Length, and surprisal from sentence-level. Our GAMs included smooth terms for current and previous word surprisal and tensor product terms to model non-linear interactions between log-frequency and word length, following the method applied by Wilcox et al. (2023a).

Thus, we compared generalized additive models (GAMs) that model surprisal effects on reading time either as linear terms or as flexible non-linear smooth functions, alongside a baseline model without surprisal, as described in Appendix C. Using 10-fold cross-validation, we calculated the prediction error (RMSE) for each model on held-out data and computed the improvement Δ over the baseline (without surprisal) for both linear and non-linear surprisal models. We then tested the significance of these improvements and differences between linear and non-linear models using paired permutation tests.

4 Results

4.1 Suprisal Models Compared to Baseline

Figure 1 presents the mean Δ log-likelihood per word for each LLM across all three reading time measures, shown as separate panels.

Regarding the different reading time measures, surprisal shows the highest predictive power for total fixation time duration, followed by gaze duration, and finally the lowest Δ values for first fixation duration. These results align with those reported by Wilcox et al. (2023a), showing similar magnitudes of Δ across reading measure conditions.

The analysis of the models concerning first fixation duration shows that the Δ values are approximately 0.0025. All Δ values are significantly different from 0 ($p < 0.001$). The pairwise comparison of the different models shows that there are no statistically significant differences among them.

Regarding the gaze duration, there is a higher variation in Δ values, with larger standard error bars. Statistical tests show that, for all models, Δ differs significantly from 0, except for bloomz-7b1 ($p = 0.0014$). When comparing the different language models, the statistical tests indicate that the models have significantly different Δ values, except for:

- llama-2-7b similar to bloom-560, llama-3-2-1B, and llama-3-2-3B
- llama-3-2-3B similar to llama-3-2-1B
- mistral_7B similar to llama-3-2-3B and llama-2-7b

Finally, when considering total fixation time, except for bloomz-7b1, Δ values are around 0.05 and are all significantly different from 0. Also, all models differ in the pairwise comparison, except for:

- mistral_7B which is similar to llama-2-7b, llama-3-2-1B, and llama-3-2-3B
- llama-2-7b, similar to llama-3-2-1B

These results show that the best models are not necessarily those with the highest number of parameters, as for gaze duration, statistically similar results were obtained for models with 560 million, 1, 3, and 7 billion parameters. This effect is even more pronounced when considering total fixation time, with some statistically similar results observed for models with 1, 3, and 7 billion parameters.

The overall analysis of Figure 1 indicates that the best model—considering both gaze and total fixation durations—is llama-3-2-3B. Moreover, it is notable that the fine-tuned model bloomz-7b1, despite having 7 billion parameters, performs the worst in predicting reading times. Additionally, although not evaluated in languages other than English, mistral_7B shows statistically similar Δ values when compared to llama-3-2-3B.

The estimated effects of surprisal, including coefficients and standard errors for each model, are presented in Table 1.

We observe some consistency among the models with the best Δ values. The most discrepant model is bloom-7B1, reflecting its poor performance. Other predictors also show significant effects, except for the log frequency of the second word preceding the target, which was statistically significant only for bloom-7B1.

4.2 Entropy Reduction Models Compared to Surprisal Baseline

The Δ results comparing baseline models (with surprisal) to models that include both entropy reduction and surprisal are presented in Figure 2.

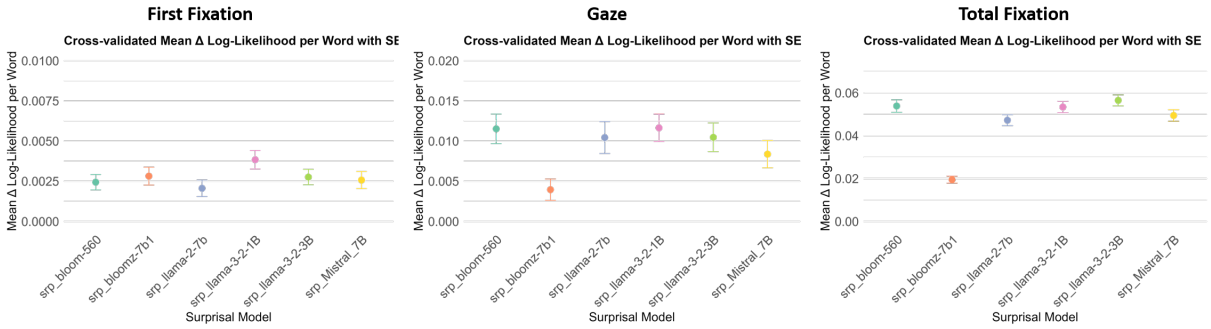

Figure 1: Predictive power of surprisal across reading time measures and LLMs. Dots indicate mean Δ log-likelihood per word; error bars show ±1 standard error of the mean. Note that each panel uses a different y-axis scale.

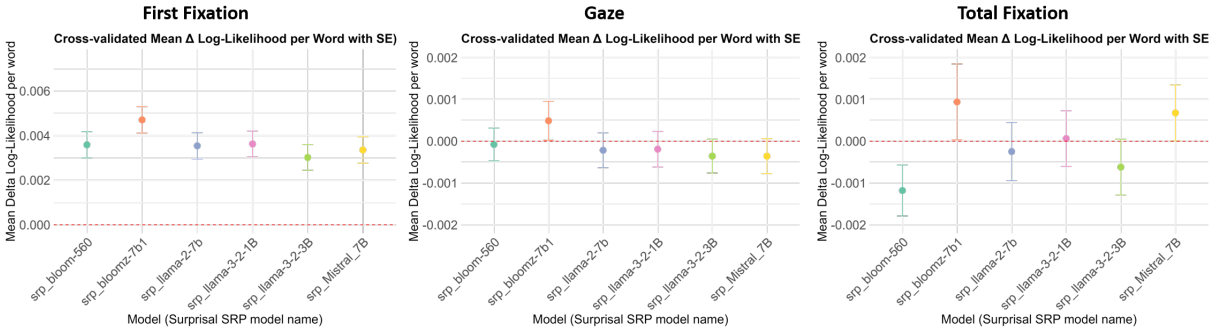

Figure 2: Predictive power of entropy reduction across reading time measures and LLMs. Dots indicate mean Δ log-likelihood per word; error bars show ±1 standard error of the mean. Note that each panel uses a different y-axis scale.

Model	Effect	Std. Error
bloom-560	9.52	0.15
bloom-7B1	5.01	0.13
llama-2-7B	8.16	0.14
llama-3-2-1B	8.57	0.15
llama-3-2-3B	8.61	0.15
mistral-7B	8.37	0.15

Table 1: Surprisal coefficient for the target word in a linear model including surprisal, frequency, and word length as predictors (considering target word and the two previous ones).

The statistical analyses show that, for first fixation duration, all Δ values are significantly different from 0, although the models perform similarly. In contrast, for both gaze duration and total fixation duration, all Δ values are close to 0, and no significant differences between models were observed.

Wilcox et al. (2023a) observed an improvement in the prediction of gaze duration when adding contextual entropy as a predictor, with a weak—albeit consistent—effect across languages. In our study, however, we do not observe the same effect. Including entropy reduction appears to have a positive impact (independent of the language model) only for first fixation duration, and even then, the Δ values are low (around 0.004).

4.3 Linearity analysis

Figure 3 presents the results of comparing the Δ obtained from a linear GAM model with surprisal to a baseline model without surprisal, as well as the corresponding Δ values from a non-linear model, for the prediction of total fixation duration.

As expected from the results presented in Section 3.3.1, the statistical tests showed that all Δ are significantly different from 0. Moreover, when comparing the linear Δ with the non-linear one for each language model, we observe that the results are not significantly different.

Thus, these results corroborate the claim that the effect of surprisal on reading time is linear, consistent with the findings of Wilcox et al. (2020). This linear effect is observed across all LLMs considered, regardless of parameter size, training data, or supported languages.

Figure 4 shows the differences in surprisal ef-

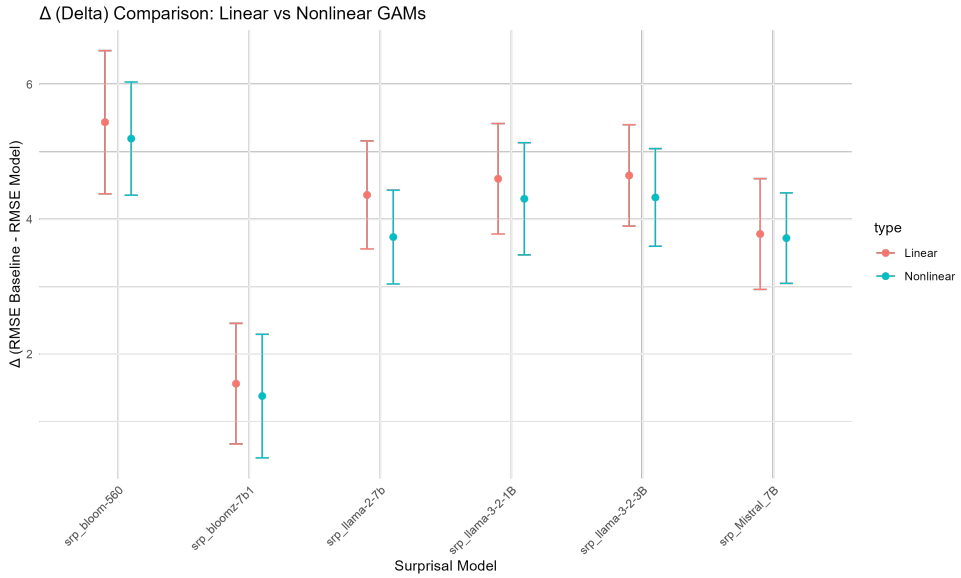

Figure 3: Comparison Between Linear and Non-linear Models for the prediction of total fixation duration. Dots indicate mean Δ log-likelihood per word; error bars represent the standard error of the mean delta RMSE across cross-validation folds..

fects between linear and non-linear models for each LLM. A linear fit can be observed, especially in the denser regions of surprisal values. Notably, the model bloom-7b-1, which showed the poorest Δ values when the effect of surprisal was analysed, also exhibits the greatest visual deviation from linearity in the non-linear model.

Similar results are observed for both first fixation and gaze durations, although the non-linear models exhibit substantially larger error bars for the first fixation measures.

4.4 Part-of-Speech Analysis

As a complementary analysis, we investigated the linearity of the relationship between surprisal estimates and eye-tracking measures across different parts of speech (PoS) in the RastrOS corpus.

To do this, we conducted ordinary least squares (OLS) linear regression analyses on data aggregated by PoS. Entries with erroneous PoS tags (i.e., "ERR", which appeared twice in the corpus) were excluded. For each PoS category, we computed the mean values of surprisal estimates from six language models, as well as mean fixation durations.

For each surprisal model, we then performed an OLS regression using SciPy's linregress function, obtaining the slope, intercept, coefficient of determination (R^2), p-value, and standard error.

Table 2 presents the slope, R^2, and p-value from the OLS regression for each language model across parts of speech.

Model	slope	R^2	p-value
bloom-560	71.88	0.86	<0.001
bloom-7B1	45.86	0.62	<0.001
llama-2-7B	38.82	0.53	<0.001
llama-3-2-1B	35.95	0.51	<0.001
llama-3-2-3B	47.91	0.57	<0.001
mistral-7B	65.05	0.71	<0.001

Table 2: Slope, R^2, and p-values from OLS regressions of surprisal estimates (per LLM) on total fixation duration aggregated by part of speech..

The LLM with the highest R^2 value is bloom-560, followed by mistral-7B. Indicating that for this aggregated analysis in terms of PoS, the smallest model gave the best results. However, in this case, we considered a simpler regression analysis, considering only the fixed effect of surprisal.

Figure 5 presents the linear regression plot obtained using bloom-560, showing the estimated mean total fixation time (i.e., IA_DWELL_TIME) as a function of the mean surprisal value for each part of speech (PoS) in RastrOS.

As expected, parts of speech typically associated with longer word forms and higher information load (e.g., verbs, nouns, and adjectives) exhibit higher values of both reading time and surprisal. In contrast, conjunctions, determiners, and pronouns show lower values, while auxiliary verbs are the least surprising and associated with the shortest reading times. The same tendency was observed

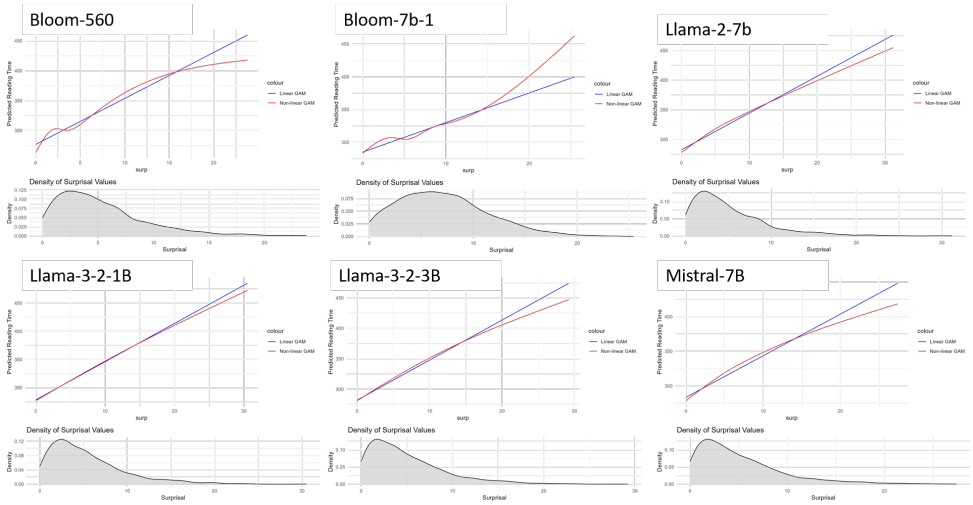

Figure 4: Surprisal versus reading time relationship: Non-linear GAMs are in red and linear control GAMs are in blue. Grey subplots indicate the distribution of surprisal values.

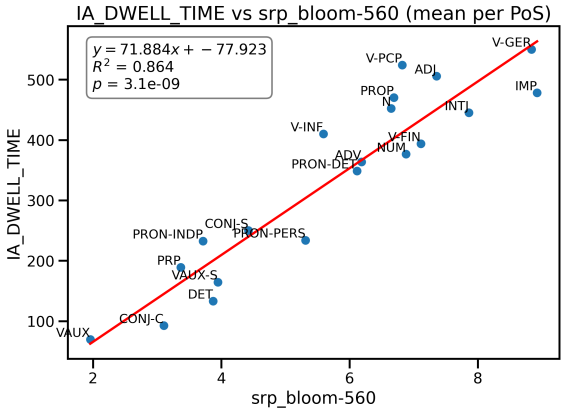

Figure 5: Linear regression plot of mean surprisal estimates against total fixation duration for each part-of-speech (PoS) category for bloom-560. Each point represents a PoS tag, labeled accordingly. Red lines indicate the best-fit regression line.

for all LLMs.

5 Conclusion

In this study, we examined the ability of surprisal estimates from six publicly available large language models (LLMs) to predict reading times in Brazilian Portuguese (BP), using eye-tracking data from the RastrOS corpus. Our findings confirm that surprisal significantly correlates with three key reading time measures (i.e., first fixation duration, gaze duration, and total fixation time) supporting the role of probabilistic predictability in BP processing.

The best-performing model, Llama-3-2-3B, appears to outperform others, including larger or fine-tuned models such as Bloomz-7b1, suggesting that model architecture and training data quality may be more important than sheer size. Moreover, the relationship between surprisal and reading times was consistently linear, aligning with previous findings. However, entropy reduction, calculated from Cloze norms, provided minimal additional predictive power.

These results extend surprisal-based research to BP and offer a baseline for model selection in future studies.

6 Limitations

Several limitations should be noted, first, the RastrOS corpus, though carefully constructed, is relatively small and genre-biased (e.g., dominated by journalistic texts), which may limit the generalizability of our findings.

Second, the language models tested were primarily trained on multilingual data with unclear proportions of Portuguese, and none were specifically optimized for BP. This raises questions about whether models trained exclusively on BP data might provide better fits.

Third, our entropy reduction analysis relied on Cloze norms rather than model-derived entropy, potentially underestimating its predictive power.

Finally, while we focused on surprisal as a key predictor, other linguistic factors, such as syntactic complexity, were not considered and may have an impact on reading time variance.

Acknowledgments

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Sandra M Aluísio, Gisele Montilha Pinheiro, Aline MP Manfrin, Leandro HM de Oliveira, Luiz C Genoves Jr, and Stella EO Tagnin. 2004. The lácio-web: Corpora and tools to advance brazilian portuguese language investigations and computational linguistic tools. In *LREC*.

Kristijan Armeni, Roel M Willems, and Stefan L Frank. 2017. Probabilistic language models in cognitive neuroscience: Promises and pitfalls. *Neuroscience & Biobehavioral Reviews*, 83:579–588.

Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Gabor Grothendieck, Peter Green, and Maintainer Ben Bolker. 2015. Package 'lme4'. *convergence*, 12(1):2.

Eckhard Bick. 2000. *The parsing system palavras: Automatic grammatical analysis of Portuguese in a constraint grammar famework*. Aarhus Universitetsforlag.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2440–2452, Marseille, France. European Language Resources Association.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

Markus J Hofmann, Steffen Remus, Chris Biemann, Ralph Radach, and Lars Kuchinke. 2022. Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, 4:730570.

Jacob Louis Hoover, Morgan Sonderegger, Steven T Piantadosi, and Timothy J O'Donnell. 2023. The plausibility of sampling as an algorithmic theory of sentence processing. *Open Mind*, 7:350–391.

JL Hoover, M Sonderegger, and TJ O'Donnell. 2022. With better language models, processing time is superlinear in surprisal (poster). york, england.

T. Jaeger and Roger Levy. 2006. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Sidney Evaldo Leal, Katerina Lukasova, Maria Teresa Carthery-Goulart, and Sandra Maria Aluísio. 2022. Rastros project: Natural language processing contributions to the development of an eye-tracking corpus with predictability norms for brazilian portuguese. *Language Resources and Evaluation*, 56(4):1333–1372.

Yi-Chien Lin and William Schuler. 2025. Surprisal from larger transformer-based language models predicts fmri data more poorly. *arXiv preprint arXiv:2506.11338*.

Tong Liu, Iza Škrjanec, and Vera Demberg. 2023. Temperature-scaling surprisal estimates improve fit to human reading times–but does it do so for the" right reasons"? *arXiv preprint arXiv:2311.09325*.

Matthew W Lowder, Wonil Choi, Fernanda Ferreira, and John M Henderson. 2018. Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive science*, 42:1166–1183.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. *arXiv preprint arXiv:2109.11635*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Sathvik Nair and Philip Resnik. 2023. Words, subwords, and morphemes: what really matters in the surprisal-reading time relationship? *arXiv preprint arXiv:2310.17774*.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

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

Tony Berber Sardinha. 2010. Corpus brasileiro. *Informática*, 708:0–1.

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

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mgpt: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Jorge A Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brwac corpus: A new open resource for brazilian portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Daphne P Wang, Mehrnoosh Sadrzadeh, Miloš Stanojević, Wing-Yee Chow, and Richard Breheny. 2025. Extracting structure from an llm-how to improve on surprisal-based models of human language processing. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4938–4944.

Tony T Wang, Miles Wang, Kaivalya Hariharan, and Nir Shavit. 2023. Forbidden facts: An investigation of competing objectives in llama-2. *arXiv preprint arXiv:2312.08793*.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023a. Testing the predictions of surprisal theory in 11 languages. *arXiv preprint arXiv:2301.12345*.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

Ethan Gotlieb Wilcox, Clara Isabel Meister, Ryan Cotterell, and Tiago Pimentel. 2023b. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511. Association for Computational Linguistics.

BigScience Workshop. 2022. Bloom: Bigscience language open-science open-access multilingual language model. https://huggingface.co/bigscience/bloom. International collaboration, May 2021–May 2022.

Weijie Xu, Jason Chon, Tianran Liu, and Richard Futrell. 2023. The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721.

A Appendix A

For the tests assessing the effect of surprisal, we use the following models: (1) Baseline and (2) with surprisal.

$$\begin{aligned}\text{reading\_time} \sim\ & \text{Freq\_brWaC\_log} \\ & + \text{Word\_Length} \\ & + \text{prev\_freq} + \text{prev\_len} \\ & + \text{prev2\_freq} + \text{prev2\_len} \\ & + (1 \mid \text{SESSION\_LABEL})\end{aligned} \tag{1}$$

$$\begin{aligned}\text{reading\_time} \sim\ & \text{prev\_surp} \\ & + \text{prev2\_surp} \\ & + \text{Freq\_brWaC\_log} \\ & + \text{Word\_Length} \\ & + \text{prev\_freq} + \text{prev\_len} \\ & + \text{prev2\_freq} + \text{prev2\_len} \\ & + (1 \mid \text{SESSION\_LABEL})\end{aligned} \tag{2}$$

B Appendix B

For the tests assessing the effect of entropy reduction, we use the model 2 in Appendix A as baseline and (3) with entropy.

$$\begin{aligned}\text{reading\_time} \sim\ & \text{prev\_surp} \\ & + \text{prev2\_surp} \\ & + \text{entropy\_Reduction} \\ & + \text{prev\_entropy} \\ & + \text{prev2\_entropy} \\ & + \text{Freq\_brWaC\_log} \\ & + \text{Word\_Length} \\ & + \text{prev\_freq} + \text{prev\_len} \\ & + \text{prev2\_freq} + \text{prev2\_len} \\ & + (1 \mid \text{SESSION\_LABEL})\end{aligned} \tag{3}$$

C Appendix C

The GAM formula used for non-linear models we use is:

$$\begin{aligned}\text{reading\_time} \sim\ & s(\text{surp},\ bs = \text{"cr"},\ k = 6) \\ & + s(\text{prev\_surp},\ bs = \text{"cr"}, \\ & k = 6) \\ & + te(\text{Freq\_brWaC\_log}, \\ & \text{Word\_Length},\ bs = \text{"cr"}) \\ & + te(\text{prev\_freq},\ \text{prev\_len}, \\ & bs = \text{"cr"})\end{aligned} \tag{4}$$

And for linear models:

$$\begin{aligned}\text{reading\_time} \sim\ & \text{surp} + \text{prev\_surp} \\ & + te(\text{Freq\_brWaC\_log}, \\ & \text{Word\_Length},\ bs = \text{"cr"}) \\ & + te(\text{prev\_freq},\ \text{prev\_len}, \\ & bs = \text{"cr"})\end{aligned} \tag{5}$$