Reading Time Modeling with Pixel-Based Language Models

Junlin Li[1] **Sara Alhajeri**[1] **Tatsuki Kuribayashi**[1,2]
[1]Mohamed bin Zayed University of Artificial Intelligence [2]Tohoku University
{junlin.li, sara.alhajeri, tatsuki.kuribayashi}@mbzuai.ac.ae

Abstract

Tokenization-free, pixel-based language models (PLMs) have recently been proposed as an alternative to conventional token-based language models (TLMs), typically motivated by practical considerations, e.g., the flexibility to support new languages or vocabularies. Nevertheless, PLMs also have the potential to serve as infrastructure for computational (psycho)linguistics, where tokenization has been a source of bias in estimating information-theoretic values, such as surprisal, and has pushed the bottom-up visual orthographic features outside the scope. In this study, we first integrate PLMs into the modeling of naturalistic human reading and, in particular, examine their use to estimate next-word predictability, which can also integrate orthographic and parafoveal information.

Our results demonstrate that next-word prediction difficulty, computed in **pixel space**, has strong predictive power for gaze duration and word skipping behavior across various languages. In particular, their advantage is complementary to existing token-level surprisal and entropy estimates. Cross-lingual analyses with script variations also clarify their distinctive visual processing biases, facilitating research into how bottom-up visual information is incrementally integrated into expectation-based reading.

reading-time-modeling-with-pixel-lm

1 Introduction

Reading coordinates a stratification of low- and high-level cognitive processes (Engbert et al., 2005) across multiple modalities (Hoover and Gough, 1990). Naturalistic reading data, for example, indicate that a considerable proportion of words is processed without a fixation; they are possibly recognized in the parafovea (Schotter et al., 2012; Rayner, 1998) with only partial visual information (Reichle et al., 2003). Such a rapid

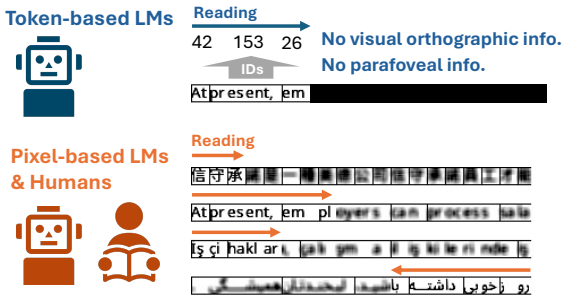

Figure 1: Simulating visual sentence processing with orthographic and parafoveal information using pixel-based LMs

pre-attentive visual processing suggests the importance of a bottom-up sensory prediction (Kok and de Lange, 2015; Tarkiainen et al., 1999), especially how graded visual presentation affects reading time and skipping (Gordon et al., 2013) via top-down prediction (Ding et al., 2025).

In computational psycholinguistics, neural language models (LMs) have been ubiquitous in estimating information-theoretic values, such as surprisal (Hale, 2001; Levy, 2008), to instantiate linguistic theories. Despite their successful interdisciplinary bridge, recent studies have also identified drawbacks of ordinal LMs defined over a fixed set of token IDs (token-based language models; TLMs) as tools for cognitive modeling. In particular, their tokenizers and vocabularies, predefined regardless of their cognitive plausibility (Haslett, 2025), potentially limit feasible experimental designs, e.g., estimating highly incremental probabilities across more fine-grained units, or implicitly bias experimental results (Nair and Resnik, 2023; Oh and Schuler, 2024; Giulianelli et al., 2024; Oh and Schuler, 2025; Kiegeland et al., 2026). For example, Nair and Resnik (2023) reported that a tokenizer or (sub)word definition easily biases the surprisal estimate $-\log p(\text{word}|\text{context})$. Another example is that a naive accumulation of subword

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 27481–27500
October 24-29, 2026 ©2026 Association for Computational Linguistics

probabilities results in their misallocation to the following word, e.g., "He give me", where the load of "give" can be dislocated to "me," if the alternative "gives" is also split into the common prefix "give" and "s" (Oh and Schuler, 2024; Giulianelli et al., 2024). Although some problems have quick solutions, tokenizer selection is inherently the source of biases in computational psycholinguistic experiments with TLMs. More fundamentally, TLM also limits the possibility of experimental design; for example, the word-to-ID conversion makes it impossible to study the visual and orthographic effects in reading.

In this paper, to handle such *tokenization bottlenecks* spread in modern TLM-based cognitive modeling, we examine the use of tokenization-free, pixel-based language models (**PLM**s) (Borenstein et al., 2023; Rust et al., 2023; Lotz et al., 2023) in computational psycholinguistics. PLMs encode text as pixel-based images, potentially without loss of visual fidelity to what humans can indeed see, enabling flexible research designs, for example, to explore how text layout, font, or orthographic complexity drives reading (Borenstein et al., 2023). This work, as a case study, specifically explores the psychometric predictive power of PLMs and parafoveal effects in reading, positing that early-stage visual processing may already begin before fixating on the exact word, and from an expectation-based reading perspective, visual word identifiability will be a strong predictor of human gaze duration and scan path. We compute word predictability and entropy with PLMs that rely on orthographic information and a blurred future context, intended to reflect parafoveal sight. These estimates are used to model reading time and word skipping behavior during naturalistic reading.

Our experiments address the following questions:

- **Q1:** Do PLMs without a tokenizer predict reading time and word skipping better than or complementary to TLM surprisal, which relies on a tokenizer?
- **Q2** : Does PLM-estimated visual word-prediction difficulty offer a general predictor for cognitive modeling across different languages and writing systems?

Our experimental results show that the visual word prediction difficulty estimated by PLMs offers complementary predictive power to conventional token-based surprisal. This advantage is observed across languages, particularly in those with scripts other than the Latin alphabet, demonstrating that PLM is a useful infrastructure for estimating word predictability and modeling human reading across diverse languages.

2 Related Work

2.1 Reading Time Modeling with Token-based Language Models

Language models (LMs) have been widely used as tools to implement and quantitatively test (neuro and psycho)linguistic theories (Caucheteux et al., 2023; Ohams et al., 2026). Given that LMs are basically next-word prediction machines, their use is typically relevant to testing an expectation-based account of human sentence processing through estimating information-theoretic measures such as surprisal (Hale, 2001; Levy, 2008; Hale, 2016). Curiously, a misalignment between human reading measures and LM-estimated predictive measures has been indicated by recent studies (Kuribayashi et al., 2021; Shain et al., 2024; Oh and Schuler, 2023), raising several possible explanations. A conceptually intuitive implication is that prediction is not a complete representation of human sentence processing (Slaats and Martin, 2025), which features structural modularistic underpinnings (Coltheart et al., 2001) and thus is not purely probabilistic. But practically, another possibility rests on the biases of the proxies, i.e., LMs, in estimating predictive measures (Nguyen and Arehalli, 2026). Existing studies (Nair and Resnik, 2023; Giulianelli et al., 2024; Oh and Schuler, 2024, 2025), for example, demonstrate that the tokenizer substantially affects LM-based cognitive modeling. Besides, language prediction based on token identity prevents any bottom-up visual and orthographic decoding processes. This study explores a more neutral, realistic design for text encoding in LM-based cognitive modeling, i.e., simply inputting the text image as we read texts with our eyes.

2.2 Pixel-Based Language Models

Fortunately, there are practical techniques for processing language using text images as input. Pixel-based language model (PLM) (Rust et al., 2023), using a visual Transformer (ViT) as its backbone, is a representative model that seamlessly models language in a tokenizer-free paradigm, regardless of vocabulary and character-inventory con-

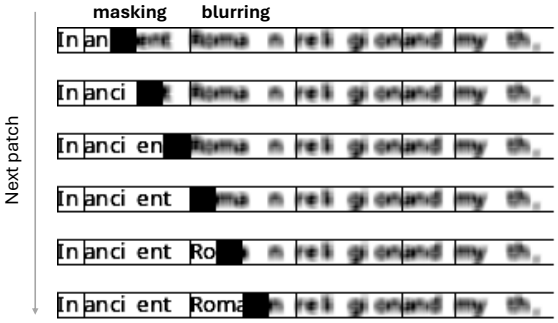

Figure 2: We approximate next-word predictability by patch-by-patch masking and blurring of the future context, simulating parafoveal sight, using PLM.

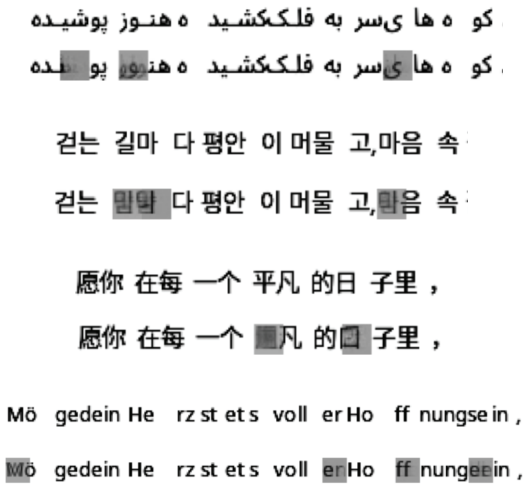

Figure 3: Examples of visual prediction from PLMs across multiple languages. The grayed areas are the PLM predictions for the regions originally masked.

straints (Lotz et al., 2023). Such PLMs have been adopted, for example, to enable a flexible addition of a new language or to process low-resource languages, such as historical languages without transcriptions (Borenstein et al., 2023). Beneath their powerful performance in the visual domain, probing analyses indicate that PLMs also learn linguistic knowledge, such as semantics and syntax (Tatariya et al., 2024) and exhibit cross-script generalizability through multilingual pre-training (Kesen et al., 2025). In contrast to these studies, we examine the use of PLMs for cognitive modeling.

3 Methods

This section briefly introduces PLMs, which model text sequences in the visual modality without relying on a tokenizer or a discrete vocabulary. We then explain how to compute next-word predictability with PLMs.

3.1 Pixel-Based Language Models

Given a token sequence $\boldsymbol{w} = [w_1, w_2, \ldots, w_n]$, a rendering function converts the sequence into an image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$, where C, H, and W denote the number of channels (e.g., RGB), height, and width, respectively. The rendered image is divided into $N = \frac{HW}{P^2}$ non-overlapping patches $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{N \times P^2 \times C}$, where P is the height/width of each patch. Each patch $\boldsymbol{x}_i \in \mathbb{R}^{P^2 \times C}$ is flattened and projected into a d-dimensional embedding $\boldsymbol{h}_i \in \mathbb{R}^d$ with an embedding layer of a PLM:

$$\boldsymbol{h}_i = \mathrm{Flatten}(\boldsymbol{x}_i)\boldsymbol{W}_E + \boldsymbol{b}_E + \boldsymbol{e}_i^{\mathrm{pos}}, \quad (1)$$

where $\boldsymbol{W}_E \in \mathbb{R}^{P^2 C \times d}$ and $\boldsymbol{b}_E \in \mathbb{R}^d$ are learnable parameters, and $\boldsymbol{e}_i^{\mathrm{pos}} \in \mathbb{R}^d$ denotes the positional embedding of the i-th patch. $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N]$ denotes the sequence of patch embeddings.

During training, a subset of patches in $\boldsymbol{H}$ is randomly masked. Let indices for masked patches be $\mathcal{I} \subset [1, \cdots, N]$. The masked input $\boldsymbol{H}_{\mathrm{masked}}$ is processed by a PLM $\mathcal{F}$ to reconstruct the original, unmasked version of the patches:[1]

$$\hat{\boldsymbol{X}} = \mathcal{F}(\boldsymbol{H}_{\mathrm{masked}}) \ , \quad (2)$$

where $\hat{\boldsymbol{X}} = \{\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_N\} \in \mathbb{R}^{N \times P^2 \times C}$. Then the reconstruction loss $\mathcal{L}$ is calculated over the indices $\mathcal{I}$ where the input was masked, and basically, the following objective is minimized:

$$\mathcal{L} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \|\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i\|_2^2 \ . \quad (3)$$

3.2 Visual Word Prediction Difficulty (VWPD) in Context

PLMs can quantify the prediction difficulty (VWPD) of the next word in the pixel space. Given a sentence $\boldsymbol{w} = [w_1, \cdots, w_n]$ and its image patches $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$, suppose a target token w_t spans from s_t-th to e_t-th patches

[1] Strictly speaking, the PLM architecture used (Rust et al., 2023; Kesen et al., 2025) adopts an encoder-decoder style, where the masked patches are excluded from the encoder input, and learnable mask embeddings are inserted in the decoder part. Note that our claim does not assume specific PLM architectures, and we begin with a unidirectional formulation in Section 3.2.

$[\boldsymbol{x}_{s_t}, \cdots, \boldsymbol{x}_{e_t}] \subset \boldsymbol{X}$ in the rendered image ($1 \leq s_t \leq e_t \leq N$). A token may span over several patches, and a straightforward way to compute the next-word prediction difficulty is:

$$\vec{I}(w_t|\boldsymbol{w}_{<t}) := \sum_{k=s_t}^{e_t} \vec{s}(\boldsymbol{x}_k|\boldsymbol{X}_{<k}) \ , \tag{4}$$

where the target patches $[\boldsymbol{x}_{s_t}, \cdots, \boldsymbol{x}_{e_t}]$ are masked one-by-one to enforce the prediction (black masking parts in Figure 2). A function $s(\cdot)$ quantifies the reconstruction error of the actual visual token (also see Figure 3 for examples of predictions). For example, one can compute the instance-wise MSE in the pixel space:

$$\vec{s}(\boldsymbol{x}_k|\boldsymbol{X}_{<k}) = \left\| \vec{\tilde{\boldsymbol{x}}}_k - \boldsymbol{x}_k \right\|_2^2 \ , \tag{5}$$

$$\vec{\tilde{\boldsymbol{x}}}_k = \mathcal{F}_k(\boldsymbol{X}_{<k} \circ m(\boldsymbol{x}_k)) \ , \tag{6}$$

where $\mathcal{F}_k$ returns the k-th patch of the output, $\circ$ denotes the concatenation of patches, and $m(\cdot)$ is a masking function. A high prediction error of patches $\vec{s}(\boldsymbol{x}_k|\boldsymbol{X}_{<k})$ is aggregated into the prediction difficulty of the word in Eq. 4; a high value of $\vec{I}(w_t|\boldsymbol{w}_{<t})$ is expected to be linked with longer reading time. Another reconstruction error is also examined in Section 4.

Off-the-shelf PLMs, however, are typically implemented with a bidirectional encoder or encoder-decoder architecture, and fully omitting future context $\boldsymbol{X}_{>t}$ from the input may incur a severe training-inference gap. Reflecting our vision, the future context is not fully invisible, but parafoveal sight still catches some superficial cues, e.g., approximate word length or sentence length. Thus, we propose the simulation of parafoveal sight by feeding blurred future context information $\boldsymbol{X}_{>t}$ with a blurring function $g(\cdot)$ (see Figure 2).[2] This blurring design is ablated in the main experiments. Now, the prediction difficulty of a word w_t is redefined as follows:

$$\overleftrightarrow{I}(w_t|\boldsymbol{w}_{\backslash t}) := \sum_{k=s_t}^{e_t} \overleftrightarrow{s}(\boldsymbol{x}_k|\boldsymbol{X}_{<k}, g(\boldsymbol{X}_{>k})) \ , \tag{7}$$

$$\overleftrightarrow{s}(\boldsymbol{x}_k|\boldsymbol{X}_{<k}, g(\boldsymbol{X}_{>k})) = \left\| \overleftrightarrow{\tilde{\boldsymbol{x}}}_k - \boldsymbol{x}_k \right\|_2^2 \ , \tag{8}$$

$$\overleftrightarrow{\tilde{\boldsymbol{x}}}_k = \mathcal{F}_k(\boldsymbol{X}_{<k} \circ m(\boldsymbol{x}_k) \circ g(\boldsymbol{X}_{>k})) \ . \tag{9}$$

[2]For brevity, $g(\cdot) : \mathbb{R}^{P^2 \times C} \rightarrow \mathbb{R}^{P^2 \times C}$ can be applied element-wise if the argument is a sequence of patches, e.g., $g(\boldsymbol{X})$ indicates $[g(\boldsymbol{x}_1), \cdots, g(\boldsymbol{x}_N)]$.

where the prediction difficulty $\overleftrightarrow{I}(w_t|\boldsymbol{w}_{\backslash t})$ is conditioned by bi-directional context (with partial blurring to the future context). Through our experiments, we use this version (Eq. 7) of visual word predictability.

3.3 SSIM Reconstruction Measure

For the reconstruction error term, $\vec{s}(\cdot)$ in Eq. 5 or $\overleftrightarrow{s}(\cdot)$ in Eq. 8, another formulation beyond instance-wise MSE can also be examined. For example, structural similarity (SSIM) (Wang et al., 2004) is widely used in the computer vision field. Given that the SSIM is a similarity measure, one minus the SSIM can be adopted for estimating VWPD:

$$\begin{aligned} &1 - \mathrm{SSIM}(\tilde{\boldsymbol{x}}_k, \boldsymbol{x}_k) \\ &= 1 - \frac{(2\mu_{\tilde{x}}\mu_x + C_1)(2\sigma_{\tilde{x}x} + C_2)}{(\mu_{\tilde{x}}^2 + \mu_x^2 + C_1)(\sigma_{\tilde{x}}^2 + \sigma_x^2 + C_2)} \ , \end{aligned} \tag{10}$$

where $\mu_{\tilde{x}}$ and μ_x denote the mean pixel intensities of the reconstructed patch $\tilde{\boldsymbol{x}}_k$ and ground-truth patch $\boldsymbol{x}_k$, respectively. $\sigma_{\tilde{x}}^2$ and σ_x^2 denote their variances, $\sigma_{\tilde{x}x}$ denotes the covariance between them, and C_1 and C_2 are small constants for numerical stability. One can apply this formula in Eq. 5 or Eq. 8 instead of instance-wise MSE, and we preliminarily examine this variant in Section 5.1.

3.4 Visual Prediction Uncertainty in Context

We also examined the entropy of next-word prediction in the pixel space, which quantifies the uncertainty in predicting black (stroke) or white (no stroke) rather than noisy gray predictions. Each dimension in a reconstructed pixel $\hat{\boldsymbol{p}}_k$ at position k is first normalized with a sigmoid function (0 for white pixel prediction and 1 for black pixel prediction):

$$\hat{\boldsymbol{p}}_k = \sigma(\overleftrightarrow{\hat{\boldsymbol{x}}}_k) \in [0, 1]^{P^2 \times C} \ . \tag{11}$$

The entropy measure is computed as follows:

$$H(w_t|\boldsymbol{w}_{\backslash t}) := \sum_{k=s_t}^{e_t} \sum_{l=1}^{P} \sum_{m=1}^{P} \sum_{c=1}^{C} h(\hat{p}_{k,l,m,c}) \ , \tag{12}$$

$$h(p) = -p \log p - (1-p) \log(1-p) \ , \tag{13}$$

where $\hat{p}_{k,l,m,c} \in [0, 1]$ denotes (l, m)-th element in c-th channel of k-th patch prediction. A high uncertainty value may parallel with the general difficulty of anticipating the next word, expected to be linked with longer reading time. We also examine the entropy measures from PLMs and TLMs for word skipping modeling.

4 Experimental Settings

4.1 Language Models

We examine PLMs as well as TLMs with a comparable model size. All model details are in Table 2 in Appendix A.

PLM Settings. For English experiments, we evaluate `pixel-m4`. For Chinese reading, we also use the off-the-shelf `pixel-m4` model (Kesen et al., 2025), which has been pretrained on Chinese script. For other languages, we finetune `pixel-m4` on 100,000 examples from the `allenai/c4` dataset (Raffel et al., 2020) for 150,000 steps with a batch size of `32`. The finetuned models reach a loss below 0.020, comparable to the official checkpoint on the pre-training data. For the blurring function $g(\cdot)$ in Eq. 8, we use Gaussian Blurring with a kernel size of 5 and a sigma of 3 (see Appendix B.2). We also examine a no-blurring condition, in which the future context is fully visible, similarly to BERT (Devlin et al., 2019).

TLM Settings. As baselines, we use monolingual GPT2 (Radford et al., 2019), monolingual BERT (Devlin et al., 2019), and multilingual mGPT (Shliazhko et al., 2024). Comparisons between TLMs and PLMs with parafoveal information clarify the general advantage of orthographically grounded prediction measures compared to discrete token prediction without any access to future word information. BERT is also compared, given the bi-directional nature of the PLM encoder. We use the surprisal (Hale, 2001; Levy, 2008) and entropy (Hale, 2016) computed by TLMs as predictors. Specifically, for the GPT and mGPT models, the surprisal of a word is calculated as:

$$\vec{I}_{\text{tok}}(w_t|\boldsymbol{w}_{<t}) = -\log_2 p(w_t \mid w_1, \ldots, w_{t-1}) \quad . \tag{14}$$

For the BERT models, the surprisal is computed with a bidirectional context:

Lang.	Dataset Code	Data	Skip Rate	Avg. TRT	#Subj.	Script
German	Deu (meco po)	MECO-2 (ge_po)	0.19	380.8	40	Latin
	Deu (meco zu)	MECO-2 (ge_zu)	0.21	390.6	45	
English	Eng (meco en)	MECO-1 (en)	0.30	346.1	46	
	Eng (meco uk)	MECO-2 (en_uk)	0.28	348.2	50	
Spanish	Esp (meco es)	MECO-1 (es)	0.33	406.7	48	
	Esp (meco ch)	MECO-1 (es_ch)	0.29	367.7	45	
Persian	Far	PSR	0.13	450.4	60	Abjad
Korean	Kor	MECO-1 (ko)	0.31	348.9	32	Hangul
Chinese	Simp (meco)	MECO-2 (ch-s)	0.43	348.7	39	Hanzi
	Trad (meco)	MECO-2 (ch-t)	0.44	435.3	38	
Turkish	Tur	MECO-2 (tur)	0.19	363.7	16	Latin

Table 1: We use multilingual and multi-script eye-movement-while-reading data across 459 readers.

$$\overleftrightarrow{I}_{\text{tok}}(w_t|\boldsymbol{w}_{\setminus t}) = -\log_2 p(w_t \mid w_1, \ldots, w_{t-1}, \texttt{[MASK]}, w_{t+1}, \ldots, w_n) \quad . \tag{15}$$

Entropy is calculated as the expected value of surprisal over the model's tokenizer vocabulary (Pimentel et al., 2023).

4.2 Human Reading Data

We model reading time and word skipping behavior of a total of 459 subjects from different eye-movement-while-reading corpora, covering German (Deu), English (Eng), Spanish (Esp), Farsi (Far), Korean (Kor), Simplified Chinese (Simp), Traditional Chinese (Trad), and Turkish (Tur). Data are from MECO Wave 1 (Siegelman et al., 2022) or Wave 2 (Siegelman et al., 2025), except for Persian reading corpus (Tekbudak et al., 2024). Data statistics and details are presented in Table 1. We take the following reading measures into consideration: First Fixation Duration (FFD), First-pass Reading Time (FPT), Total Reading Time (TRT), First-pass Skipping (FSKP), and Overall Skipping (SKP). For durative measures, any zero fixation duration and zero reading time are excluded. We adopt two word skipping measures: (i) first-pass skipping — whether an interest area is skipped before a later interest area is fixated (1=yes, 0=no); and (ii) Overall skipping — whether an interest area is not fixated throughout the current trial (1=yes, 0=no).

4.3 Regression Models

Predictors of Interest. We extract visual prediction difficulty (Eq. 7) and visual entropy (Eq. 12) from PLMs, as well as the widely used token-level

surprisal and entropy from TLMs. For each regression model, we only use one predictor of interest, denoted as x. The regression formula is as follows:

$$\{\mathtt{y_n}\} \sim \{\mathtt{x_n}\} + \mathrm{l_n} + \mathrm{f}_n + \mathrm{l_{n-1}} + \mathrm{f_{n-1}} + \mathrm{idx_n} + (1 + \{\mathtt{x_n}\} + \mathrm{l_{n-1}} + \mathrm{f_{n-1}}|\mathrm{subject_id}) + (1|\mathrm{stimuli_id}) \quad , \tag{16}$$

where l is word length, and f is zipf word frequency, $\mathrm{idx_n}$ is the word index in sentence. The regression formula for the baseline is as follows:

$$\{\mathtt{y_n}\} \sim \mathrm{l_n} + \mathrm{f}_n + \mathrm{l_{n-1}} + \mathrm{f_{n-1}} + \mathrm{idx_n} + (1 + \mathrm{l_{n-1}} + \mathrm{f_{n-1}}|\mathrm{subject_id}) + (1|\mathrm{stimuli_id}) \quad . \tag{17}$$

Here, y is the reading measures, i.e., FFD, FPT, TRT, FSKP, or SKP. For reading measures (FFD, FPT, and TRT), we applied log transformation, and skipping measures (FSKP and SKP) are converted with binomial family. We conduct our main experiments with generalized additive mixed models (GAMM) (Sóskuthy, 2017; Wood et al., 2015), using `mgcv` packages (Wood, 2017) and setting all effects as linear. For main experiments, we use the following designs of x:

- pixel: VWPD from PLM, without blurring the future context.
- pixel_uni: VWPD from PLM, with blurring the future context.
- gpt: surprisal from the respective monolingual GPT2.
- mgpt: surprisal from multilingual GPT.
- bert: bi-directional surprisal from the respective monolingual BERT.
- pixel_uni + gpt: pixel_uni and gpt are both included, without interaction.
- pixel_uni ∗ gpt: Interaction between pixel_uni and gpt with their main effects.

4.4 Evaluation Metrics

Following existing studies (Goodkind and Bicknell, 2018), we use the difference in loglikelihood between two nested models with and without the predictor of interest $\mathtt{x_n}$ from TLMs or PLMs. In our case, the joint contribution of the predictor and its subject-specific random slope is measured. We refer to this metric as psychometric predictive power (PPP). To test robustness, we rerun each formula 100 times using a crossed random-effects bootstrap with replacement to obtain a distribution of the log-likelihood difference. Similarly to Goodkind and Bicknell (2018), we fit regression models to explain the data of interest, but in Appendix D, we also adopt 10-fold cross-validation, as in (Wilcox et al., 2023), yielding consistent results. When reporting a single PPP value for each language with several datasets (Table 1), PPPs for different datasets in the target language are averaged.

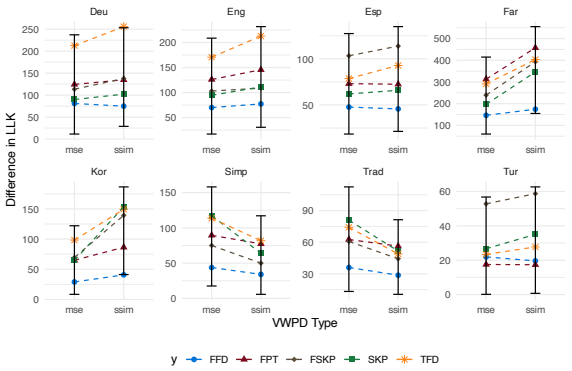

Figure 4: MSE-based and SSIM-based VWPD estimates show slight, but different PPP patterns across languages. The error bars in the figures indicate 95% percentile bootstrap confidence intervals of PPP across the settings.

5 Experimental Results

5.1 Preliminary Experiments: Comparing Two VWPD Measures

In Section 3.2, we introduced MSE and SSIM as an image reconstruction error in Eq. 8. As preliminary experiments, we compare MSE-based and SSIM-based image reconstruction errors as general indicators of next-word prediction difficulty. Figure 4 shows that different PPPs are obtained with different error formulations, and there was a clear empirical trend that SSIM is slightly better or comparable to MSE, except for the Chinese (Simplified and Traditional) language. This might reflect a special visual processing of logograms or the complexity of Chinese characters. In our main experiments, we show SSIM results across all languages, except for MSE for the Chinese (Simplified and Traditional) data. Full SSIM and MSE comparison is shown in Appendix E.

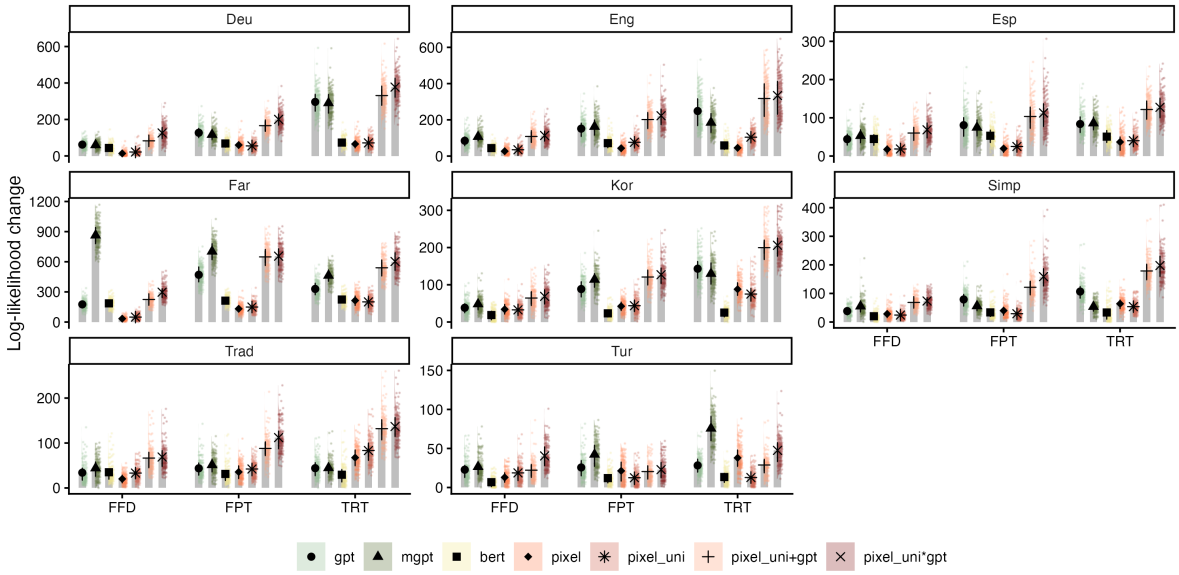

Figure 5: PPP (larger better) of VWPD and token surprisal for fixation measures (FFD = First Fixation Duration, FPT = First-pass Reading time, TRT = Total Reading Time). All zero-duration or zero-reading-time is filtered to disentangle from skipping measures. The Y-axis is PPP. The error bars in the figures indicate standard deviation.

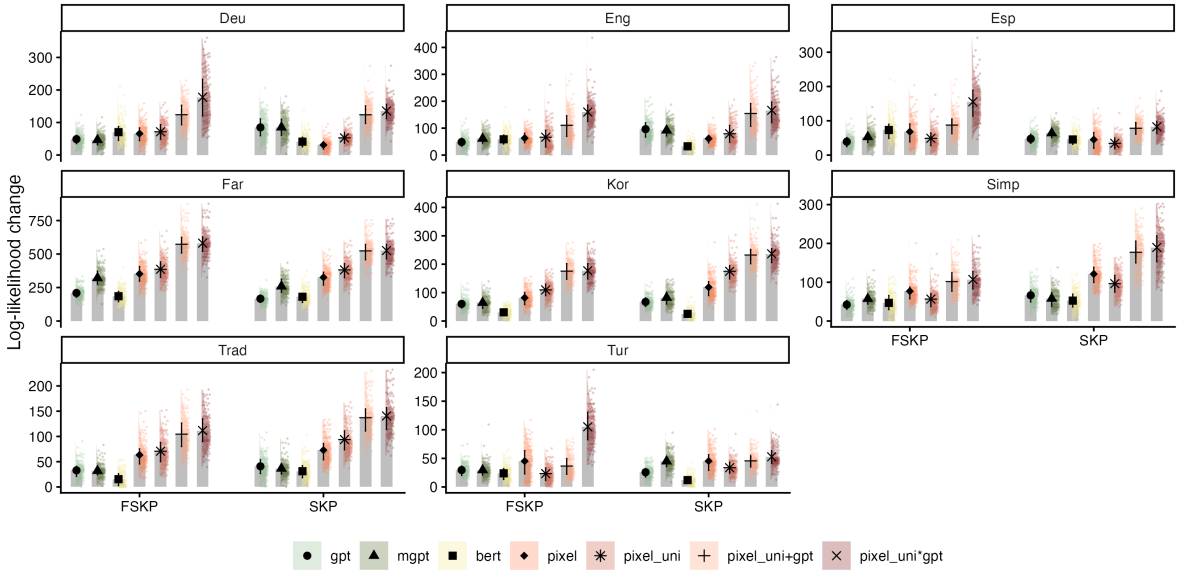

Figure 6: PPP (larger better) of VWPD and token surprisal for skipping measures (FSKP = First-pass skipping, SKP = Overall skipping). The Y-axis is PPP. The error bars in the figures indicate standard deviation.

5.2 Comparing PLMs and TLMs

We evaluate two versions of PLMs with and without blurring for the future context, and three TLMs (monolingual GPT2, BERT, and mGPT).

Fixation and Reading Time Figure 5 shows the results for FFD, FPT, and TRT modeling. Visual word predictability (VWPD) is used for PLMs, and surprisal is used for TLMs. PPPs for PLMs are somewhat worse than TLMs, but when these two predictors are jointly used, the PPP substantially increases, indicating their complementary advantage in reading time modeling.

Word Skipping Figure 6 shows the results for FSKP and SKP; whether the word is skipped or not during reading is predicted. In contrast to gaze duration measures, PLMs yield comparable or better predictive power, especially for the Farsi (Far), Korean (Kor), and Chinese (Simp and Trad) languages; interestingly, these languages do not adopt the Latin alphabet. In these languages, in

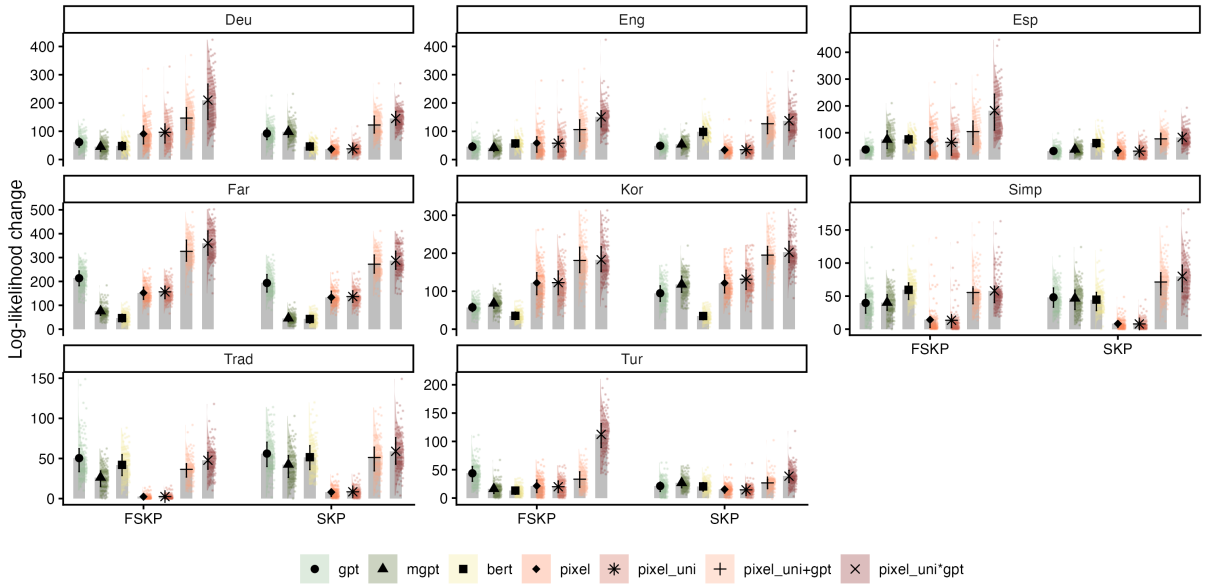

Figure 7: PPP (larger better) of visual and token entropy for skipping measures (FSKP = First-pass skipping, SKP = Overall skipping). The Y-axis is PPP. The error bars in the figures indicate standard deviation.

other words, if a word can easily be predicted visually from the surrounding context, it is likely to be skipped during reading. This idea is also compatible with entropy measures (how the next word is uncertain), and Figure 7 shows the comparison of entropy computed by PLMs and TLMs. The results generally reproduce our observations with surprisal, i.e., token-level and visual entropy are complementary. We report the cross-validation analysis in appendix D, which substantiates the tendencies we have pointed out above.

5.3 Comparing Addition and Interaction

Comparing the PPP of the additive model ($\text{pixel_uni} + \text{gpt}$) and the interactive model ($\text{pixel_uni} * \text{gpt}$), we sometimes observe an advantage of the interaction term, especially in the first-pass measures (FFD and FSKP) in German (Deu), Spanish (Esp), and Turkish (Tur). These have somewhat longer word length and richer morphology than those of other languages (Siegelman et al., 2022). Thus, a strong interaction may imply a more general hierarchical engagement of prediction in reading, spanning from shallow visual perception to morphological or richer linguistic processing. In the following section, we look into how this interaction shapes first-pass reading behavior. The detailed breakdown of our main results into individual dataset is in Figures 13 and 14.

6 Analysis: Characteristics of VWPD

6.1 Interaction with TLM Surprisal

Given the complementary advantage of predictors from PLM and TLM, we exploratorily extend the analysis, specifically on how PLM and TLM predictors interact. We focus on early-stage reading measures, such as first fixation duration (FFD), which typically aligns with early lexical access (Kurnik, 2016), first-pass reading time (FPT), as well as first-pass refixation duration (FPT minus FFD; FPREF), which indicates the "failure to achieve lexical access" (Lowell and Morris, 2014). For simplicity, we only consider the fixed effect of PLM surprisal (namely VWPD) and TLM surprisal (we use GPT2 surprisal in each language). We split the data into those with relatively high surprisal (+1 SD) and low surprisal (-1 SD) and analyze how the VWPD effect differs between the two subsets.

Figure 8 shows the interaction between VWPD and GPT2 surprisal. Interestingly, for FFD, we found that high surprisal (+1 SD in the figure) can reverse (e.g., in Deu and Tur) the VWPD–FFD relationship to be negative. In contrast to FFD, FPREF results indicate a positive mediation of VWPD, as high surprisal (+1SD in the figure) leads to a positive VWPD-FPREF relationship. That is, words with a high surprisal (lexical difficulty) tend to weaken the VWPD effect for first fixation but strengthen it for refixation. One possible interpretation of the neutralized or negative interac-

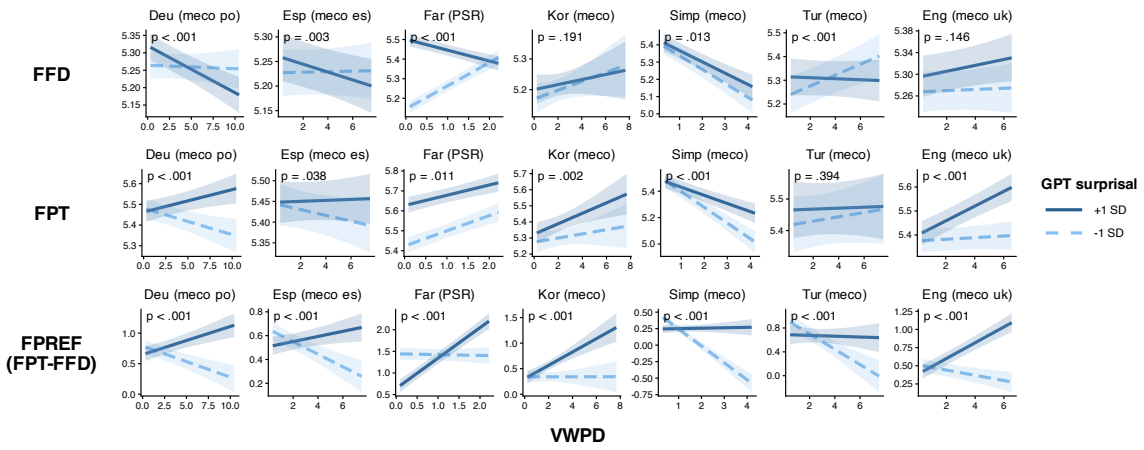

Figure 8: Interaction between VWPD and GPT2 surprisal (+1 SD = high, −1 SD = low) to FFD, FPT, and FPREF.

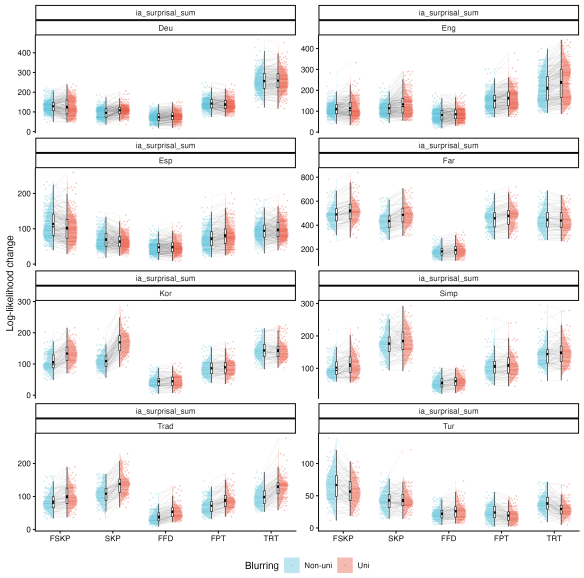

Figure 9: Effect of Unidirectional Gaussian Blurring. The points are obtained through bootstrapping, and the blue (left) ones are from the `pixel` setting without blurring, while the red (right) ones are from the `pixel_uni` one with blurring. Y-axis is PPP.

tion between VWPD and surprisal in FFD prediction is that lexical prediction (token-level surprisal) may involve pre-activation of visual representations (Buggy et al., 2026); if more effort is devoted to predicting the next word, early-stage visual processing can be skipped. The positive interaction in FPREF aligns with the finding that orthographic confusion is significant only for unpredictable words (Yao et al., 2022), as the positive interaction was only observed when people need multiple fixations (FPREF), potentially when lexical access initially failed.

6.2 Gaussian Blurring Effects

We also analyze how Gaussian blurring of the future context affects PPP. The improvement in predictive power (See Figure 9) with the blurring treatment is not substantial. Still, the advantage is slightly pronounced for SKP in Farsi (Far), Korean (Kor), and Traditional Chinese (Trad), where non-Latin alphabets are used. The advantage of blurring in Trad has also been suggested by the finding that reading Traditional Chinese characters narrows the perceptual span (Pan and Yan, 2024), supporting an information-processing account of visual perception during reading, as assumed in the OB1-Reader model (Snell et al., 2018). Nevertheless, our Gaussian Blurring is our tentative implementation, and more cognitively motivated designs should be explored in future studies.

7 Conclusion

We demonstrate that pixel-based LMs can estimate expectation-based visual word-processing measures that complement those estimated by token-based LMs, which are especially relevant to pre-fixation processing on partial input. This study is the first to introduce a pixel-based language model into computational cognitive modeling. The results demonstrate, at least partial cognitive plausibility of PLM, especially its alignment with early-stage language processing, such as first-pass skipping.

Limitations

First, a pixel-based language model has various rendering strategies (Lotz et al., 2023). We only use the most cross-lingually generalizable setting, namely bigram rendering, which is the foundation

of the only multilingual pixel-based model so far. We agree that calculating visual word prediction metrics is highly sensitive to the text rendering strategy. We encourage future work to explore different rendering strategies and analyze their effect on the cognitive plausibility of pixel-based language modeling. Second, for pixel blurring, we have not conducted an intensive hyperparameter analysis regarding the effects of kernel size and sigma on model performance. We also only tried Gaussian Blurring, without exploring other image corruption methods such as Gaussian Noise. Future work could also compare more cognitively motivated parafoveal blurring and noise. Besides, the human parafovea exhibits graded visual acuity, which is not accounted for in this study. We hope that the current study initiates the primary work and establishes the foundational settings, and that future work will continue to explore parafoveal vision modeling by manipulating the gradedness of visual acuity.

Ethical Considerations

This study uses open-source eye-movement-while-reading data collected from IRB-approved studies. No personally identifiable information is used in our experiment. Our study aims to gain knowledge about human cognition. We do not curate any system to replace humans in terms of reading.

AI Writing/Coding Assistance Policy

AI writing assistance systems like Grammarly and ChatGPT are occasionally used to correct grammatical and spelling errors and to adjust phrasing. Our use of writing assistants complies with the AI writing assistance policy described in ARR guidelines under category (a), assistance purely with the language of the paper.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers 23K16938 and 26H02511.

References

Nadav Borenstein, Phillip Rust, Desmond Elliott, and Isabelle Augenstein. 2023. PHD: Pixel-based language modeling of historical documents. In *Proceedings of EMNLP 2023*, pages 87–107.

Ryan Buggy, Stephanie Cho, Cory Shain, and Adrian Staub. 2026. Predictability Effects in Natural Reading are Logarithmic: Evidence from an Eye-Movement Replication of Brothers and Kuperberg (2021). *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 7(3):430–441.

Max Coltheart, Kathleen Rastle, Conrad Perry, Robyn Langdon, and Johannes Ziegler. 2001. DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204–256.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings NAACL 2019*, pages 4171–4186.

Cui Ding, Yanning Yin, Lena Ann Jäger, and Ethan Wilcox. 2025. Modeling bottom-up information quality during language processing. In *Proceedings of EMNLP 2025*, pages 11709–11721.

Ralf Engbert, Antje Nuthmann, Eike M Richter, and Reinhold Kliegl. 2005. SWIFT: a dynamical model of saccade generation during reading. *Psychological review*, 112(4):777.

Mario Giulianelli, Luca Malagutti, Juan Luis Gastaldi, Brian DuSell, Tim Vieira, and Ryan Cotterell. 2024. On the proper treatment of tokenization in psycholinguistics. In *Proceedings of EMNLP 2024*, pages 18556–18572.

Adam Goodkind and Klinton Bicknell. 2018. Predictive Power of Word Surprisal for Reading Times is a Linear Function of Language Model Quality. In *Proceedings of CMCL 2018*, pages 10–18.

Peter C Gordon, Patrick Plummer, and Wonil Choi. 2013. See before you jump: full recognition of parafoveal words precedes skips during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(2):633–641.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL 2001*.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

David A. Haslett. 2025. Tokenization Changes Meaning in Large Language Models: Evidence from Chinese. *Computational Linguistics*, 51(3):785–814.

Wesley A Hoover and Philip B Gough. 1990. The simple view of reading. *Reading and Writing*, 2(2):127–160.

Ilker Kesen, Jonas F. Lotz, Ingo Ziegler, Phillip Rust, and Desmond Elliott. 2025. Multilingual pretraining for pixel language models. In *Proceedings of EMNLP 2025*, pages 29594–29611.

Samuel Kiegeland, Vésteinn Snæbjarnarson, Tim Vieira, and Ryan Cotterell. 2026. On the Proper Treatment of Units in Surprisal Theory. In *Proceedings of ACL 2026*, pages 32202–32224.

Peter Kok and Floris P. de Lange. 2015. Predictive coding in sensory cortex. In Birte U. Forstmann and Eric-Jan Wagenmakers, editors, *An Introduction to Model-Based Cognitive Neuroscience*, pages 221–244.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of ACL-IJCNLP 2021*, pages 5203–5217.

Mattias Kurnik. 2016. Bilingual lexical access in reading : Analyzing the effect of semantic context on non-selective access in bilingual memory. Master's thesis, Stockholm University, Centre for Research on Bilingualism.

Roger Levy. 2008. Expectation-based Syntactic Comprehension. *Cognition*, 106(3):1126–1177.

Jonas Lotz, Elizabeth Salesky, Phillip Rust, and Desmond Elliott. 2023. Text Rendering Strategies for Pixel Language Models. In *Proceedings of EMNLP 2023*, pages 10155–10172.

Randy Lowell and Robin K Morris. 2014. Word length effects on novel words: Evidence from eye movements. *Attention, Perception, & Psychophysics*, 76(1):179–189.

Sathvik Nair and Philip Resnik. 2023. Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship? In *Findings of EMNLP 2023*, pages 11251–11260.

Kien Nguyen and Suhas Arehalli. 2026. Word predictability estimates from language models are not robust to tokenizer vocabulary. In *Proceedings of CoNLL 2026*, pages 34–44.

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.

Byung-Doh Oh and William Schuler. 2024. Leading Whitespaces of Language Models' Subword Vocabulary Pose a Confound for Calculating Word Probabilities. In *Proceedings of EMNLP 2024*, pages 3464–3472.

Byung-Doh Oh and William Schuler. 2025. The impact of token granularity on the predictive power of language model surprisal. In *Proceedings of ACL 2025*, pages 4150–4162.

Chiebuka Ohams, Sathvik Nair, Shohini Bhattasali, and Philip Resnik. 2026. A predictive coding model for online sentence processing. *Journal of Memory and Language*, 146:104705.

Jinger Pan and Ming Yan. 2024. The perceptual span in traditional chinese. *Language and Cognition*, 16(1):134–147.

Tiago Pimentel, Clara Meister, Ethan G Wilcox, Roger P Levy, and Ryan Cotterell. 2023. On the effect of anticipation on reading times. *Transactions of the Association for Computational Linguistics*, 11:1624–1642.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Keith Rayner. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, 124(3):372–422.

Erik D. Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, 26(4):445–476.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. Language Modelling with Pixels. In *Proceedings of ICLR 2023*.

Elizabeth R Schotter, Bernhard Angele, and Keith Rayner. 2012. Parafoveal processing in reading. *Attention, Perception, & Psychophysics*, 74(1):5–35.

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).

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.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Maria Da Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost, Carolina A. Gattei, Areti Kalaitzi, Nayoung Kwon, Kaidi Lõo, and 12 others. 2022. Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, 54(6):2843–2863.

Noam Siegelman, Sascha Schroeder, Yaqian Borogjoon Bao, Cengiz Acartürk, Niket Agrawal, Lena S. Bolliger, Jan Brasser, César Campos-Rojas, Denis Drieghe, Dušica Filipović Đurđević, Sofya Goldina, Romualdo Ibáñez Orellana, Lena A. Jäger, Ómar I. Jóhannesson, Anurag Khare, Nik Kharlamov, Hanne B. S. Knudsen, Árni Kristjánsson, Charlotte E. Lee, and 30 others. 2025. Wave 2 of the Multilingual Eye-Movement Corpus (MECO): New text reading data across languages. *Scientific Data*, 12(1):1183.

Sophie Slaats and Andrea E Martin. 2025. What's surprising about surprisal. *Computational Brain & Behavior*, 8(2):233–248.

Joshua Snell, Sam van Leipsig, Jonathan Grainger, and Martijn Meeter. 2018. Ob1-reader: A model of word recognition and eye movements in text reading. *Psychological review*, 125(6):969–984.

Márton Sóskuthy. 2017. Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. *arXiv preprint arXiv:1703.05339*.

Antti Tarkiainen, Päivi Helenius, Peter C Hansen, Piers L Cornelissen, and Riitta Salmelin. 1999. Dynamics of letter string perception in the human occipitotemporal cortex. *Brain*, 122(11):2119–2132.

Kushal Tatariya, Vladimir Araujo, Thomas Bauwens, and Miryam De Lhoneux. 2024. Pixology: Probing the linguistic and visual capabilities of pixel-based language models. In *Proceedings of EMNLP 2024*, pages 3307–3320.

Zohre Soleymani Tekbudak, Mehdi Purmohammad, Ayşegül Özkan, and Cengiz Acartürk. 2024. The PSR corpus: A Persian sentence reading corpus of eye movements. *Behavior Research Methods*, 57(1):14.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Simon N. Wood. 2017. *Generalized Additive Models: An Introduction with R*, 2 edition. Chapman and Hall/CRC.

Simon N. Wood, Yannig Goude, and Simon Shaw. 2015. Generalized additive models for large data sets. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 64(1):139–155.

Panpan Yao, Adrian Staub, and Xingshan Li. 2022. Predictability eliminates neighborhood effects during Chinese sentence reading. *Psychonomic Bulletin & Review*, 29(1):243–252.

Lang.	Unidirectional LMs	Model size (#parameters)	Bidirectional LMs	Model size (#parameters)
Deu	dbmdz/german-gpt2	510MB (∼137M)	google-bert/bert-base-german-cased	439MB (∼109M)
Eng	openai-community/gpt2	548MB (∼137M)	google-bert/bert-base-cased	436MB (∼108M)
Esp	DeepESP/gpt2-spanish	548MB (∼124M)	dccuchile/bert-base-spanish-wwm-cased	440MB (∼110M)
Far	gpt2-persian	1.31GB (∼328M)	HooshvareLab/bert-base-parsbert-uncased	654MB (∼163M)
Kor	skt/kogpt2-base-v2	513MB (∼125M)	kykim/bert-kor-base	476MB (∼118M)
Simp (zh)	uer/gpt2-chinese-cluecorpussmall	421MB (∼102M)	google-bert/bert-base-chinese	412MB (∼102M)
Trad (zh)	ckiplab/gpt2-base-chinese	421MB (∼102M)	ckiplab/bert-base-chinese	409MB (∼102M)
Tur	ytu-ce-cosmos/turkish-gpt2	498MB (∼124M)	dbmdz/bert-base-turkish-cased	445MB (∼111M)
All	mGPT	3.45GB (∼1.4B)	Pixel-m4 (including finetuned ones)	449MB (∼112M)

Table 2: Models with their sizes in our experiment. All checkpoints of GPT and BERT are from https://huggingface.co/.

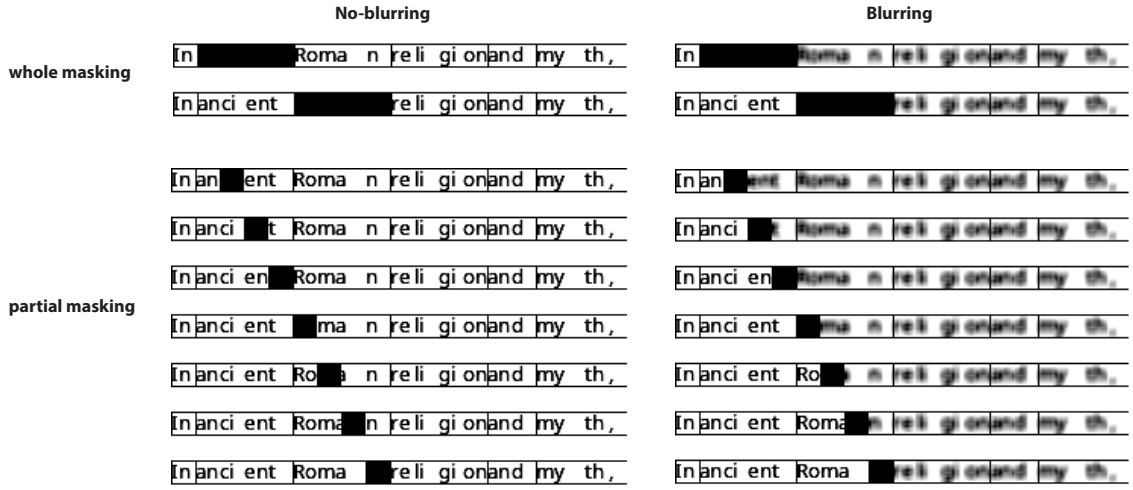

Figure 10: We approximate next-word predictability with patch-by-patch masking and add noise for the future context, simulating parafoveal sight, using PLM.

A Models

In Table 2, we present the details of models used in this study.

B PLM details

B.1 Conditions

Figure 10 shows all four conditions explored in this study. The first option is whether to mask all the patches related to the target word (whole masking) or only the target patch (partial masking) when computing patch predictability. The second option is whether to apply the noisy blurring to the future context, potentially simulating the parafoveal information, or not.

B.2 Gaussian Blurring

We apply the following noise function to the original image $\boldsymbol{X}$:

$$\boldsymbol{X}'(i,j) = \sum_{k=-m}^{m} \sum_{l=-m}^{m} \boldsymbol{X}(i-k,\ j-l) \cdot G(k,\ l) \tag{18}$$

$$m = \left\lfloor \frac{\texttt{kernel_size-1}}{2} \right\rfloor \tag{19}$$

$$G(k,l) = \frac{1}{2\pi\sigma^2} e^{-\frac{k^2+l^2}{2\sigma^2}} \tag{20}$$

$$\tag{21}$$

As a cognitively implausible baseline, we also examine a no-blurring condition in which the right-side context is visible.

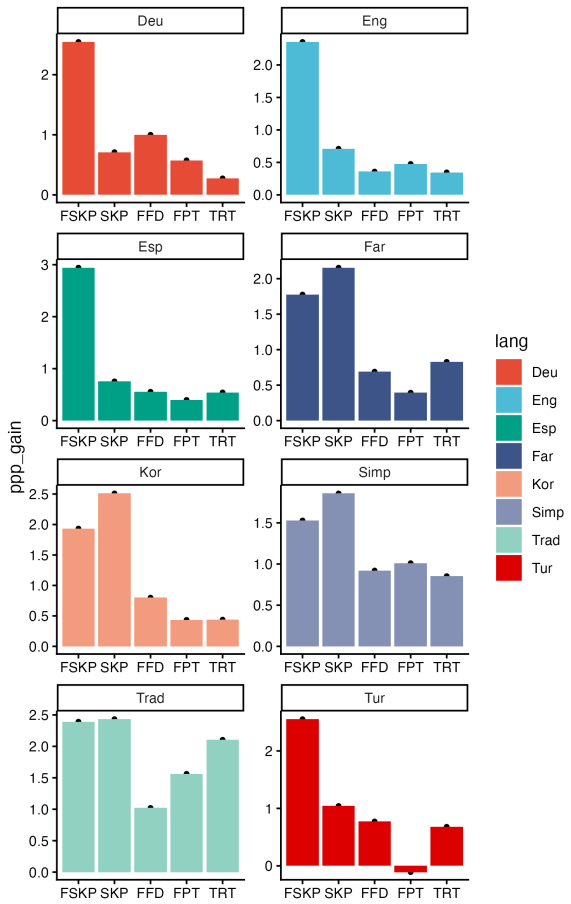

Figure 11: Gain of PPP after introducing PLM predictability to TLM ($\frac{\mathrm{PPP}_{\mathrm{pixel_uni*gpt}} - \mathrm{PPP}_{\mathrm{gpt}}}{\mathrm{PPP}_{\mathrm{gpt}}}$), which suggests multimodal prediction encodes rich metacognitive knowledge about skipping.

B.3 Others

To adapt Persian to the bigram renderer, we leave the space between two characters, even if they are connected visually. However, the space does not change the pre-space word into an unconnected form if it is originally in a connected form. To gather VWPD from Persian, we did not include the white-area. For the results in the current paper, we used the checkpoint after the 33,500th step results of the PSR dataset.

C Detailed Results

C.1 Interactive Effect of PLM prediction difficulty to TLM surprisal

Figure 11 shows a more detailed comparison between the regression model with GPT surprisal and that with GPT surprisal and the interaction between GPT surprisal and PLM VWPD.

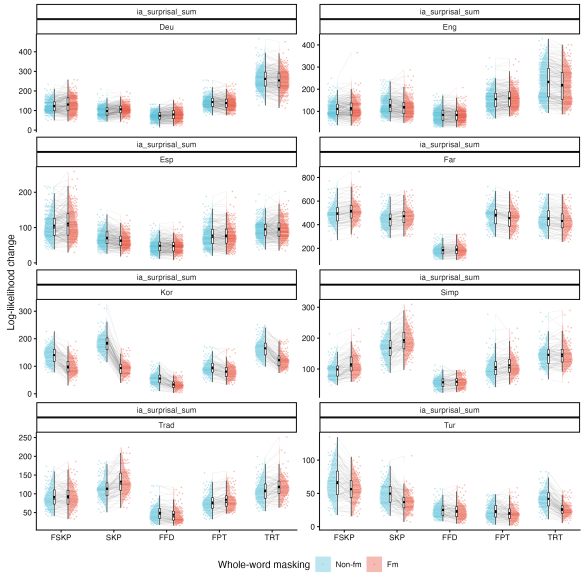

Figure 12: Effect of Whole Masking on PPP (Paired-t test of PPP is implemented, with matched regression model formula)

C.2 Effects of Whole Word Masking

Compared with blurring, the effect of masking is more nuanced across languages. We notice an advantage of whole-word masking in word skipping in Traditional Chinese, but a disadvantage in Korean. This potentially indicates a language-dependent sub-word processing strategy during naturalistic reading.

C.3 Main Results by Dataset

Below are the detailed main results for each dataset:

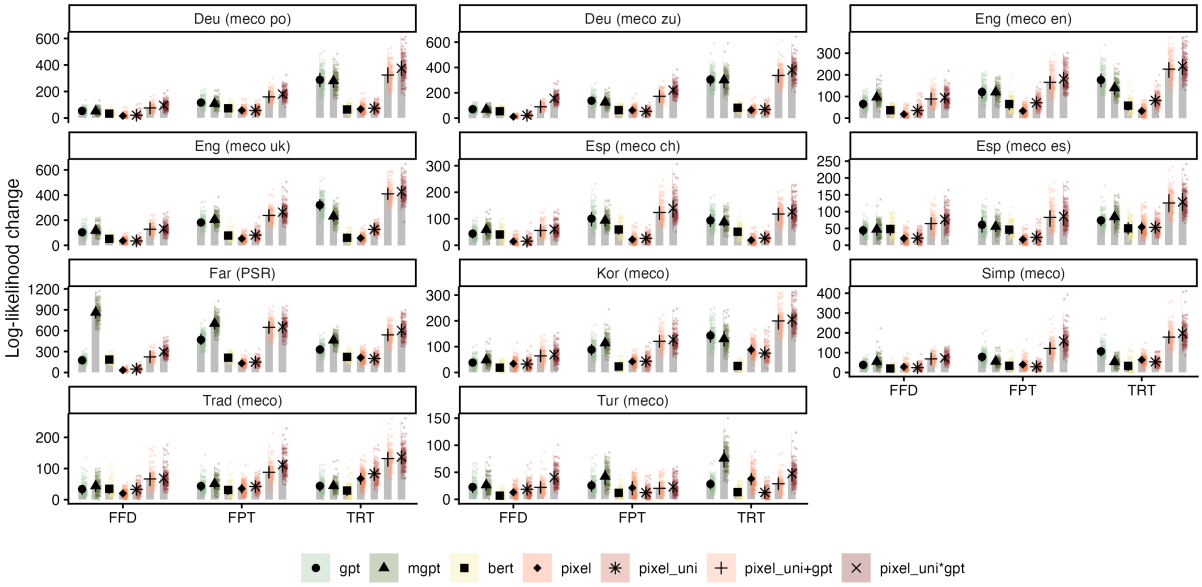

Figure 13: Main results by dataset (Reading Time)

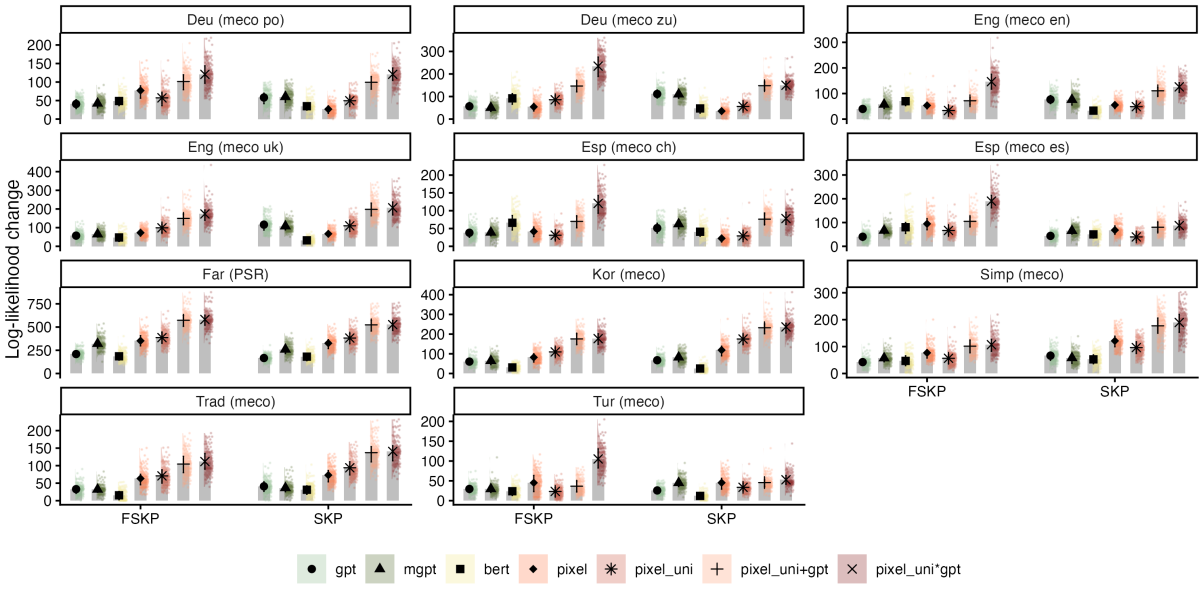

Figure 14: Main results by dataset (Skipping)

D Cross-validation

In Figure 15, we reran the inter-measure comparison results in Figure 4 through 10-fold cross-validation. In Figure 16, Figure 17 and Figure 18, we reproduce the results in Figure 5, Figures 6 and 7, and Table 4.

The cross-validation results do not contradict the observation that the MECO Chinese data show a stronger advantage of MSE-based VWPD. A revisit of the 10-fold cross-validation result supports the identical tendencies we have captured in Section 5. In the first place, an integration, especially an interaction, between token and visual prediction improves reading-time and skipping modeling. Besides, visual prediction accounts for skipping better than token prediction in certain scripts, notably Persian (Far), Korean, and Chinese scripts.

E Comparing MSE and SSIM across datasets

We report the performance of MSE and SSIM in terms of their advantage against the GPT2 model in Table 3.

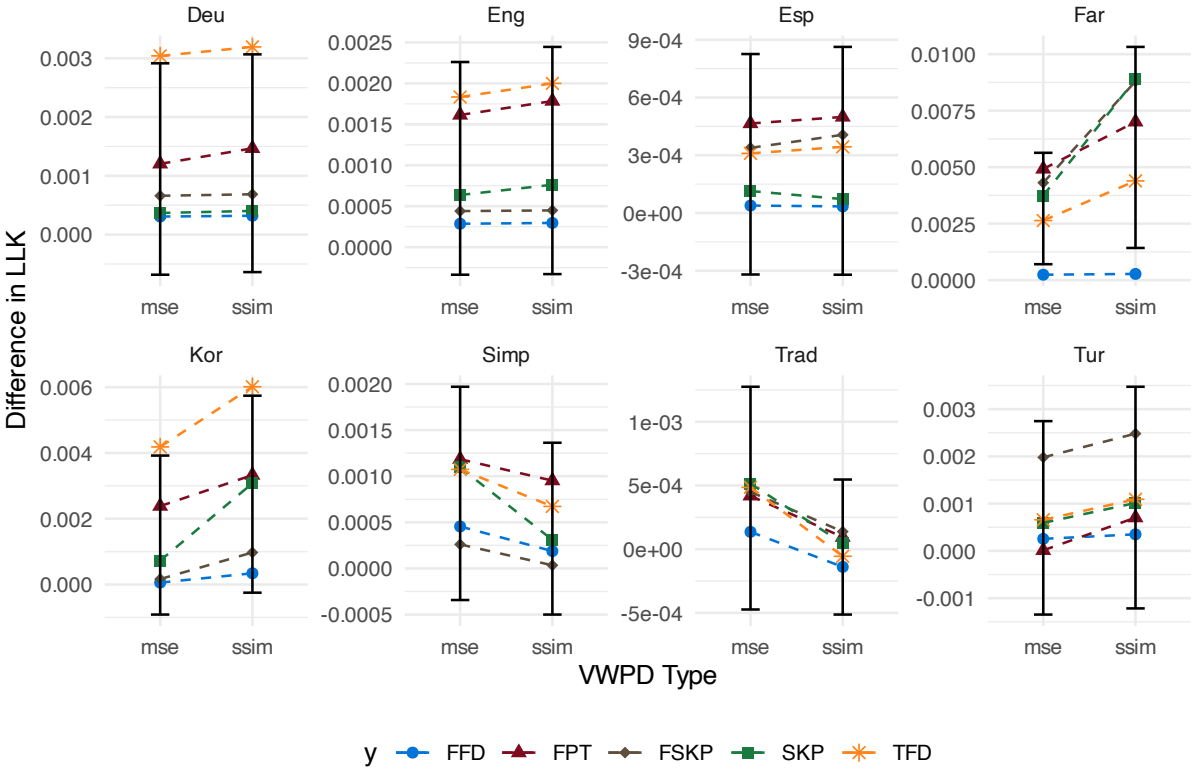

Figure 15: Comparison of MSE and SSIM measures in 10-fold cross-validation settings.

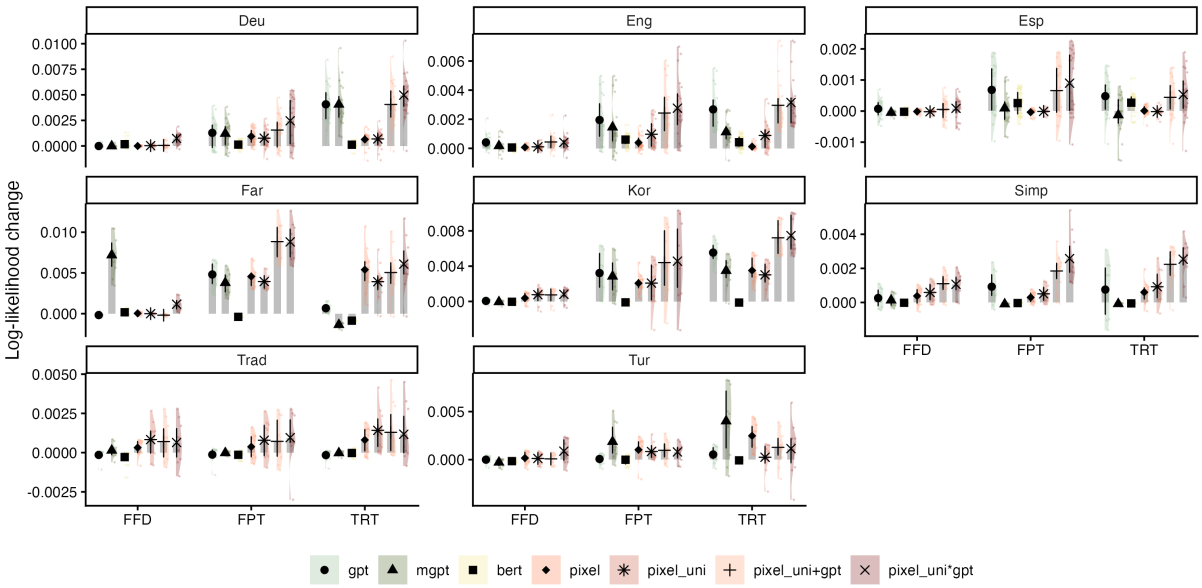

Figure 16: Main results by dataset (Reading Time) using 10-fold cross-validation based on visual and token surprisal. (Log-likelihood change is case-wise averaged)

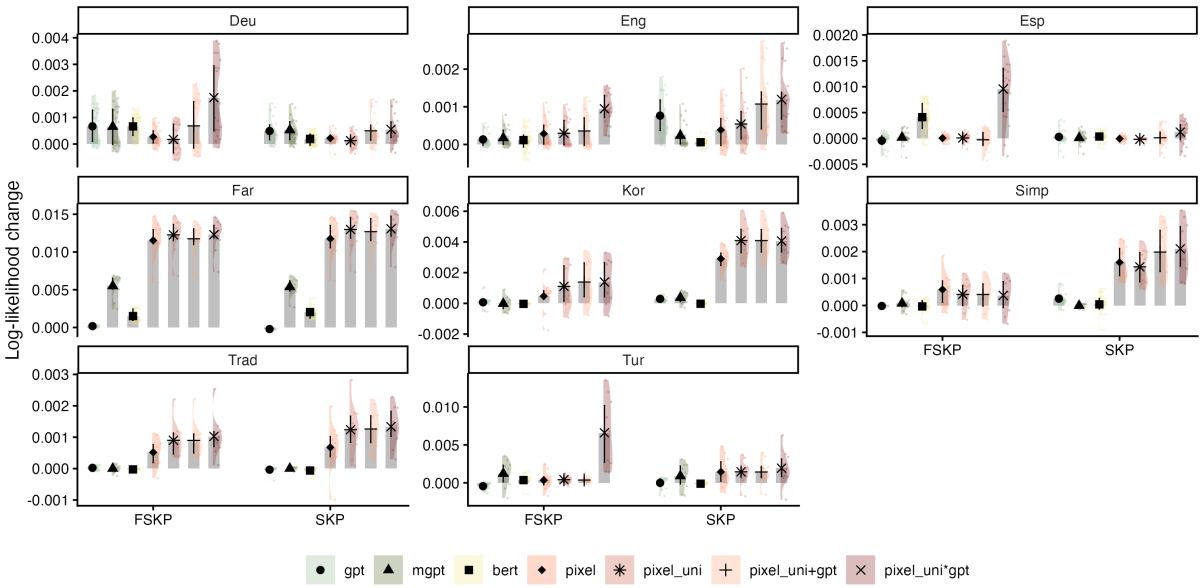

Figure 17: Main results by dataset (Skipping) using 10-fold cross-validation based on visual and token surprisal. (Log-likelihood change is case-wise averaged)

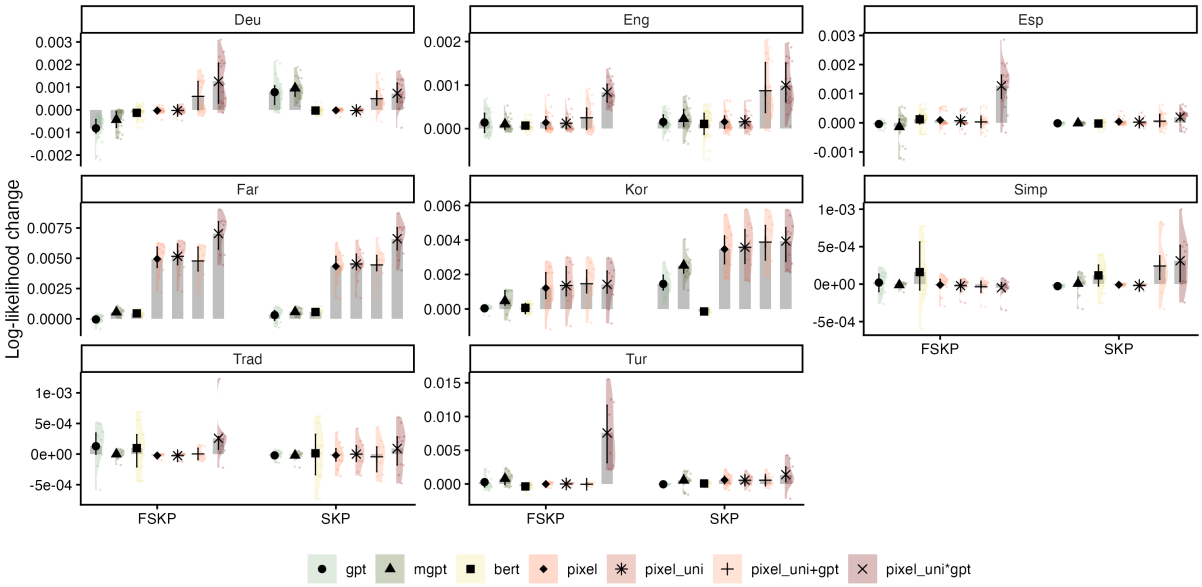

Figure 18: Main results by dataset (Skipping) using 10-fold cross-validation, based on visual and token entropy. (Log-likelihood change is case-wise averaged)

VWPD	Measure	Esp (es)	Esp (ch)	Deu (po)	Deu (zu)	Eng (uk)	Eng (en)	Kor	Far (PSR)	Simp	Trad	Tur
pixel_uni vs. gpt												
SSIM	FFD	-23.0***	-28.3***	-32.7***	-48.4***	-68.9***	-30.6***	-5.6*	-126.6***	-27.5***	-24.3***	-4.1*
	FPT	-37.3***	-74.3***	-60.1***	-84.1***	-100.4***	-49.9***	-44.8***	-321.2***	-63.8***	-24.0***	-13.2***
	TRT	-20.8***	-66.7***	-214.0***	-235.6***	-194.3***	-94.6***	-68.1***	-129.1***	-90.6***	-30.9***	-15.5***
	FSKP	+26.4***	-6.9**	+16.4***	+28.9***	+41.8***	-6.1*	+49.2***	+177.0***	-23.1***	-16.1***	-6.3***
	SKP	-4.3$^{\cdot}$	-22.1***	-8.7***	-55.4***	-6.2	-28.0***	+107.6***	+214.9***	-45.6***	-21.1***	+7.8***
MSE	FFD	-20.3***	-21.0***	-25.9***	-30.1***	-75.6***	-34.3***	-31.4***	-140.0***	-13.9***	-0.9	-11.2***
	FPT	-33.7***	-63.9***	-73.6***	-94.2***	-126.2***	-72.4***	-73.0***	-408.7***	-49.9***	-1.5	-12.4***
	TRT	-23.8***	-70.3***	-250.2***	-257.0***	-268.6***	-128.4***	-123.7***	-249.6***	-52.9***	+39.2***	-22.6***
	FSKP	+22.6***	-4.6$^{\cdot}$	+10.3**	+7.3*	+16.1***	-0.3	-14.2***	-109.4***	+14.2***	+37.9***	-14.2***
	SKP	-9.1***	-21.7***	-14.5***	-63.9***	-68.1***	-35.8***	-40.5***	-70.7***	+30.0***	+53.0***	-8.1***
pixel_uni × gpt vs. gpt												
SSIM	FFD	+26.6***	+13.7***	+31.7***	+37.0***	+18.1***	+12.1***	+21.7***	+103.8***	+11.1***	+4.0***	-6.0*
	FPT	+20.2***	+28.7***	+55.9***	+78.9***	+76.5***	+53.0***	+36.2***	+181.7***	+37.0***	+34.7***	-17.1***
	TRT	+43.1***	+26.6***	+74.4***	+65.9***	+100.3***	+54.2***	+59.8***	+228.4***	+19.3***	+7.3***	-4.4*
	FSKP	+134.7***	+49.0***	+76.1***	+143.9***	+65.1***	+56.2***	+114.5***	+328.7***	+16.2***	+23.7***	+16.1***
	SKP	+42.5***	+16.9***	+43.0***	+19.1***	+85.4***	+42.0***	+167.4***	+337.6***	+19.2***	+23.7***	+15.7***
MSE	FFD	+30.6***	+21.8***	+34.8***	+45.0***	+16.1***	+8.7***	-0.7	+56.2***	+33.4***	+28.9***	-11.1***
	FPT	+23.1***	+37.1***	+57.2***	+84.9***	+64.0***	+40.0***	+18.4***	+57.7***	+66.3***	+59.9***	-14.8***
	TRT	+41.2***	+21.4***	+64.6***	+73.6***	+51.8***	+31.1***	+23.3***	+79.2***	+78.4***	+82.7***	-10.8***
	FSKP	+132.7***	+54.0***	+66.1***	+112.0***	+36.8***	+56.1***	+44.2***	+56.2***	+61.0***	+72.2***	+5.6*
	SKP	+37.1***	+19.5***	+41.9***	+27.5***	+35.4***	+46.6***	+31.1***	+63.0***	+119.1***	+97.4***	+2.9

Table 3: Mean paired log-likelihood difference relative to GPT. Positive values favour the reported model. The estimation is based on boostrapping. The significance is based on paired t-test.* $p < .05$, ** $p < .01$, *** $p < .001$, $^{\cdot}$ $p < .1$.

VWPD	Measure	Esp (es)	Esp (ch)	Deu (po)	Deu (zu)	Eng (uk)	Eng (en)
pixel_uni vs. gpt							
SSIM	FFD	-0.000033	-0.000153	+0.000029	+0.000045$^{\cdot}$	-0.000421$^{\cdot}$	-0.000188$^{\cdot}$
	FPT	-0.000374	-0.001012**	-0.000885$^{\cdot}$	-0.000158	-0.001189$^{\cdot}$	-0.000768**
	TRT	-0.000125	-0.000871***	-0.004425***	-0.002344***	-0.002263***	-0.001323***
	FSKP	+0.000098	+0.000007	-0.000392***	-0.000599***	+0.000323*	+0.000003
	SKP	-0.000009	-0.000087	-0.000167$^{\cdot}$	-0.000562***	-0.000131	-0.000313**
MSE	FFD	-0.000036	-0.000146$^{\cdot}$	+0.000089	+0.000148	-0.000500*	-0.000215*
	FPT	-0.000376$^{\cdot}$	-0.001025**	-0.001090*	-0.001178*	-0.001984**	-0.001001**
	TRT	-0.000134	-0.000860***	-0.005057***	-0.002882***	-0.003238***	-0.001759***
	FSKP	+0.000123	+0.000002	-0.000342**	-0.001444***	+0.000232$^{\cdot}$	+0.000027
	SKP	-0.000004	-0.000005	-0.000246*	-0.000756**	-0.000857***	-0.000382**
pixel_uni × gpt vs. gpt							
SSIM	FFD	-0.000008	+0.000003	+0.000429*	+0.000408$^{\cdot}$	-0.000064	-0.000207$^{\cdot}$
	FPT	+0.000040	+0.000230*	+0.000511	+0.001832***	+0.001246***	+0.000365***
	TRT	-0.000155	+0.000232**	+0.000863*	+0.000743*	+0.000505	+0.000352*
	FSKP	+0.001094***	+0.000184	+0.000259*	+0.001035***	+0.000131	+0.000318**
	SKP	+0.000186**	-0.000164*	-0.000454*	-0.000338***	+0.000578**	+0.000175*
MSE	FFD	-0.000006	+0.000027	+0.000397**	+0.000283	-0.000072	-0.000219$^{\cdot}$
	FPT	-0.000005	+0.000108	+0.000677	+0.001291**	+0.000839***	+0.000222**
	TRT	-0.000179	+0.000171$^{\cdot}$	+0.000861*	+0.000642*	+0.000021	+0.000042
	FSKP	+0.001180***	+0.000225	+0.000162	+0.000210	+0.000030	+0.000331**
	SKP	+0.000169**	-0.000063	-0.000309$^{\cdot}$	-0.000299**	+0.000060	+0.000258**

VWPD	Measure	Kor	Far (PSR)	Simp	Trad	Tur
pixel_uni vs. gpt						
SSIM	FFD	+0.000710$^{\cdot}$	+0.000173	-0.000287	+0.000074	+0.000127
	FPT	-0.001143$^{\cdot}$	-0.000866	-0.000944**	+0.000049	+0.000777$^{\cdot}$
	TRT	-0.002521**	+0.003256***	-0.000795	+0.000122	-0.000284
	FSKP	+0.001036$^{\cdot}$	+0.012088***	-0.000023	-0.000052	+0.000854*
	SKP	+0.003796***	+0.013201***	-0.000286*	-0.000024	+0.001453*
MSE	FFD	-0.000061	+0.000217	+0.000329	+0.000978*	+0.000007
	FPT	-0.003004**	-0.003013***	-0.000419	+0.000917$^{\cdot}$	-0.000087
	TRT	-0.005010***	+0.000656*	+0.000167	+0.001579**	-0.000645*
	FSKP	+0.000079	+0.002944***	+0.000431*	+0.000877***	+0.000464*
	SKP	+0.000351	+0.003176***	+0.001179***	+0.001274**	+0.000643$^{\cdot}$
pixel_uni × gpt vs. gpt						
SSIM	FFD	+0.000555	+0.001022***	-0.000181$^{\cdot}$	-0.000408	+0.000556
	FPT	+0.001384$^{\cdot}$	+0.003938***	+0.000578	+0.000228	+0.000777*
	TRT	+0.001847**	+0.004196**	+0.000071	+0.000149	+0.000076
	FSKP	+0.001260*	+0.009560***	-0.000053	+0.000180	+0.003422**
	SKP	+0.003754***	+0.010907***	-0.000003	+0.000073	+0.001872*
MSE	FFD	-0.000212$^{\cdot}$	+0.000435*	+0.000687*	+0.000520	+0.000395
	FPT	+0.000341	+0.000923**	+0.001424**	+0.000952	-0.000039
	TRT	+0.000694**	+0.000879*	+0.001578**	+0.001596$^{\cdot}$	-0.000039
	FSKP	+0.000048	+0.001996***	+0.000391	+0.000988***	+0.003323*
	SKP	+0.000702***	+0.002047***	+0.001822***	+0.001355***	+0.001367*

Table 4: Mean paired log-likelihood difference relative to GPT. Positive values favour the reported model. * $p < .05$, ** $p < .01$, *** $p < .001$, $^{\cdot}$ $p < .1$. The estimations are based on cross-validation. The significance is based on paired t-test.