Probing for Reading Times

Eleftheria Tsipidi[1] **Samuel Kiegeland**[1] **Francesco Ignazio Re**[1] **Tianyang Xu**[2]
Mario Giulianelli[3] **Karolina Stańczak**[1] **Ryan Cotterell**[1]
[1]ETH Zürich [2]Toyota Technological Institute at Chicago [3]University College London
{eleftheria.tsipidi, francesco.re, ryan.cotterell}@inf.ethz.ch
samuel.kiegeland@gmail.com sallyxu@ttic.edu
m.giulianelli@ucl.ac.uk karolinaewa.stanczak@ai.ethz.ch

Abstract

Probing has shown that language model representations encode rich linguistic information, but it remains unclear whether they also capture cognitive signals about human processing. In this work, we probe language model representations for human reading times. Using regularized linear regression on two eye-tracking corpora spanning five languages (English, Greek, Hebrew, Russian, and Turkish), we compare the representations from every model layer against scalar predictors—surprisal, information value, and logit-lens surprisal. We find that the representations from early layers outperform surprisal in predicting early-pass measures such as first fixation and gaze duration. The concentration of predictive power in the early layers suggests that human-like processing signatures are captured by low-level structural or lexical representations, pointing to a functional alignment between model depth and the temporal stages of human reading. In contrast, for late-pass measures such as total reading time, scalar surprisal remains superior, despite its being a much more compressed representation. We also observe performance gains when using both surprisal and early-layer representations. Overall, we find that the best-performing predictor varies strongly depending on the language and eye-tracking measure.

https://github.com/rycolab/llm-representations-rt

1 Introduction

How long a reader's eyes linger on a linguistic unit is posited to reflect the cognitive effort required to process it (Just and Carpenter, 1980; Rayner, 1998). One prominent way to measure these durations is eye-tracking, which records fixation times at fine temporal resolution. A key question in psycholinguistics is which textual features best predict these reading times, and the predictive fit of a feature set serves as a measure of its *psychometric power* (Smith and Levy, 2013). To date, the most successful neural language model-based predictor has been surprisal (Hale, 2001; Levy, 2008; Wilcox et al., 2023).

Independently, a large body of work on *probing* has demonstrated that the internal representations of neural language models encode a wealth of linguistic information, including syntactic structure, morphological features, and semantic properties (Alain and Bengio, 2017; White et al., 2021; Immer et al., 2022; Kim et al., 2025). Yet probing studies have overwhelmingly focused on predicting properties of the linguistic signal itself from representations. While recent work has shown that language model representations align with neural signals measured via fMRI and EEG (Schrimpf et al., 2021; Caucheteux and King, 2022), it remains unclear to what extent the language model's internal representations can directly predict *behavioral* reading times—the fine-grained, unit-level processing effort that readers expend, as reflected in eye-tracking measures.

In this work, we probe language model representations for human reading times. Using regularized linear regression, we predict unit-level reading times directly from the neural language models' representations extracted at every layer of a language model. We compare these representation-based predictors against scalar baselines—surprisal, information value (Giulianelli et al., 2024b), and logit-lens surprisal (nostalgebraist, 2020; Kuribayashi et al., 2025)—which compress the model's internal state into a single dimension. An illustration of this predictive task is provided in Figure 1, which shows the true aggregated unit-by-unit gaze duration of human readers and the gaze duration predicted by various predictor variables. We conduct our evaluation on two eye-tracking corpora, Provo (Luke and Christianson, 2018) and MECO (Siegelman et al., 2022), spanning five languages: English, Greek, Hebrew, Russian, and Turkish, using mGPT (Shli-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12618–12642

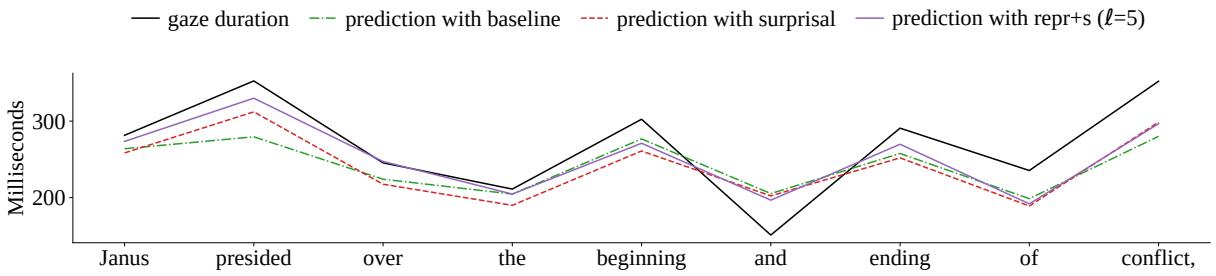

Figure 1: Gaze duration and its prediction by different mGPT-derived feature settings. The excerpt is from a document in the MECO dataset. The y-axis represents reading time measured in milliseconds. True gaze duration is represented by a black line. The purple line represents the predictions of a linear model trained on 5th-layer representations and standard surprisal. Note how the gaze duration and its predictions spike on units with high information content, such as *presided* and *conflict*.

azhko et al., 2024), GPT-2 (Radford et al., 2019), and cosmosGPT (Kesgin et al., 2024). We evaluate the predictive power of representations from all layers for three reading time metrics: first fixation duration, gaze duration, and total reading time.

Our results reveal clear differences across reading time modalities. In English, representations from early layers tend to outperform surprisal in predicting early-pass measures, such as first fixation duration and gaze duration, suggesting that features relevant to initial lexical access and local structural encoding are accessible in internal states beyond what surprisal captures. In contrast, for late-pass measures such as total reading time, scalar predictors, especially surprisal and logit-lens surprisal, are often competitive with or superior to high-dimensional representations. We also observe substantial cross-lingual variation in the relative predictive power of scalar and representation-based predictors. In Greek, Hebrew, Russian, and Turkish, scalar predictors are frequently as strong as or stronger than representations, depending on the eye-tracking measure. We further find that combining surprisal with layer-wise representations frequently improves predictive performance over representations alone, although the gains over scalar baselines are less consistent. Overall, our findings show that the psychometric power of language models depends strongly on the reading-time measure, the model layer, and the language under study, rather than being captured by a single predictor across all settings.

2 Preliminaries

Language Models. We adopt the formulation of Kiegeland et al. (2026), who distinguish the abstract linguistic **units** that humans process, over which reading times are modeled, and **symbols**, which the language model outputs. Throughout this section, we present surprisal theory and our predictors in terms of units. We discuss how to reconcile this formulation with a language model defined over tokens in § 5.1. Let U be a countable set of units. A **string** $\boldsymbol{u} = u_1 \ldots u_T$ is a finite sequence of units $u_t \in U$. We write $\boldsymbol{u}_{<t} = u_1 \cdots u_{t-1}$ for the prefix of $\boldsymbol{u}$ up to but not including position t. We denote string concatenation by juxtaposition, i.e., $\boldsymbol{u}\boldsymbol{u}'$ denotes the concatenation of $\boldsymbol{u}$ and $\boldsymbol{u}'$. With U^*, we denote the Kleene closure of U, i.e., the set of all finite strings over U. A **language model** is a probability distribution p over U^*. Every language model p induces a **prefix probability**, defined as

$$\overrightarrow{p}(\boldsymbol{u}) \stackrel{\text{def}}{=} \sum_{\boldsymbol{u}' \in U^*} p(\boldsymbol{u}\boldsymbol{u}'). \tag{1}$$

Then, define the **conditional prefix probability** as

$$\overrightarrow{p}(u \mid \boldsymbol{u}) \stackrel{\text{def}}{=} \frac{\overrightarrow{p}(\boldsymbol{u}u)}{\overrightarrow{p}(\boldsymbol{u})}. \tag{2}$$

By the chain rule of probability, the language model factorizes autoregressively as

$$p(\boldsymbol{u}) = \overrightarrow{p}(\text{EOS} \mid \boldsymbol{u}) \prod_{t=1}^{T} \overrightarrow{p}(u_t \mid \boldsymbol{u}_{<t}), \tag{3}$$

where EOS is a distinguished end-of-string symbol. Let $\overline{U} \stackrel{\text{def}}{=} U \cup \{\text{EOS}\}$.

Neural Language Models. Modern language models, such as those based on the transformer architecture (Vaswani et al., 2017), parameterize the conditional distributions above through a stack of L layers. The input layer maps each symbol $u \in \overline{U}$

to a vector $\mathbf{h}_0(u) \in \mathbb{R}^D$, and each subsequent layer computes a representation as a function of the previous layer's representations. Let $\boldsymbol{u} = u_1 \cdots u_T$ be a string over $\overline{U}$, then for each layer $\ell \in \{1, ..., L\}$, we define

$$\mathbf{h}_\ell(\boldsymbol{u}) \stackrel{\text{def}}{=} f_\ell\left(\mathbf{h}_{\ell-1}(u_1), ..., \mathbf{h}_{\ell-1}(u_T)\right), \quad (4)$$

where $\mathbf{h}_\ell(\boldsymbol{u})$ denotes the ℓ-layer representation at the final unit position T and $f_\ell \colon (\mathbb{R}^D)^* \to \mathbb{R}^D$ denotes the transformation at layer ℓ. The last layer representation is then projected onto $\Delta(\overline{U})$ as

$$\overrightarrow{p}(\cdot \mid \boldsymbol{u}) = \text{softmax}\left(\mathbf{W} g(\mathbf{h}_L(\boldsymbol{u})) + \mathbf{b}\right), \quad (5)$$

where $g \colon \mathbb{R}^D \to \mathbb{R}^D$ is a final (non-linear) transformation applied before the linear projection (e.g., layer normalization), $\mathbf{W} \in \mathbb{R}^{(|\overline{U}|) \times D}$ is the projection matrix, and $\mathbf{b} \in \mathbb{R}^{|\overline{U}|}$ is a bias term. The final representation $\mathbf{h}_L$ is directly used to compute the distribution over the next symbols.[1] However, the intermediate representations $\mathbf{h}_1(\boldsymbol{u}), ..., \mathbf{h}_{L-1}(\boldsymbol{u})$ encode linguistic information themselves (Alain and Bengio, 2017; Immer et al., 2022). In this work, we investigate whether these representations also encode information predictive of human reading behavior.

3 Psychometric Data

In this work, we study how well we can predict real-valued measurements of human processing effort collected during natural reading. Formally, for a unit $u_t \in \overline{U}$ read in context $\boldsymbol{u}_{<t} \in U^*$, we observe a reading time $r(u_t, \boldsymbol{u}_{<t}) \in \mathbb{R}$; when $u_t = \text{EOS}$, this corresponds to utterance-final **wrap-up** cost (Rayner et al., 2000; Meister et al., 2022), reflecting the additional processing cost associated with integrating the full utterance. Eye-tracking experiments yield several such measurements per unit, corresponding to different stages of processing: first-fixation duration (the duration of the initial fixation), gaze duration (the sum of all fixations before the eyes leave the unit), and total reading time (the sum of all fixations including regressions). Our goal is to predict these reading times from features derived from a language model.

[1] Parameterized LMs typically include a special end-of-sequence token, for which representations can be computed and on which next-token predictions can in principle be conditioned. From the perspective of an LM as a distribution over strings, however, conditioning on EOS is not well defined. In this work, EOS may only appear as the final unit of a string $\boldsymbol{u}$, where it is used to model wrap-up effects (see § 3).

3.1 Previously Proposed Predictors

We now discuss three previously proposed *scalar* predictors of human reading time.

Surprisal Theory. Surprisal theory (Hale, 2001; Levy, 2008) posits that reading times are an affine function of **surprisal**, the negative log-probability of a unit under the reader's implicit language model. Formally, let p_H denote the **human language model**—the probability distribution that characterizes a reader's expectations over upcoming linguistic material. The **surprisal** of unit u_t in context $\boldsymbol{u}_{<t}$ is then

$$s(u_t, \boldsymbol{u}_{<t}) \stackrel{\text{def}}{=} -\log p_\text{H}(u_t \mid \boldsymbol{u}_{<t}), \quad (6)$$

and the theory predicts that reading time is an affine function of this quantity (Smith and Levy, 2013; Shain et al., 2024). Since p_H is not directly observable, it is standard practice to approximate it with a trained language model p, and empirical support for the resulting predictions has been found across diverse datasets and languages (Wilcox et al., 2023). Note that the predictive power of p-derived surprisal depends on how well p approximates p_H, and this approximation quality likely varies across languages and models.

Information Value. Shannon surprisal is the standard metric for quantifying the unexpectedness of a linguistic unit under a model p, but other operationalizations of information content exist; for an overview, see Giulianelli et al. (2024b). In this paper, we include next-unit information value in addition to standard surprisal. Next-unit information value measures the expected distance between the observed next unit u_t and alternative continuations $u \in \overline{U}$ sampled from the model's predictive distribution. This corresponds to a special case of the general string-level formulation of information value, where continuations are restricted to a single unit (cf. Giulianelli et al., 2023, 2024b). Formally, it is defined as

$$v(u_t, \boldsymbol{u}_{<t}) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{u \sim \overrightarrow{p}(\cdot \mid \boldsymbol{u}_{<t})} \left[\text{d}(u_t, u)\right], \quad (7)$$

where $\text{d} \colon \overline{U} \times \overline{U} \to \mathbb{R}_{\geq 0}$ is a distance function, typically operationalized as the cosine distance between the contextual representations $\mathbf{h}_\ell(\boldsymbol{u}_{<t} u_t)$ and $\mathbf{h}_\ell(\boldsymbol{u}_{<t} u)$ at a given layer ℓ (Giulianelli et al., 2024b, 2026). This makes information value a natural point of comparison for our representation-based predictors.

Logit Lens. Standard surprisal is computed from the final layer's next-token distribution. The **logit lens** (nostalgebraist, 2020; Kuribayashi et al., 2025) asks what distribution an intermediate layer would induce if its representation were fed directly to the output head. Concretely, it applies the *same* projection matrix $\mathbf{W}$, bias $\mathbf{b}$, and layer normalization LN that are used after the final layer to the representation of an earlier layer ℓ:

$$q_\ell(\cdot \mid \boldsymbol{u}) \stackrel{\text{def}}{=} \text{softmax}\left(\mathbf{W}\mathbf{h}_\ell(\boldsymbol{u}) + \mathbf{b}\right).^{2} \quad (8)$$

Because $\mathbf{W}$ and $\mathbf{b}$ are estimated only to decode the final layer's representation, there is no guarantee that this projection yields a meaningful distribution at earlier layers—the intermediate representations may not be linearly decodable in vocabulary space. In practice, however, the logit lens has been found to produce interpretable predictions at many layers (nostalgebraist, 2020). We define the **logit-lens surprisal** s^{LL} at layer ℓ as

$$s^{\text{LL}}_\ell(u_t, \boldsymbol{u}_{<t}) \stackrel{\text{def}}{=} -\log q_\ell(u_t \mid \boldsymbol{u}_{<t}). \quad (9)$$

3.2 The Limitations of Scalar Predictors

All three predictors introduced above—surprisal, information value, and logit-lens surprisal—share a fundamental limitation: each compresses a representation extracted from a language model into a single scalar. While such scalar predictors have served as useful proxies for human processing effort, it is natural to suspect that using the entire representation as a predictor may be more useful. Moreover, in the case of surprisal, larger models that achieve lower cross-entropy on held-out text often evince poorer fit to human reading times (Oh and Schuler, 2023; Kuribayashi et al., 2024), and recent fine-grained modeling suggests that much of the variance typically attributed to surprisal may instead be explained by skip rates (Re et al., 2025). Finally, recent evidence also indicates that the internal layers of language models contain representations that align more closely with human behavioral and neural signals than any single scalar derived from them (Schrimpf et al., 2021; Caucheteux and King, 2022; Kuribayashi et al., 2025). Taken together, this suggests that scalar compression—whether through surprisal (which reduces the final layer to a log-probability), information value (which summarizes representational distance as a single expectation), or logit-lens surprisal (which projects an intermediate layer through the output head)—discards much of the psychometrically relevant information contained in the model's internal representations.

4 Methods

To evaluate whether the representations induced by neural language models serve as useful predictors of human processing effort, we apply various forms of regularized linear regression. By controlling for standard psycholinguistic factors, we compare the predictive power of representations, information value, standard surprisal, and layer-wise surprisal.

4.1 Linear Regression

To predict reading times, we follow standard psycholinguistic practices and use linear models (Goodkind and Bicknell, 2018; Wilcox et al., 2020). Formally, let $r(u_t, \boldsymbol{u}_{<t}) \in \mathbb{R}$ be a real-valued reading time measurement for a unit $u_t \in \overline{U}$ in context $\boldsymbol{u}_{<t} \in U^*$, and let $\mathbf{x}(u_t, \boldsymbol{u}_{<t}) \in \mathbb{R}^D$ be a column vector of predictor variables.[3] We predict reading times as

$$\widehat{r_{\boldsymbol{\beta}}}(u_t, \boldsymbol{u}_{<t}) \stackrel{\text{def}}{=} \mathbf{x}(u_t, \boldsymbol{u}_{<t})^\top \boldsymbol{\beta}, \quad (10)$$

where $\boldsymbol{\beta} \in \mathbb{R}^D$ is a parameter vector. Let the corpus consist of N strings $\boldsymbol{u}^{(1)}, ..., \boldsymbol{u}^{(N)}$, where string $\boldsymbol{u}^{(n)}$ has $T^{(n)}$ units plus EOS. We estimate $\boldsymbol{\beta}$ by minimizing the per-string squared loss:

$$L_n(\boldsymbol{\beta}) \stackrel{\text{def}}{=} \sum_{t=1}^{T^{(n)}} \left(r(u_t^{(n)}, \boldsymbol{u}_{<t}^{(n)}) - \widehat{r_{\boldsymbol{\beta}}}(u_t^{(n)}, \boldsymbol{u}_{<t}^{(n)})\right)^2 + \left(r(\text{EOS}, \boldsymbol{u}^{(n)}) - \widehat{r_{\boldsymbol{\beta}}}(\text{EOS}, \boldsymbol{u}^{(n)})\right)^2. \quad (11)$$

Ordinary least squares estimates $\boldsymbol{\beta}$ by minimizing $L(\boldsymbol{\beta}) \stackrel{\text{def}}{=} \sum_{n=1}^{N} L_n(\boldsymbol{\beta})$. Following Wilcox et al. (2023) and Opedal et al. (2024), we do not apply any transformation (e.g., log or z-score) to the reading times before fitting the model, so that $\widehat{r_{\boldsymbol{\beta}}}$ is directly interpretable in milliseconds.

Regularized Linear Regression. We also consider regularized variants. Ridge regression adds an $\|\cdot\|_2^2$ penalty:

$$L_{\text{R}}(\boldsymbol{\beta}) \stackrel{\text{def}}{=} L(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_2^2, \quad (12)$$

[2] Even at the final layer $\ell = L$, the logit-lens distribution q_L need not equal the model's true output distribution $\overrightarrow{p}$, because Eq. (5) includes the non-linear transformation g (e.g., layer normalization) which is absent from the logit lens.

[3] All vectors in this paper are column vectors.

where $\lambda \geq 0$ controls the strength of regularization. LASSO regression instead uses an $\|\cdot\|_1$ penalty:

$$L_{\mathrm{L}}(\boldsymbol{\beta}) \stackrel{\text{def}}{=} L(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1. \tag{13}$$

In contrast to ridge regression, LASSO encourages sparsity in $\boldsymbol{\beta}$, inducing sparse solutions and acting as a form of feature selection.

Tuning. We tune the regression models by selecting (i) whether to apply regularization and, if so, (ii) whether to use LASSO or ridge regression, along with (iii) the corresponding penalty weight. Model selection is performed using the test **mean squared error** (MSE) on a fixed train–test split:

$$\mathrm{MSE}(\boldsymbol{\beta}) \stackrel{\text{def}}{=} \frac{L(\boldsymbol{\beta})}{\sum_{n=1}^{N}(T^{(n)}+1)}. \tag{14}$$

To avoid leakage, the documents used in this tuning test split (5 documents in Provo and 2 documents in MECO) are excluded from all subsequent experiments. We evaluate penalty weights in the range $[0.001, 10]$, performing hyperparameter selection independently for each predictor type, layer, and dependent variable. This procedure is applied to the baseline and surprisal models, and to each layer-wise instance (layers 1–24 for mGPT and 1–12 for GPT-2 and cosmosGPT) of information value, logit lens, and representation predictors.

Cross-Validation. We evaluate each combination of predictor type and reading time measure using 10-fold cross-validation, run separately on Provo and on each language subset of MECO.

5 Experimental Setup

5.1 Feature Estimation

Models. We use surprisal estimates from mGPT (Shliazhko et al., 2024), a multilingual model based on the GPT-3 (Brown et al., 2020) architecture. mGPT was trained on 61 languages from 25 language families, which enables us to experiment on both Provo (Luke and Christianson, 2018) and the MECO (Siegelman et al., 2022) data. It has 24 layers, each with an embedding dimension of 2048. For additional experiments, we use two monolingual models: the English monolingual GPT-2 Small (Radford et al., 2019) on the Provo data and the English subset of MECO; and the Turkish cosmosGPT (Kesgin et al., 2024) on the Turkish subset of MECO. Both monolingual models have 12 layers with an embedding dimension of 768.

From Tokens to Units. Let p_Σ denote the token-level[4] language model: a probability distribution over token strings Σ^*, where Σ is a finite token alphabet, and $\varphi\colon U^* \to \Sigma^*$ is a function that maps a unit string to a token string. We assume that φ respects unit boundaries: no token spans two units, so any tokenization decomposes as $\varphi(u_1 \cdots u_T) = \boldsymbol{\sigma}_1 \cdots \boldsymbol{\sigma}_T$, where $\boldsymbol{\sigma}_t = \sigma_{t,1} \cdots \sigma_{t,n_t}$ is the token sequence corresponding to unit u_t.[5] Following standard practice (Wilcox et al., 2023), we define unit-level surprisal and logit-lens surprisal as the sum over $\boldsymbol{\sigma}_t$ of the per-token surprisals under p_Σ and q_ℓ (Eq. (8)), respectively.[6] Unlike Kuribayashi et al. (2025), we do not include tuned lens (Belrose et al., 2025), since mGPT does not have a pre-trained tuned-lens model; we include logit-lens surprisal of the last layer, as it may differ from standard surprisal.[7]

Writing $\mathbf{h}_\ell(\boldsymbol{\sigma})$ for the layer-ℓ hidden state of p_Σ at the final token of a token string $\boldsymbol{\sigma} \in \Sigma^*$, we define the unit-level representation of $\boldsymbol{u}_{<t}u_t$ as the mean over the n_t tokens that correspond to u_t:

$$\mathbf{h}_\ell(\boldsymbol{u}_{<t}u_t) \stackrel{\text{def}}{=} n_t^{-1} \sum_{k=M-n_t+1}^{M} \mathbf{h}_\ell(\varphi(\boldsymbol{u}_{<t}u_t)_{\leq k}) \tag{15}$$

Mean-pooling is one of several possible aggregations; we discuss this in the limitations section. For information value, Eq. (7) then applies with d computed as the cosine distance between these pooled representations, and we approximate the expectation by Monte Carlo with $k = 50$ continuations sampled from p_Σ.[8]

5.2 Data

We use reading time data from two commonly used corpora in psycholinguistics. The Provo corpus (Luke and Christianson, 2018) is a dataset of eye-movement behavior comprising eye-tracking recordings from 84 native English readers as they

[4] See Gastaldi et al. (2025) and Vieira et al. (2025) for a formal treatment of tokenized and token-level language models.

[5] For the whitespace-delimited words used in our corpora and the token alphabets of mGPT, GPT-2, and cosmosGPT, this holds in practice: each token sits within a single word.

[6] See Giulianelli et al. (2024a), Pimentel and Meister (2024), Oh and Schuler (2024), and Kiegeland et al. (2026) for discussion of this choice and alternative approaches to calculating unit-level surprisal.

[7] The Hugging Face documentation states that some models do in fact apply a function g or further processing to the last state when it is returned; this could affect surprisal and render it different from the corresponding final layer logit lens.

[8] For each continuation, we generate tokens until an end-of-unit marker is encountered, or 3 tokens have been produced.

Figure 2: MSE for baseline, surprisal, representations, information value, and logit-lens surprisal on the Provo and MECO data across the 24 layers of mGPT and eye-tracking measures.

read 55 short English passages drawn from a range of fiction and nonfiction sources. The Multilingual Eye Movement Corpus (MECO; Siegelman et al., 2022) is a large multilingual corpus with eye movement data from L1 speakers reading 12 Wikipedia-style passages in 13 languages. To capture a variety of different language families, we select the English, Greek, Hebrew, Russian, and Turkish portions of the dataset for our experiments.

Data Preprocessing. We quantify the unit-by-unit reading time, using three standard eye-tracking measures: **first fixation duration**, the duration of the first fixation on a unit during first pass reading; **gaze duration**, the sum of all consecutive fixations on the unit from first entry, until a fixation leaves it for the first time; and **total reading time**, the sum of all fixations on the unit across the entire trial, including any later re-reading due to regressions. In line with established psycholinguistic theory, data collected during eye-tracking experiments can be divided into early-pass and late-pass measures (Rayner and Fischer, 1996). First fixation and gaze duration are considered early-pass measures, as they are primarily sensitive to the initial stages of unit recognition and lexical access (Rayner, 1998; Cook and Wei, 2019). Specifically, first

fixation duration is viewed as a marker of initial orthographic and phonetic activation (Rayner, 2009). Conversely, total reading time is a late-pass measure, which is interpreted as a marker of higher-level post-lexical processing, reflecting the cognitive effort required for syntactic integration, discourse-level comprehension, and the resolution of processing difficulties (Clifton et al., 2007).

6 Results

We now present results with surprisal, representations, information value, and logit lens computed using mGPT. For results using GPT-2 Small and cosmosGPT, see App. C.

6.1 Predictive Power of Representations

Figure 2 and Table 1 compare predictive power across layers and eye-tracking measures for Provo. Performance is highest for early-layer representations (1–10), declines in intermediate layers, and recovers at the final layer. The best representation layer is comparable to or stronger than surprisal, though surprisal sometimes wins despite its much lower dimensionality. Combining representations with surprisal improves over either alone; these gains tend to be significant over representations but not over scalars (App. B).

6.2 Information Value and Logit Lens

Figure 2 and Table 1 compare the predictive power of information value and logit-lens surprisal against model representations and surprisal. Overall, these predictors show less variability across layers compared to model representations. For first fixation duration and total reading time, we find that information value is more predictive when computed from the early to intermediate layers than from the later ones. In contrast, logit-lens surprisal tends to perform best in the last layers.

6.3 Predictive Power across Languages

To assess the predictive power of representations across languages, we repeat the experiments for five languages from the MECO dataset: English, Greek, Hebrew, Russian, and Turkish. Consistent with the results for Provo, Figure 2 and Table 1 show that representations from early layers are the most predictive, with performance decreasing in intermediate layers before recovering in the final layer. An exception to this is gaze duration and total reading time for Greek, as well as total reading time for Turkish, where the representations from the last layer perform best. Similarly, we find that information value is most predictive at early and intermediate layers, while logit-lens surprisal has the highest predictive power at later layers. Moreover, we observe that Russian and Turkish are the only languages where the combination of representation and surprisal (Table 2) is not the best predictor for first-fixation duration and gaze duration. Overall, we find that the most predictive predictor varies strongly depending on the language and eye-tracking measure.

6.4 Permuting Reading Times

To further test whether the observed predictive performance reflects a meaningful relationship between the predictors derived from language models and the reading time data, we conduct a permutation test on the dependent variable. Specifically, we refit the same models from § 6.1, on training sets with their reading times randomly permuted across words, but without permuting the reading times of the respective validation sets. There are similarities to non-permuted trends across layers for representations; however, mean squared error scores are higher across the board, and differences between the various sets of predictors tend to be smaller (App. E).

6.5 Linear Mixed-Effects Models

We extend Eq. (10) to account for subject- and document-level variability by fitting linear mixed-effects models (LMMs) with random intercepts for participants and documents on the MECO data; for high-dimensional representations, we first reduce the representations to 25 principal components. We selected the number of principal components using the scree-plot elbow criterion (Cattell, 1966). The results (Table 9) are broadly consistent with our main analyses, and significance over permuted reading times is preserved. At the same threshold ($\alpha = 0.001$), however, fewer predictors significantly outperform the baseline. This may follow from LMMs yielding more conservative effect estimates: by explicitly modeling reader and document variability, they isolate the predictor's contribution net of these sources, which simpler models can conflate into the effect itself.

7 Discussion

We now turn to discussing our results and their main implications. Across predictors (representations, information value, and logit-lens surprisal),

Measure	Surprisal	Best $\mathbf{h}$ (ℓ)	Best v (ℓ)	Best s^{LL} (ℓ)
		Provo—English		
FFD	$-2.28_{4.55}{}^{*}$	$\mathbf{-17.92}_{8.76}{}^{*\bullet}$ **(2)**	$-7.13_{7.02}{}^{*}$ (1)	$-2.31_{4.54}{}^{*}$ (24)
GD	$-39.96_{34.34}{}^{*}$	$\mathbf{-57.21}_{31.77}{}^{*}$ **(2)**	$-23.62_{19.43}{}^{*}$ (24)	$-41.34_{35.16}{}^{*}$ (24)
TRT	$-290.61_{134.19}{}^{*\bullet}$	$-198.09_{196.67}{}^{*\bullet}$ (2)	$-189.52_{100.75}{}^{*\bullet}$ (10)	$\mathbf{-298.87}_{135.03}{}^{*\bullet}$ **(24)**
		MECO—English		
FFD	$-11.09_{18.93}{}^{*}$	$\mathbf{-27.74}_{29.25}{}^{*}$ **(2)**	$-9.91_{20.48}{}^{*}$ (11)	$-11.29_{18.60}{}^{*}$ (23)
GD	$-50.98_{120.48}{}^{*}$	$\mathbf{-139.92}_{169.75}{}^{*\bullet}$ **(5)**	$-44.78_{33.77}{}^{*\bullet}$ (10)	$-97.28_{112.91}{}^{*}$ (24)
TRT	$-605.14_{433.80}{}^{*\bullet}$	$\mathbf{-914.85}_{746.53}{}^{*\bullet}$ **(3)**	$0.00_{0.00}{}^{*}$ (1)	$-626.49_{441.29}{}^{*\bullet}$ (24)
		MECO—Greek		
FFD	$-3.75_{9.13}{}^{*}$	$-4.72_{22.70}{}^{*}$ (2)	$\mathbf{-6.10}_{10.71}{}^{*}$ **(1)**	$-4.67_{10.32}{}^{*}$ (23)
GD	$\mathbf{-275.80}_{212.98}{}^{*\bullet}$	$-158.94_{306.51}{}^{*}$ (24)	$-32.10_{38.16}{}^{*}$ (5)	$-268.69_{192.38}{}^{*\bullet}$ (23)
TRT	$-1079.68_{944.67}{}^{*\bullet}$	$-914.45_{1072.60}{}^{*\bullet}$ (24)	$-145.36_{152.23}{}^{*\bullet}$ (6)	$\mathbf{-1080.86}_{939.97}{}^{*\bullet}$ **(24)**
		MECO—Hebrew		
FFD	$-5.04_{5.08}{}^{*}$	$\mathbf{-18.31}_{24.28}{}^{*}$ **(2)**	$0.00_{0.00}{}^{*}$ (1)	$-4.84_{4.68}{}^{*}$ (24)
GD	$-50.97_{58.09}{}^{*}$	$-20.96_{114.62}{}^{*}$ (1)	$-53.54_{55.93}{}^{*}$ (3)	$\mathbf{-64.09}_{59.38}{}^{*\bullet}$ **(23)**
TRT	$-621.84_{786.76}{}^{*\bullet}$	$\mathbf{-1431.52}_{2384.96}{}^{*\bullet}$ **(2)**	$-53.61_{32.12}{}^{*\bullet}$ (24)	$-610.27_{781.71}{}^{*\bullet}$ (24)
		MECO—Russian		
FFD	$-18.74_{22.90}{}^{\bullet}$	$-21.86_{28.87}{}^{*}$ (5)	$-22.31_{26.95}{}^{*}$ (2)	$\mathbf{-28.75}_{27.04}{}^{*\bullet}$ **(8)**
GD	$-88.70_{110.92}{}^{*\bullet}$	$-3.59_{101.50}{}^{*}$ (2)	$-36.36_{38.45}{}^{*}$ (7)	$\mathbf{-89.84}_{102.91}{}^{*\bullet}$ **(23)**
TRT	$\mathbf{-619.68}_{741.79}{}^{*\bullet}$	$-438.41_{605.05}{}^{*\bullet}$ (24)	$-153.04_{149.30}{}^{*}$ (7)	$-615.77_{735.20}{}^{*\bullet}$ (24)
		MECO—Turkish		
FFD	$-2.63_{6.62}$	$3.67_{29.35}{}^{*}$ (2)	$-3.79_{9.74}$ (12)	$\mathbf{-5.21}_{6.39}{}^{*}$ **(19)**
GD	$-186.91_{119.24}{}^{*\bullet}$	$14.06_{167.93}{}^{*}$ (24)	$-78.60_{47.64}{}^{*\bullet}$ (11)	$\mathbf{-187.66}_{118.60}{}^{*\bullet}$ **(23)**
TRT	$-1642.88_{1410.20}{}^{*\bullet}$	$-973.47_{1871.74}{}^{*\bullet}$ (3)	$-403.02_{217.54}{}^{*\bullet}$ (6)	$\mathbf{-1644.42}_{1409.93}{}^{*\bullet}$ **(24)**

Table 1: Δ_{MSE} (baseline–target) of ten-fold cross-validation for models trained on baseline features and mGPT-derived surprisal, representations ($\mathbf{h}$), information value (v), and logit-lens surprisal (s^{LL}) on the Provo and MECO data across the 24 layers of mGPT and eye-tracking measures. For each measure, we report the lowest MSE over layers and the corresponding layer index ℓ. Bold indicates the best condition per row. Asterisks (*) denote models that significantly outperform the respective models trained on permuted reading times, according to a one-sided paired t-test ($\alpha = 0.001$). Similarly, bullets ($\bullet$) indicate significance over the baseline.

we find substantial differences in their predictive power when extracted from different model layers.

Opposite Trends for Logit Lens. Kuribayashi et al. (2025) report that logit-lens surprisal performs better at earlier layers for larger models and at later layers for smaller ones. We corroborate this for mGPT (1.3B) and GPT-2 (117M), where logit-lens surprisal peaks at later layers. However, representations show the opposite pattern—higher predictive power at earlier layers—indicating that psychometric predictive power depends on the choice of predictor, not just model size. This contrast should be interpreted with care: representations are evaluated via linear regression on raw vectors, whereas logit-lens predictors undergo additional transformations (layer normalization, projection into vocabulary space), so the difference may partly reflect accessibility to a linear decoder rather than representational content alone. The superiority of early-layer representations for early-pass measures is consistent with the view that initial reading stages rely on lexical and morpho-syntactic features preserved in earlier layers.[9]

Cross-Lingual Differences. On the MECO dataset, the performance of the different settings on first fixation duration is generally consistent between languages, as well as with the Provo data. However, we do not observe the same consistency across languages for gaze duration or total reading time, despite the fact that the non-baseline predictors are all derived from mGPT.[10] In fact, Kurib-

[9] More broadly, this pattern suggests a functional alignment between the layer hierarchy of the transformer and the temporal stages of human reading—early layers correspond to early-pass measures, while later layers correspond to late-pass measures—reminiscent of the depth–time correspondence observed in neuroscience (Caucheteux and King, 2022).

[10] Shliazhko et al. (2024) report worse downstream performance for mGPT on Greek and Turkish, while Arnett and Bergen (2025) find that differences in downstream performance of language models seem to be caused by data quantity disparities, rather than model architecture.

ayashi et al. (2025) also observed mixed results in multilingual settings, possibly due to latent effects of English on the processing of the target language (Wendler et al., 2024). We conducted experiments with a monolingual Turkish model to investigate this possibility (App. C); yet, we saw no discernible difference to the mGPT results for Turkish. In a qualitative example of a particular clause in different languages (App. D), we observed some intra-lingual patterns that seem to modulate reading times: Greek seems more verbose and more likely to begin a structural unit with lower information content, because it is a language that favors the use of articles (e.g., even ahead of proper names). Turkish word order places verbs at the end of clauses. However, it is unclear if there is a connection between these patterns and the differences in the performance of representations on reading times.

Variance of Predictors across Layers. Finally, the validation MSE of both information value and logit-lens surprisal varies less across layers than representation-based predictors. One plausible reason is that they compress each layer's state into a single scalar, whereas representations expose a high-dimensional feature whose usefulness can change more sharply with layer depth.

Future Work. While this work provides a controlled comparison between several language-model-derived representations across layers and reading time measures, it leaves possible extensions for analyzing internal representations using dimensionality reduction or feature selection techniques. In particular, kernel principal component analysis could be used to map representations into a non-linear feature space prior to probing, allowing us to assess whether reading-time-relevant structure is present but not linearly accessible in the original representations. Comparing layer-wise performance before and after such transformations would help disentangle representational content from linear decodability and clarify whether some of the observed trends are due to differences in how easily the information can be recovered by the probe, rather than differences in the information itself. Moreover, our experiments are limited to mGPT, GPT-2, and cosmosGPT, and thus could be extended to other monolingual and/or larger language models. Finally, we believe that the diverging performance of representations and logit-lens surprisal across layers offers interesting avenues to test with combinations of different predictors.

8 Conclusion

In this work, we investigated the psychometric power of language models by revisiting the hypothesis that reading times are best predicted by the scalar measure of surprisal by testing whether a neural language model's internal representations serve as more accurate predictors of human processing effort. While controlling for standard psycholinguistic factors, we compared the predictive power of language model representations, information value, and layer-wise surprisal. Across two eye-tracking corpora and five typologically distinct languages, we identify differences across reading time modalities: early-layer representations of mGPT and GPT-2 are superior at predicting early-pass measures (first fixation and gaze duration), while scalar surprisal remains superior for late-pass measures (total reading time). These results suggest that some psychometric power of language models is encoded within their internal representations beyond what surprisal captures. Notably, language model representations show more variability across layers compared to information value and logit-lens surprisal. Finally, we find that the most effective predictor varies across the languages and eye-tracking measures analyzed.

Limitations

Our study is subject to several limitations. First, our analysis is restricted to eye-tracking data. While these measures provide high-fidelity temporal markers of reading effort, future work is needed to determine if the observed patterns generalize to other modalities, such as Self-Paced Reading or neuroimaging (EEG/fMRI). Second, computational constraints limited our evaluation to models up to 1.3B parameters (mGPT). It remains unclear whether the observed functional divergence between representations and surprisal persists or evolves in larger models with higher dimensionality. While prior work (Oh and Schuler, 2023; Shain et al., 2024; Kuribayashi et al., 2024) found that surprisal from smaller models often fits reading times better than that of larger language models, this scaling behavior may not generalize to a neural language model's internal representations. Next, we acknowledge that probing experiments can lead to false discoveries if random noise in representations is not properly accounted for (Méloux et al., 2025). While comparing our results against randomly initialized mGPT, GPT-2, and cosmos-

GPT baselines was not computationally feasible, we mitigate this concern through our permutation testing, which controls for random associations between predictors and reading times. Furthermore, we observe consistent trends across various experimental setups that random noise would not reproduce. However, we note that future research using randomly initialized model representations could examine our findings. In addition, our evaluation relies on raw MSE, which is scale-dependent: although this allows meaningful comparisons among predictors within the same eye-tracking measure, the magnitude of MSE differences should not be compared directly across first fixation duration, gaze duration, and total reading time, since these measures lie on different numerical scales. Finally, we use a language model's raw internal representations without experimenting with dimensionality reduction methods, except in the case of mixed-effects models (§ 6.5). Exploring lower-rank subspaces (e.g., via principal component analysis) could further isolate the specific features within the internal states that are accountable for the alignment with human processing effort. Relatedly, for multi-token units, we aggregate hidden states by mean-pooling; alternative aggregations (e.g., max-pool, first- or last-token pooling, or concatenation) may yield different results, and a systematic comparison is left to future work.

Ethics Statement

We foresee no ethical problems with our work.

Acknowledgments

We would like to thank Alex Warstadt for helpful discussions, and Taiga Someya and Andreas Opedal for pointing us to Kuribayashi et al. (2025). We also thank the anonymous reviewers for their useful comments, suggestions, and references to related work. Eleftheria Tsipidi was supported by the SNSF grant number 204667. Karolina Stańczak was supported by the ETH AI Center postdoctoral fellowship. We disclose the use of generative AI tools for light editing and rephrasing; the original text was our own, and we carefully reviewed all suggested edits.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations*.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the International Conference on Computational Linguistics*.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2025. Eliciting latent predictions from transformers with the tuned lens.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33.

Raymond B. Cattell. 1966. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276. PMID: 26828106.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1).

Charles Clifton, Adrian Staub, and Keith Rayner. 2007. Eye movements in reading words and sentences. In *Eye Movements*. Elsevier.

Anne E. Cook and Wencl Wei. 2019. What can eye movements tell us about higher level comprehension? *Vision*, 3(3).

Juan Luis Gastaldi, John Terilla, Luca Malagutti, Brian DuSell, Tim Vieira, and Ryan Cotterell. 2025. The foundations of tokenization: Statistical and computational concerns. In *The International Conference on Learning Representations*.

Mario Giulianelli, Luca Malagutti, Juan Luis Gastaldi, Brian DuSell, Tim Vieira, and Ryan Cotterell. 2024a. On the proper treatment of tokenization in psycholinguistics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Mario Giulianelli, Andreas Opedal, and Ryan Cotterell. 2024b. Generalized measures of anticipation and responsivity in online language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Mario Giulianelli, Sarenne Wallbridge, Ryan Cotterell, and Raquel Fernández. 2026. Incremental alternative sampling as a lens into the temporal and representational resolution of linguistic prediction. *Journal of Memory and Language*, 148.

Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. 2023. Information value: Measuring utterance predictability as distance from plausible alternatives. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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 the Workshop on Cognitive Modeling and Computational Linguistics*.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics*.

Alexander Immer, Lucas Torroba Hennigen, Vincent Fortuin, and Ryan Cotterell. 2022. Probing as quantifying inductive bias. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4).

H. Toprak Kesgin, M. Kaan Yuce, Eren Dogan, M. Egemen Uzun, Atahan Uz, H. Emre Seyrek, Ahmed Zeer, and M. Fatih Amasyali. 2024. Introducing cosmosGPT: Monolingual training for Turkish language models. In *International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*.

Samuel Kiegeland, Vésteinn Snæbjarnarson, Tim Vieira, and Ryan Cotterell. 2026. On the proper treatment of units in surprisal theory. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Junsol Kim, James Evans, and Aaron Schein. 2025. Linear representations of political perspective emerge in large language models. In *The International Conference on Learning Representations*.

Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. 2024. Psychometric predictive power of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *Transactions of the Association for Computational Linguistics*, 13.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3).

Steven G. Luke and Kiel Christianson. 2018. The Provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50.

Clara Meister, Tiago Pimentel, Thomas Clark, Ryan Cotterell, and Roger Levy. 2022. Analyzing wrap-up effects through an information-theoretic lens. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. 2025. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *The International Conference on Learning Representations*.

nostalgebraist. 2020. Interpreting GPT: The logit lens.

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.

Byung-Doh Oh and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Andreas Opedal, Eleanor Chodroff, Ryan Cotterell, and Ethan Wilcox. 2024. On the role of context in reading time prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

Keith Rayner. 2009. Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, 62(8).

Keith Rayner and Martin H. Fischer. 1996. Mindless reading revisited: Eye movements during reading and scanning are different. *Perception & Psychophysics*, 58(5).

Keith Rayner, Gretchen Kambe, and Susan A. Duffy. 2000. The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology Section A*, 53(4).

Francesco Ignazio Re, Andreas Opedal, Glib Manaiev, Mario Giulianelli, and Ryan Cotterell. 2025. A spatio-temporal point process for fine-grained modeling of reading behavior. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45).

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.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, and Daria Chernova. 2022. Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, 54(6).

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Tim Vieira, Benjamin Lebrun, Mario Giulianelli, Juan Luis Gastaldi, Brian Dusell, John Terilla, Timothy J. O'Donnell, and Ryan Cotterell. 2025. From language models over tokens to language models over characters. In *Proceedings of the International Conference on Machine Learning*, volume 267.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Jennifer C. White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. 2021. A non-linear structural probe. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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.

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. In *Proceedings of the Cognitive Science Society*.

A Reproducibility

Data Sources. The license of both datasets is CC-By Attribution 4.0 International, which we adhere to. We obtained the Provo Corpus on the Open Science Framework.[11] We used the preprocessed version of the MECO data provided in Opedal et al. (2024).[12]

Compute. To estimate our predictors (surprisal, representations, information value, and logit lens), we used an RTX 2080 Ti GPU with 11GB VRAM for circa 30 hours. For our predictive modeling experiments, we used the same GPU, but reserving between 1–8GB of VRAM depending on the feature setting (with mGPT representations requiring the most compute), for approximately two months of compute time total.

Predictive Modeling. We implemented our linear modeling using the `statsmodels`[13] package. We attach the tuning hyperparameters in the supplementary material.

[11] https://osf.io/sjefs
[12] https://github.com/rycolab/context-reading-time/tree/main/merged_data_no_zero
[13] https://www.statsmodels.org/stable/index.html

B Results for mGPT—Combined Settings

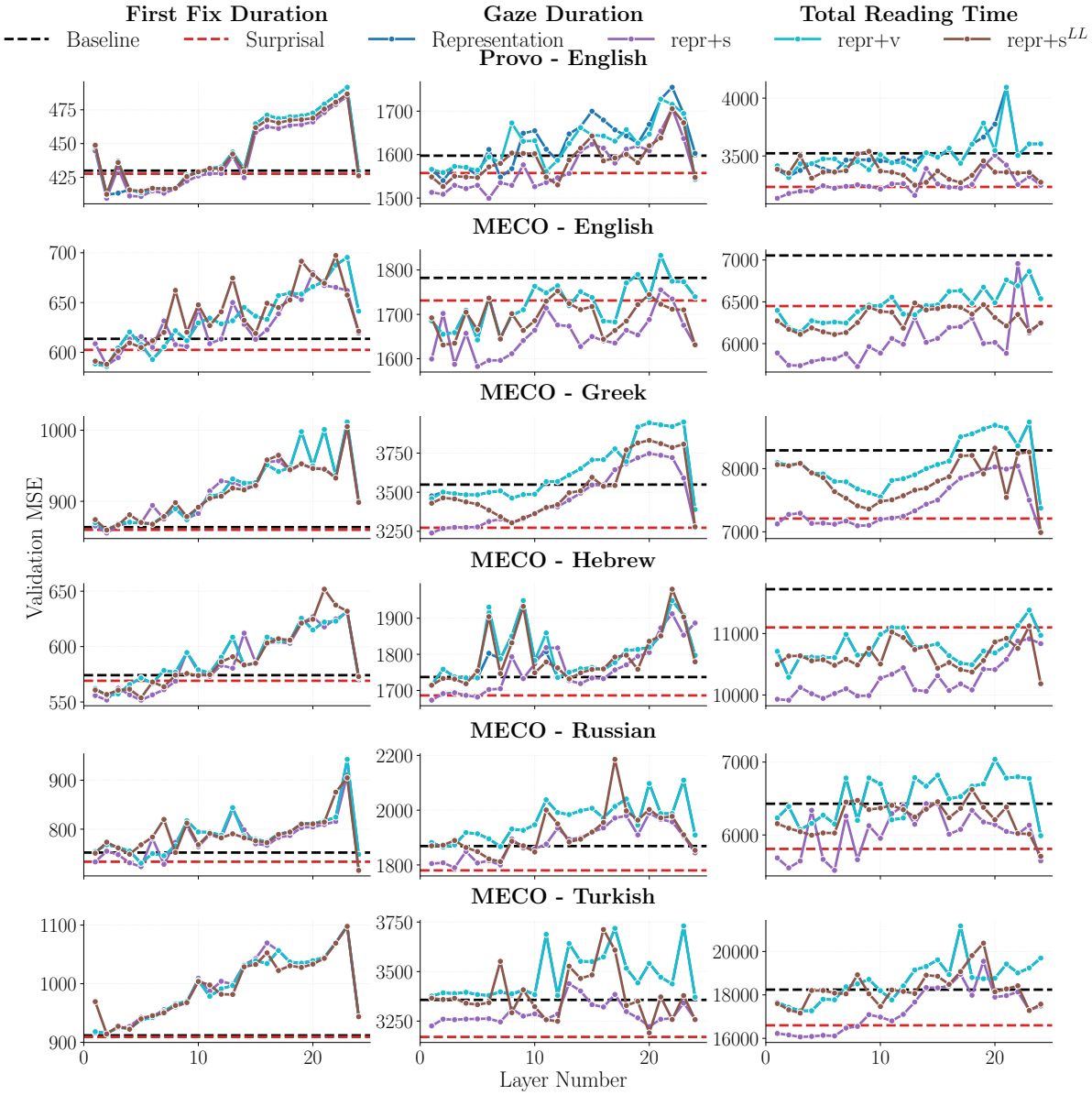

Figure 3: MSE for baseline, surprisal, and combined settings (representations with surprisal, information value, and logit-lens surprisal) on the Provo and MECO data across the 24 layers of mGPT and eye-tracking measures.

Measure	Surprisal	Best $\mathbf{h}$ (ℓ)	Best repr+s (ℓ)	Best repr+v (ℓ)	Best repr+s^{LL} (ℓ)
			Provo—English		
FFD	$-2.28_{4.55}{}^{*}$	$-17.92_{8.76}{}^{*\bullet}$ (2)	$\mathbf{-20.44}_{9.98}{}^{*\bullet}$ **(2)**	$-17.92_{8.76}{}^{*\bullet}$ (2)	$-17.46_{8.50}{}^{*\bullet}$ (2)
GD	$-39.96_{34.34}{}^{*}$	$-57.21_{31.77}{}^{*}$ (2)	$\mathbf{-98.01}_{85.12}{}^{*\bullet\ddagger}$ **(6)**	$-54.60_{76.83}{}^{*\ddagger}$ (24)	$-70.98_{52.66}{}^{*}$ (2)
TRT	$-290.61_{134.19}{}^{*\bullet}$	$-198.09_{196.67}{}^{*\bullet}$ (2)	$\mathbf{-389.00}_{167.39}{}^{*\bullet\ddagger}$ **(1)**	$-208.87_{190.68}{}^{*\bullet}$ (2)	$-275.41_{214.61}{}^{*\bullet\ddagger}$ (13)
			MECO—English		
FFD	$-11.09_{18.93}{}^{*}$	$\mathbf{-27.74}_{29.25}{}^{*}$ **(2)**	$-26.60_{36.19}{}^{*}$ (2)	$\mathbf{-27.74}_{29.25}{}^{*}$ **(2)**	$-25.65_{26.84}{}^{*\bullet}$ (2)
GD	$-50.98_{120.48}{}^{*}$	$-139.92_{169.75}{}^{*\bullet}$ (5)	$\mathbf{-199.65}_{170.26}{}^{*\bullet}$ **(5)**	$-139.92_{169.75}{}^{*\bullet}$ (5)	$-151.29_{142.44}{}^{*\bullet}$ (2)
TRT	$-605.14_{433.80}{}^{*\bullet}$	$-914.85_{746.53}{}^{*\bullet}$ (3)	$\mathbf{-1327.41}_{976.18}{}^{*\bullet\ddagger}$ **(8)**	$-914.85_{746.53}{}^{*\bullet}$ (3)	$-944.99_{857.93}{}^{*\bullet}$ (6)
			MECO—Greek		
FFD	$-3.75_{9.13}{}^{*}$	$-4.72_{22.70}{}^{*}$ (2)	$\mathbf{-8.23}_{25.06}{}^{*}$ **(2)**	$-4.72_{22.70}{}^{*}$ (2)	$-4.09_{26.86}{}^{*}$ (2)
GD	$-275.80_{212.98}{}^{*\bullet}$	$-158.94_{306.51}{}^{*}$ (24)	$\mathbf{-309.45}_{269.98}{}^{*}$ **(1)**	$-158.95_{306.52}{}^{*}$ (24)	$-270.07_{314.15}{}^{*}$ (24)
TRT	$-1079.68_{944.67}{}^{*\bullet}$	$-914.45_{1072.60}{}^{*\bullet}$ (24)	$\mathbf{-1308.51}_{1378.06}{}^{*\bullet}$ **(24)**	$-914.60_{1072.56}{}^{*\bullet}$ (24)	$-1299.87_{1316.34}{}^{*\bullet\ddagger}$ (24)
			MECO—Hebrew		
FFD	$-5.04_{5.08}{}^{*}$	$-18.31_{24.28}{}^{*}$ (2)	$\mathbf{-22.56}_{22.10}{}^{*\ddagger}$ **(5)**	$-18.31_{24.28}{}^{*}$ (2)	$-20.45_{27.61}{}^{*\ddagger}$ (5)
GD	$-50.97_{58.09}{}^{*}$	$-20.96_{114.62}{}^{*}$ (1)	$\mathbf{-64.13}_{94.59}{}^{*}$ **(1)**	$-21.57_{117.99}{}^{*}$ (1)	$-22.71_{116.94}{}^{*}$ (1)
TRT	$-621.84_{786.76}{}^{*\bullet}$	$-1431.52_{2384.96}{}^{*\bullet}$ (2)	$\mathbf{-1804.89}_{1972.43}{}^{*\bullet}$ **(2)**	$-1431.52_{2384.96}{}^{*\bullet}$ (2)	$-1537.43_{1621.02}{}^{*\bullet\ddagger}$ (24)
			MECO—Russian		
FFD	$-18.74_{22.90}{}^{\bullet}$	$-21.86_{28.87}{}^{*}$ (5)	$-36.11_{49.96}{}^{*\ddagger}$ (24)	$-21.86_{28.87}{}^{*}$ (5)	$\mathbf{-36.37}_{49.17}{}^{*\ddagger}$ **(24)**
GD	$\mathbf{-88.70}_{110.92}{}^{*\bullet}$	$-3.59_{101.50}{}^{*}$ (2)	$-78.96_{175.41}{}^{*\ddagger}$ (3)	$-3.59_{101.50}{}^{*}$ (2)	$-56.70_{144.80}{}^{*}$ (7)
TRT	$-619.68_{741.79}{}^{*\bullet}$	$-438.41_{605.05}{}^{*\bullet}$ (24)	$\mathbf{-914.07}_{1179.86}{}^{*\bullet\ddagger}$ **(6)**	$-438.61_{605.01}{}^{*\bullet}$ (24)	$-721.94_{1213.85}{}^{*\bullet}$ (24)
			MECO—Turkish		
FFD	$\mathbf{-2.63}_{6.62}$	$3.67_{29.35}{}^{*}$ (2)	$4.01_{31.17}{}^{*}$ (1)	$3.67_{29.35}{}^{*}$ (2)	$2.35_{32.29}{}^{*}$ (2)
GD	$\mathbf{-186.91}_{119.24}{}^{*\bullet}$	$14.06_{167.93}{}^{*}$ (24)	$-134.56_{163.52}{}^{*\ddagger}$ (20)	$13.96_{167.93}{}^{*}$ (24)	$-166.31_{182.10}{}^{*\ddagger}$ (20)
TRT	$-1642.88_{1410.20}{}^{*\bullet}$	$-973.47_{1871.74}{}^{*\bullet}$ (3)	$\mathbf{-2176.55}_{2190.16}{}^{*\bullet\ddagger}$ **(3)**	$-973.78_{1871.72}{}^{*\bullet\ddagger}$ (3)	$-1079.82_{1906.71}{}^{*}$ (3)

Table 2: Δ_{MSE} of ten-fold cross-validation for models trained on baseline features and mGPT-derived surprisal and representations, as well as combined settings: representations + surprisal (repr+s), representations + information value (repr+v), and representations + logit-lens surprisal (repr+s^{LL}). Experiments were conducted on the Provo and MECO data across the 24 layers of mGPT and eye-tracking measures. For each measure, we report the lowest MSE over layers and the corresponding layer index ℓ. Bold indicates the best condition per row. Asterisks (*) denote models that significantly outperform the respective models trained on permuted reading times, according to a one-sided paired t-test ($\alpha = 0.001$). Similarly, bullets (•) indicate significance over the baseline. In combined settings, double daggers (‡) indicate significance over representation-trained models. None of the models in the combined settings were significant over their respective scalars, e.g. repr+s over surprisal.

C Results for Monolingual Models

C.1 Individual Predictors

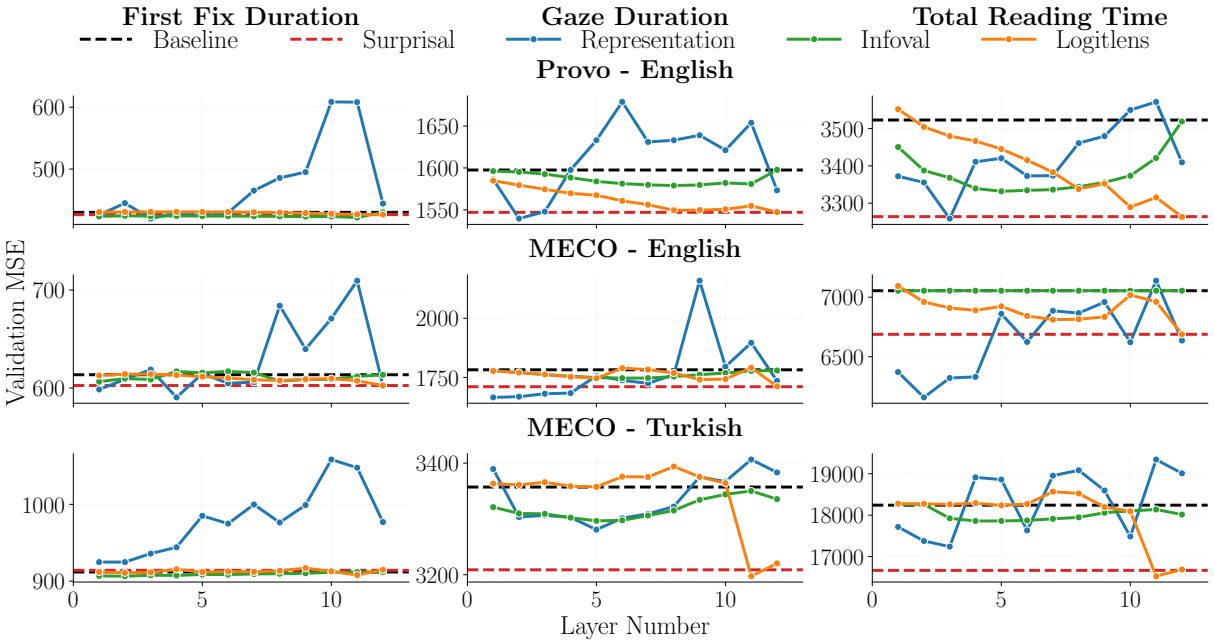

Figure 4: MSE for baseline, surprisal, representations, information value, and logit-lens surprisal on the Provo and English MECO with GPT-2 and Turkish MECO data with cosmosGPT, across the 12 layers of each language model and eye-tracking measures.

Measure	Surprisal	Best $\mathbf{h}$ (ℓ)	Best v (ℓ)	Best s^{LL} (ℓ)
		Provo—English		
FFD	$-3.70_{7.01}{}^{*}$	$\mathbf{-10.05}_{13.84}{}^{*}$ **(3)**	$-8.39_{11.53}{}^{*}$ (11)	$-3.55_{7.05}{}^{*}$ (12)
GD	$-50.48_{54.64}{}^{*}$	$\mathbf{-57.92}_{90.88}{}^{*}$ **(2)**	$-18.51_{28.90}{}^{*}$ (8)	$-50.22_{54.88}{}^{*}$ (12)
TRT	$-258.98_{205.34}{}^{*\bullet}$	$\mathbf{-264.17}_{139.95}{}^{*\bullet}$ **(3)**	$-191.08_{140.91}{}^{*\bullet}$ (5)	$-259.69_{199.76}{}^{*\bullet}$ (12)
		MECO—English		
FFD	$-11.00_{19.90}{}^{*}$	$\mathbf{-23.12}_{19.55}{}^{*}$ **(4)**	$-7.06_{11.95}{}^{*}$ (1)	$-11.00_{19.85}{}^{*}$ (12)
GD	$-71.56_{100.89}{}^{*}$	$\mathbf{-116.72}_{136.39}{}^{*\bullet}$ **(1)**	$-35.20_{41.36}{}^{*\bullet}$ (6)	$-70.19_{100.40}{}^{*}$ (12)
TRT	$-366.37_{445.23}{}^{*}$	$\mathbf{-896.11}_{761.82}{}^{*\bullet}$ **(2)**	$0.00_{0.00}{}^{*}$ (1)	$-365.45_{439.37}{}^{*}$ (12)
		MECO—Turkish		
FFD	$2.20_{13.84}$	$12.91_{24.99}$ (1)	$\mathbf{-5.51}_{7.51}$ **(2)**	$-3.97_{6.28}{}^{*}$ (11)
GD	$-148.33_{78.18}{}^{*\bullet}$	$-76.27_{196.86}{}^{*}$ (5)	$-60.60_{48.24}{}^{*}$ (5)	$\mathbf{-159.79}_{86.08}{}^{*\bullet}$ **(11)**
TRT	$-1575.56_{1668.76}{}^{*\bullet}$	$-1000.99_{1953.89}{}^{*\bullet}$ (3)	$-381.99_{319.72}{}^{*\bullet}$ (4)	$\mathbf{-1716.61}_{1877.91}{}^{*\bullet}$ **(11)**

Table 3: Δ_{MSE} (baseline − target) of ten-fold cross-validation for models trained on surprisal, representations ($\mathbf{h}$), information value (v), and logit-lens surprisal (s^{LL}) derived from GPT-2 for the Provo and English MECO data, and from cosmosGPT for Turkish MECO data, across the 12 embedding layers of each language model and eye-tracking measures. For each measure, we report the lowest MSE over layers and the corresponding layer index ℓ. Bold indicates the best condition per row. Asterisks (*) denote models that significantly outperform the respective models trained on permuted reading times, according to a one-sided paired t-test ($\alpha = 0.001$). Similarly, bullets (•) indicate significance over the baseline.

C.2 Combined Settings

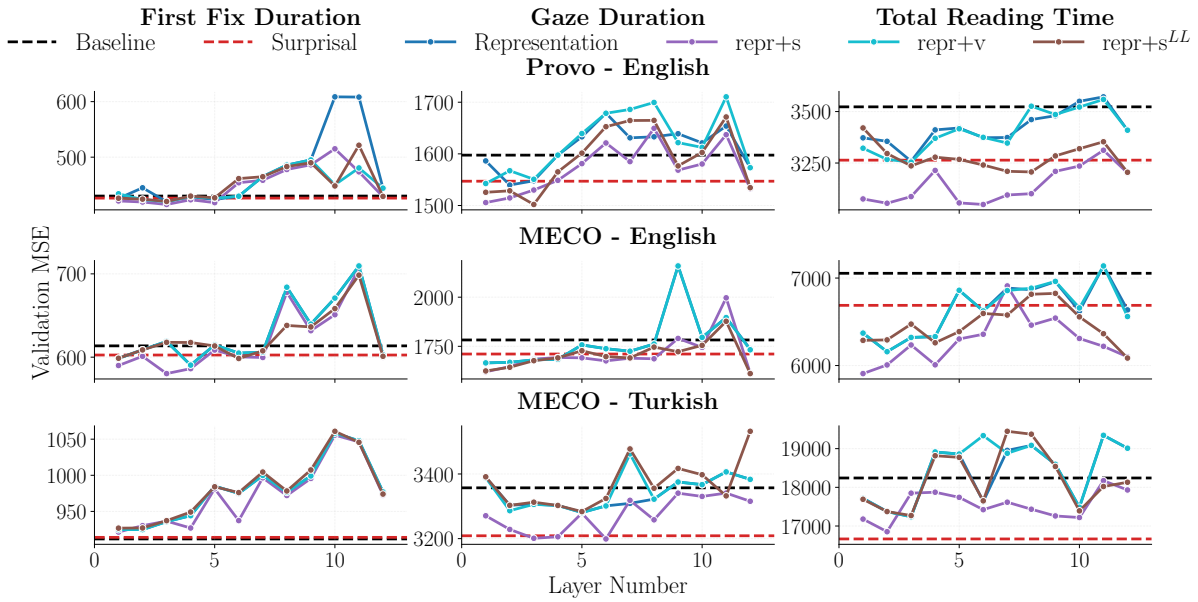

Figure 5: MSE for baseline, surprisal, and combined settings (representations with surprisal, information value, and logit-lens surprisal) on the Provo and English MECO data with GPT-2 and Turkish MECO data with cosmosGPT, across the 12 layers of each language model and eye-tracking measures.

Measure	Surprisal	Best **h** (ℓ)	Best repr+s (ℓ)	Best repr+v (ℓ)	Best repr+s^{LL} (ℓ)
			Provo—English		
FFD	$-3.70_{7.01}^{*}$	$-10.05_{13.84}^{*}$ (3)	$\mathbf{-15.18}_{19.58}^{*}$ **(3)**	$-10.05_{13.84}^{*}$ (3)	$-9.77_{14.20}^{*}$ (3)
GD	$-50.48_{54.64}^{*}$	$-57.92_{90.88}^{*}$ (2)	$-91.86_{96.53}^{*\ddagger}$ (1)	$-54.96_{76.80}^{*}$ (1)	$\mathbf{-95.54}_{117.95}^{*}$ **(3)**
TRT	$-258.98_{205.34}^{*\bullet}$	$-264.17_{139.95}^{*\bullet}$ (3)	$\mathbf{-474.29}_{291.50}^{*\bullet\ddagger}$ **(6)**	$-268.57_{135.91}^{*\bullet}$ (3)	$-317.86_{292.19}^{*\bullet\ddagger}$ (12)
			MECO—English		
FFD	$-11.00_{19.90}^{*}$	$-23.12_{19.55}^{*}$ (4)	$\mathbf{-33.29}_{36.91}^{*\bullet\ddagger}$ **(3)**	$-23.12_{19.55}^{*}$ (4)	$-15.20_{27.29}^{*}$ (6)
GD	$-71.56_{100.89}^{*}$	$-116.72_{136.39}^{*\bullet}$ (1)	$-168.44_{176.21}^{*\bullet\ddagger}$ (12)	$-116.72_{136.39}^{*\bullet}$ (1)	$\mathbf{-171.10}_{176.75}^{*\bullet\ddagger}$ **(12)**
TRT	$-366.37_{445.23}^{*}$	$-896.11_{761.82}^{*\bullet}$ (2)	$\mathbf{-1146.38}_{1069.94}^{*\bullet\ddagger}$ **(1)**	$-896.11_{761.82}^{*\bullet}$ (2)	$-968.82_{822.72}^{*\bullet\ddagger}$ (12)
			MECO—Turkish		
FFD	$2.20_{13.84}$	$12.91_{24.99}$ (1)	$9.50_{23.80}$ (1)	$12.91_{24.99}$ (1)	$14.96_{26.12}$ (1)
GD	$-148.33_{78.18}^{*\bullet}$	$-76.27_{196.86}^{*}$ (5)	$\mathbf{-158.12}_{176.68}^{*}$ **(6)**	$-76.55_{196.89}^{*\ddagger}$ (5)	$-73.45_{199.55}^{*}$ (5)
TRT	$\mathbf{-1575.56}_{1668.76}^{*\bullet}$	$-1000.99_{1953.89}^{*\bullet}$ (3)	$-1388.19_{3077.40}^{*}$ (2)	$-1003.79_{1953.62}^{*\bullet\ddagger}$ (3)	$-970.59_{1945.13}^{*\bullet}$ (3)

Table 4: As in Table 3, except we now consider the Δ_{MSE} of surprisal, representations, and combined settings: representations + surprisal (repr+s), representations + information value (repr+v), and representations + logit-lens surprisal (repr+s^{LL}). In combined settings, double daggers (‡) indicate statistical significance over models trained on representations. Note that all of these predictors performed worse than the baseline for first fixation duration on Turkish.

D Qualitative Example

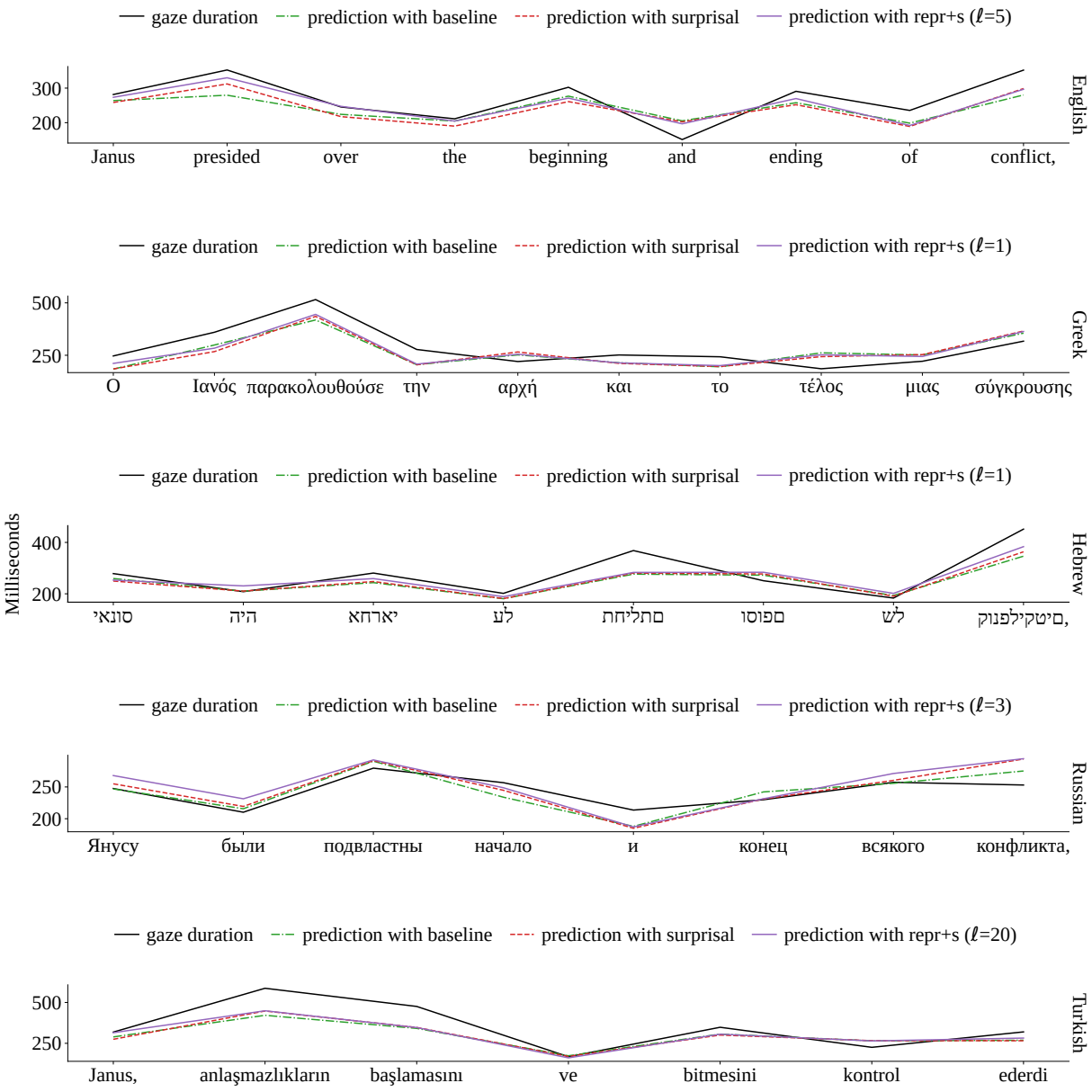

Figure 6: Gaze duration and its prediction by different mGPT-derived feature settings. We show the same excerpt from a document in the MECO dataset in different languages. The y-axis represents reading time measured in milliseconds. True gaze duration is represented by a black line. The purple line represents the prediction of a linear model trained on ℓ^{th} layer representations and standard surprisal. The chosen layer was the best for this feature setting and eye tracking measure per Table 2. Note that Hebrew is in reverse order, as Hebrew is read and written right-to-left, and MECO data has words indexed by the order they are read.

E Permuted Results

E.1 mGPT—Permuted

E.1.1 Individual Predictors

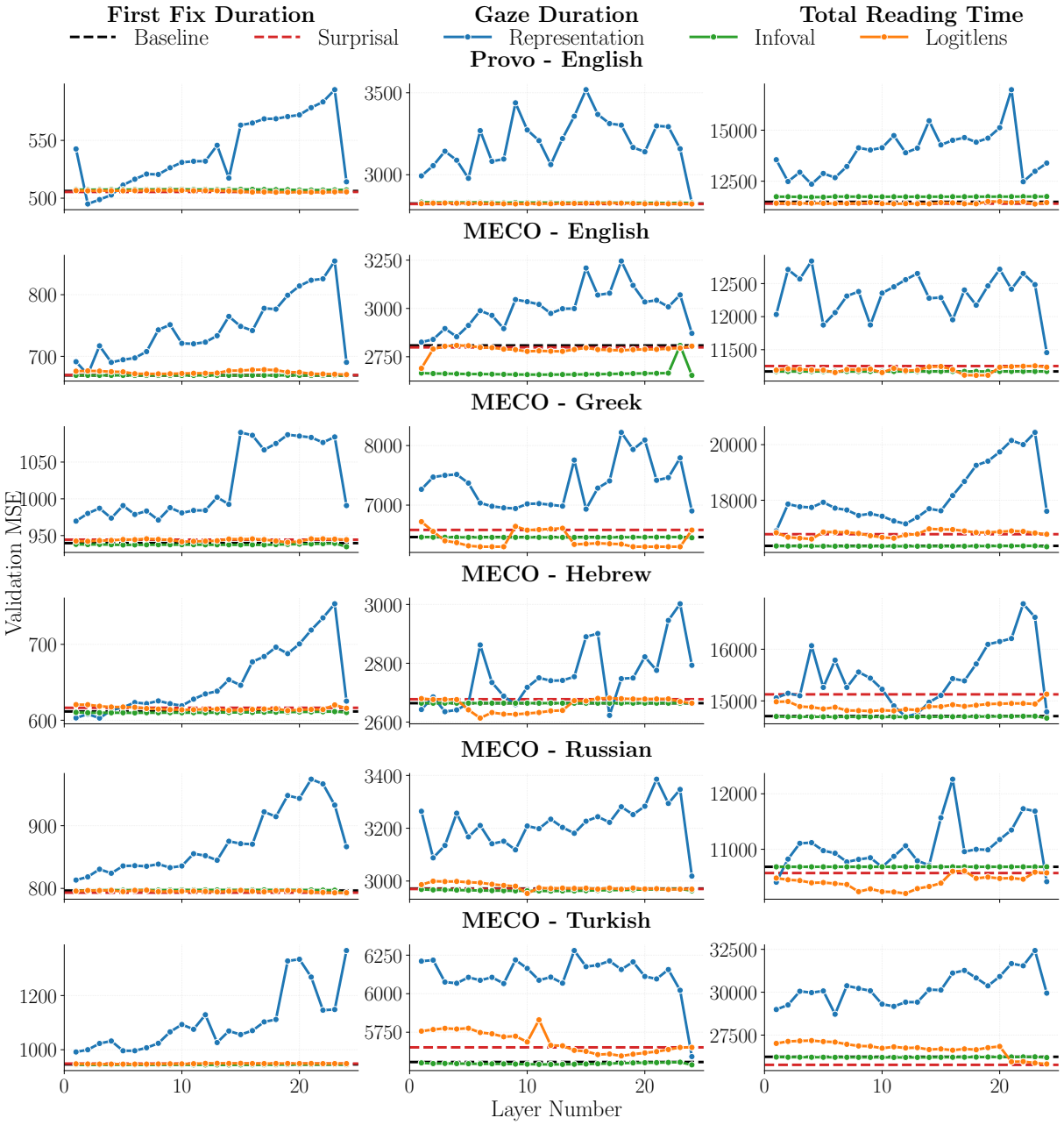

Figure 7: MSE for baseline, surprisal, representations, information value, and logit-lens surprisal on the Provo and MECO data across the 24 layers of mGPT and eye-tracking measures **with reading times randomly permuted during training**.

Measure	Surprisal	Best $\mathbf{h}$ (ℓ)	Best v (ℓ)	Best s^{LL} (ℓ)
		Provo—English		
FFD	$-0.79_{2.96}$	$\mathbf{-11.23}_{16.03}^{\bullet}$ **(2)**	$0.80_{4.66}$ (20)	$-1.07_{3.90}$ (18)
GD	$\mathbf{-2.64}_{10.05}$	$2.87_{158.05}$ (24)	$3.26_{14.14}$ (22)	$-2.61_{10.06}$ (24)
TRT	$-84.88_{131.25}^{\bullet}$	$869.44_{286.65}$ (4)	$232.09_{211.80}$ (4)	$\mathbf{-102.83}_{182.46}^{\bullet}$ **(23)**
		MECO—English		
FFD	$0.47_{5.30}$	$1.30_{31.42}$ (2)	$0.00_{0.00}$ (1)	$1.27_{3.26}$ (24)
GD	$-9.90_{31.13}^{\bullet}$	$17.19_{46.97}$ (1)	$\mathbf{-154.72}_{103.73}^{\bullet}$ **(24)**	$-118.38_{102.49}^{\bullet}$ (1)
TRT	$81.39_{185.30}$	$284.92_{441.58}$ (24)	$0.00_{0.00}$ (1)	$\mathbf{-59.33}_{425.79}$ **(18)**
		MECO—Greek		
FFD	$4.68_{4.57}$	$29.90_{31.19}$ (1)	$\mathbf{-4.91}_{8.69}$ **(24)**	$1.06_{10.30}$ (19)
GD	$120.19_{119.49}$	$438.84_{475.50}$ (24)	$-11.17_{20.89}$ (24)	$\mathbf{-162.05}_{111.87}^{\bullet}$ **(23)**
TRT	$419.74_{392.78}$	$540.16_{1144.00}$ (1)	$\mathbf{-30.47}_{28.26}$ **(24)**	$247.26_{667.37}$ (4)
		MECO—Hebrew		
FFD	$4.84_{17.24}$	$\mathbf{-8.71}_{27.72}$ **(3)**	$-1.84_{1.72}$ (6)	$1.19_{17.98}$ (19)
GD	$13.46_{28.99}$	$-41.55_{390.14}$ (17)	$0.00_{0.00}$ (1)	$\mathbf{-50.71}_{111.73}^{\bullet}$ **(6)**
TRT	$421.20_{992.75}$	$-12.78_{1579.43}$ (12)	$\mathbf{-38.06}_{38.82}^{\bullet}$ **(24)**	$97.95_{123.34}$ (9)
		MECO—Russian		
FFD	$-3.28_{7.87}$	$16.77_{23.60}$ (1)	$\mathbf{-4.63}_{6.13}^{\bullet}$ **(24)**	$-3.48_{8.98}$ (23)
GD	$-1.69_{74.90}$	$47.14_{159.98}$ (24)	$-9.45_{10.76}$ (12)	$\mathbf{-18.68}_{84.95}$ **(10)**
TRT	$-111.87_{260.18}^{\bullet}$	$-277.17_{724.43}$ (1)	$0.00_{0.00}$ (1)	$\mathbf{-480.24}_{638.05}^{\bullet}$ **(12)**
		MECO—Turkish		
FFD	$2.90_{4.52}$	$45.91_{30.86}$ (1)	$\mathbf{-1.71}_{3.66}$ **(14)**	$0.92_{4.37}$ (5)
GD	$95.13_{99.69}$	$35.26_{407.39}$ (24)	$\mathbf{-18.35}_{22.87}$ **(24)**	$39.64_{48.78}$ (18)
TRT	$\mathbf{-469.13}_{886.55}$	$2484.25_{2163.78}$ (6)	$-43.42_{55.23}$ (24)	$-418.17_{887.85}$ (24)

Table 5: Δ_{MSE} (baseline − target) of ten-fold cross-validation for models trained on baseline features and mGPT-derived surprisal, representations ($\mathbf{h}$), information value (v), and logit-lens surprisal (s^{LL}) **with reading times randomly permuted during training** on the Provo and MECO data across the 24 layers of mGPT. For each eye-tracking measure, we report the lowest MSE over layers and the corresponding layer index ℓ. Bold indicates the best condition per row. Bullets ($\bullet$) denote models that significantly outperform the baseline, according to a one-sided paired t-test ($\alpha = 0.001$).

E.1.2 Combined Settings

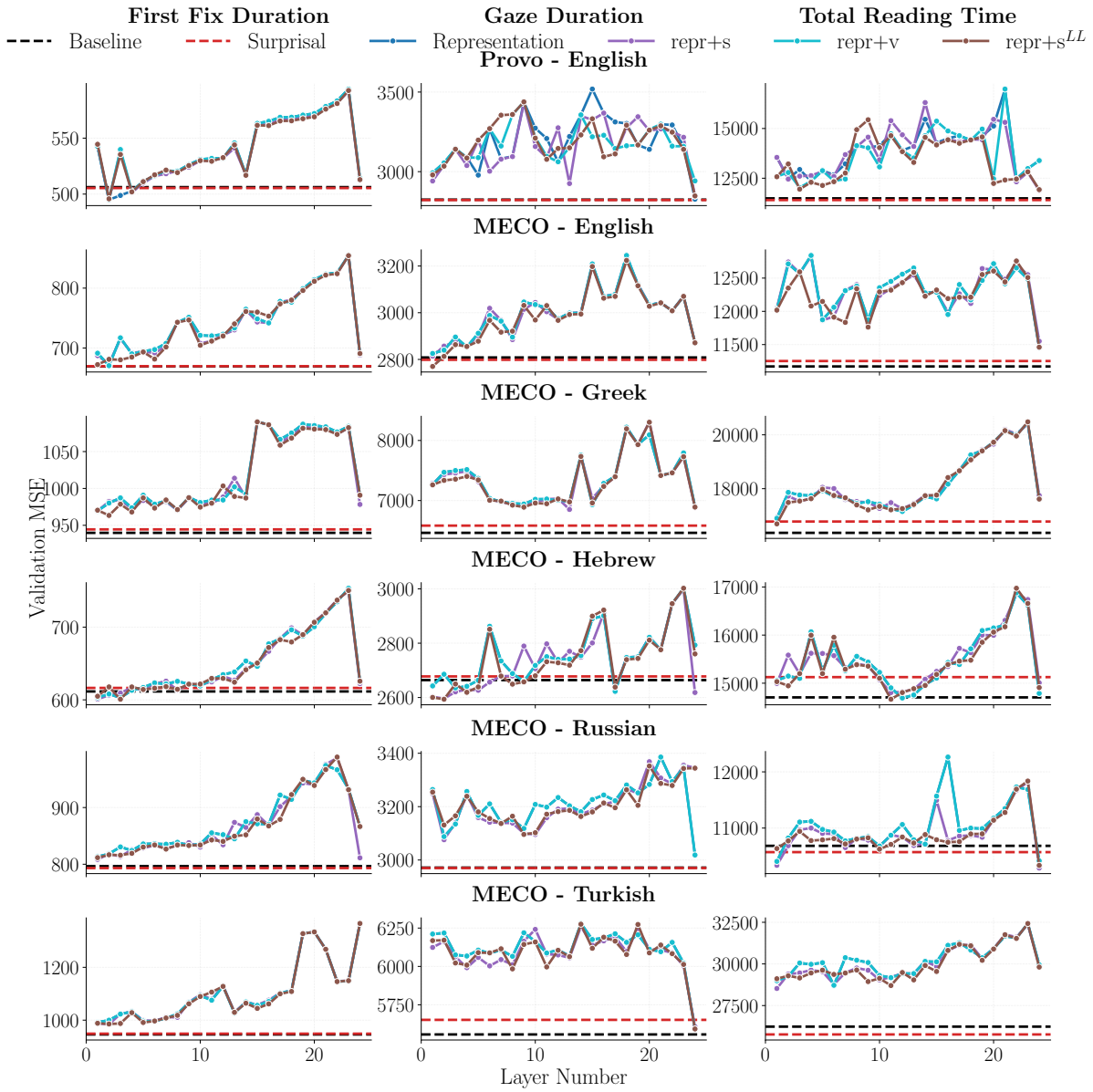

Figure 8: MSE for baseline, surprisal, and combined settings (representations with surprisal, information value, and logit-lens surprisal) on the Provo and MECO data across the 24 layers of mGPT and eye-tracking measures **with reading times randomly permuted during training**.

Measure	Surprisal	Best $\mathbf{h}$ (ℓ)	Best repr+s (ℓ)	Best repr+v (ℓ)	Best repr+s^{LL} (ℓ)
			Provo—English		
FFD	$-0.79_{2.96}$	$-11.23_{16.03}^{\bullet}$ (2)	**$-11.53_{16.29}$ (2)**	$-11.23_{16.03}^{\bullet}$ (2)	$-10.44_{15.46}$ (2)
GD	**$-2.64_{10.05}$**	$2.87_{158.05}$ (24)	$22.04_{87.40}$ (24)	$117.57_{98.88}$ (24)	$23.41_{86.33}$ (24)
TRT	**$-84.88_{131.25}^{\bullet}$**	$869.44_{286.65}$ (4)	$399.78_{508.09}^{\ddagger}$ (24)	$541.40_{287.46}^{\ddagger}$ (3)	$438.03_{572.60}^{\ddagger}$ (24)
			MECO—English		
FFD	$0.47_{5.30}$	$1.30_{31.42}$ (2)	$4.55_{26.14}$ (2)	$1.30_{31.42}$ (2)	$3.09_{17.74}^{\ddagger}$ (1)
GD	$-9.90_{31.13}^{\bullet}$	$17.19_{46.97}$ (1)	$5.79_{50.99}$ (1)	$17.19_{46.97}$ (1)	**$-38.42_{58.76}$ (1)**
TRT	$81.39_{185.30}$	$284.92_{441.58}$ (24)	$379.85_{481.71}$ (24)	$284.92_{441.58}$ (24)	$289.60_{447.32}$ (24)
			MECO—Greek		
FFD	$4.68_{4.57}$	$29.90_{31.19}$ (1)	$29.76_{29.49}$ (1)	$29.90_{31.19}$ (1)	$23.61_{29.14}^{\ddagger}$ (2)
GD	$120.19_{119.49}$	$438.84_{475.50}$ (24)	$389.14_{380.88}$ (13)	$438.84_{475.50}$ (24)	$427.27_{323.76}$ (9)
TRT	$419.74_{392.78}$	$540.16_{1144.00}$ (1)	$576.88_{1128.93}$ (1)	$540.16_{1144.00}$ (1)	$333.44_{648.44}$ (1)
			MECO—Hebrew		
FFD	$4.84_{17.24}$	$-8.71_{27.72}$ (3)	$-10.58_{29.66}$ (1)	$-8.71_{27.72}$ (3)	**$-10.72_{32.18}$ (3)**
GD	$13.46_{28.99}$	$-41.55_{390.14}$ (17)	$-68.35_{181.72}^{\ddagger}$ (2)	$-41.55_{390.14}$ (17)	**$-70.53_{166.65}^{\ddagger}$ (2)**
TRT	$421.20_{992.75}$	$-12.78_{1579.43}$ (12)	$99.43_{1393.98}$ (11)	$-12.78_{1579.43}$ (12)	**$-38.39_{1493.58}^{\ddagger}$ (11)**
			MECO—Russian		
FFD	**$-3.28_{7.87}$**	$16.77_{23.60}$ (1)	$11.91_{25.19}$ (1)	$16.77_{23.60}$ (1)	$15.07_{22.85}$ (1)
GD	**$-1.69_{74.90}$**	$47.14_{159.98}$ (24)	$104.39_{103.45}$ (2)	$47.14_{159.98}$ (24)	$124.32_{196.86}$ (9)
TRT	$-111.87_{260.18}^{\bullet}$	$-277.17_{724.43}$ (1)	**$-395.25_{1043.14}^{\bullet}$ (24)**	$-277.17_{724.43}$ (1)	$-347.73_{1064.00}^{\bullet}$ (24)
			MECO—Turkish		
FFD	$2.90_{4.52}$	$45.91_{30.86}$ (1)	$38.61_{32.40}$ (2)	$45.91_{30.86}$ (1)	$39.56_{40.33}$ (2)
GD	$95.13_{99.69}$	$35.26_{407.39}$ (24)	$51.21_{413.28}$ (24)	$35.26_{407.39}$ (24)	$35.81_{408.01}$ (24)
TRT	**$-469.13_{886.55}$**	$2484.25_{2163.78}$ (6)	$2291.41_{2137.59}$ (1)	$2484.25_{2163.78}$ (6)	$2468.11_{2508.89}^{\ddagger}$ (11)

Table 6: Δ_{MSE} (baseline $-$ target) of ten-fold cross-validation for models trained on baseline features and mGPT-derived surprisal, as well as combined settings: representations + surprisal (repr+s), representations + information value (repr+v), and representations + logit-lens surprisal (repr+s^{LL}), **with reading times randomly permuted during training** on the Provo and MECO data across the 24 layers of mGPT. For each eye-tracking measure, we report the lowest MSE over layers and the corresponding layer index ℓ. Bold indicates the best condition per row. Bullets ($\bullet$) denote models that significantly outperform the baseline, according to a one-sided paired t-test ($\alpha = 0.001$). In combined settings, double daggers ($\ddagger$) indicate statistical significance over models trained on representations.

E.2 Monolingual Models—Permuted

E.2.1 Individual Predictors

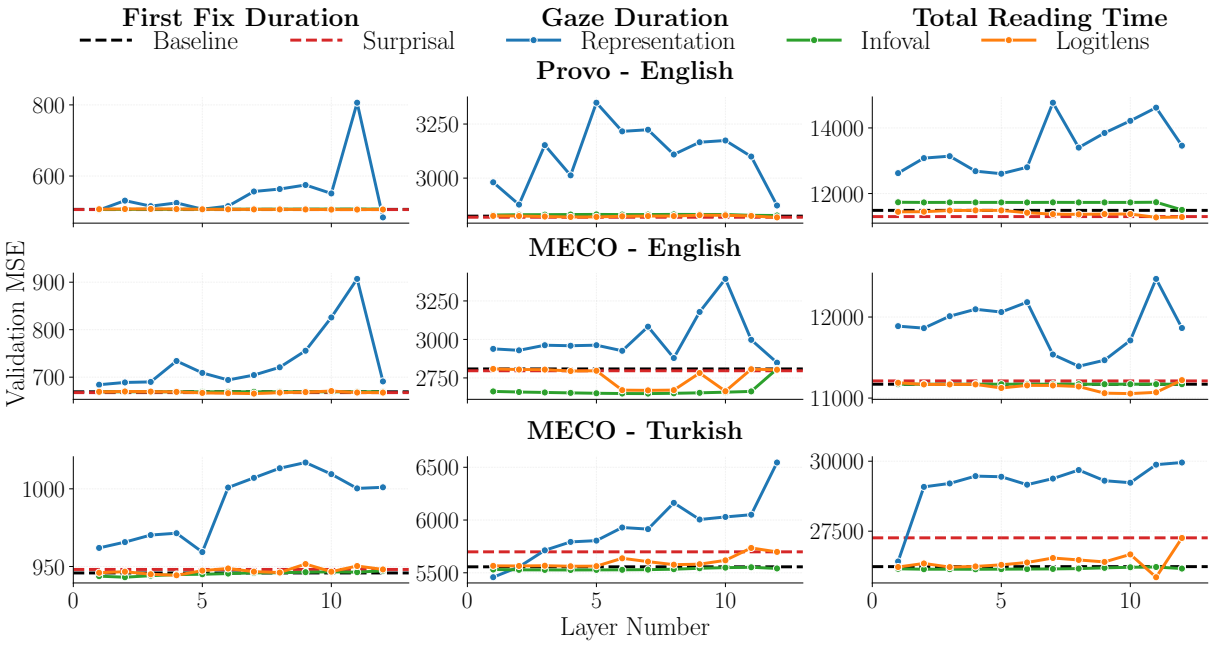

Figure 9: MSE for baseline, surprisal, representations, information value, and logit-lens surprisal on the Provo and English MECO with GPT-2 and Turkish MECO data with cosmosGPT, across the 12 layers of each language model and eye-tracking measures **with reading times randomly permuted during training**.

Measure	Surprisal	Best $\mathbf{h}$ (ℓ)	Best v (ℓ)	Best s^{LL} (ℓ)
		Provo—English		
FFD	$-0.54_{3.45}$	$\mathbf{-22.56}_{18.65}^{\bullet}$ **(12)**	$-0.20_{3.47}$ (1)	$-0.61_{3.40}$ (10)
GD	$-5.04_{11.92}$	$49.20_{91.78}$ (12)	$3.09_{2.91}$ (12)	$\mathbf{-5.83}_{13.60}$ **(12)**
TRT	$-189.62_{137.32}^{\bullet}$	$1117.49_{686.19}$ (5)	$14.75_{160.53}$ (12)	$\mathbf{-214.38}_{170.94}^{\bullet}$ **(11)**
		MECO—English		
FFD	$-1.73_{8.67}$	$14.93_{36.99}$ (1)	$0.00_{0.00}$ (1)	$\mathbf{-3.79}_{7.83}^{\bullet}$ **(7)**
GD	$-11.95_{30.49}^{\bullet}$	$40.69_{81.35}$ (12)	$\mathbf{-160.25}_{103.02}^{\bullet}$ **(7)**	$-144.79_{100.08}^{\bullet}$ (10)
TRT	$40.57_{151.86}$	$221.43_{666.71}$ (8)	$0.00_{0.00}$ (1)	$\mathbf{-115.94}_{390.15}^{\bullet}$ **(10)**
		MECO—Turkish		
FFD	$2.27_{8.06}$	$13.45_{31.13}$ (5)	$\mathbf{-2.74}_{7.58}$ **(2)**	$-1.49_{4.66}$ (4)
GD	$141.95_{158.25}$	$\mathbf{-98.55}_{123.02}$ **(1)**	$-29.89_{28.11}$ (4)	$5.47_{13.11}$ (4)
TRT	$1028.82_{1222.60}$	$184.14_{976.70}$ (1)	$-95.05_{60.78}^{\bullet}$ (4)	$\mathbf{-378.28}_{822.30}$ **(11)**

Table 7: Δ_{MSE} (baseline − target) of ten-fold cross-validation for models trained on baseline features and surprisal, representations ($\mathbf{h}$), information value (v), and logit-lens surprisal (s^{LL}) derived from GPT-2 for the Provo and English MECO data, and from cosmosGPT for Turkish MECO data **with reading times randomly permuted during training**. For each measure, we report the lowest MSE over layers and the corresponding layer index ℓ. Bold indicates the best condition per row. Bullets ($\bullet$) denote models that significantly outperform the baseline, according to a one-sided paired t-test ($\alpha = 0.001$).

E.2.2 Combined Settings

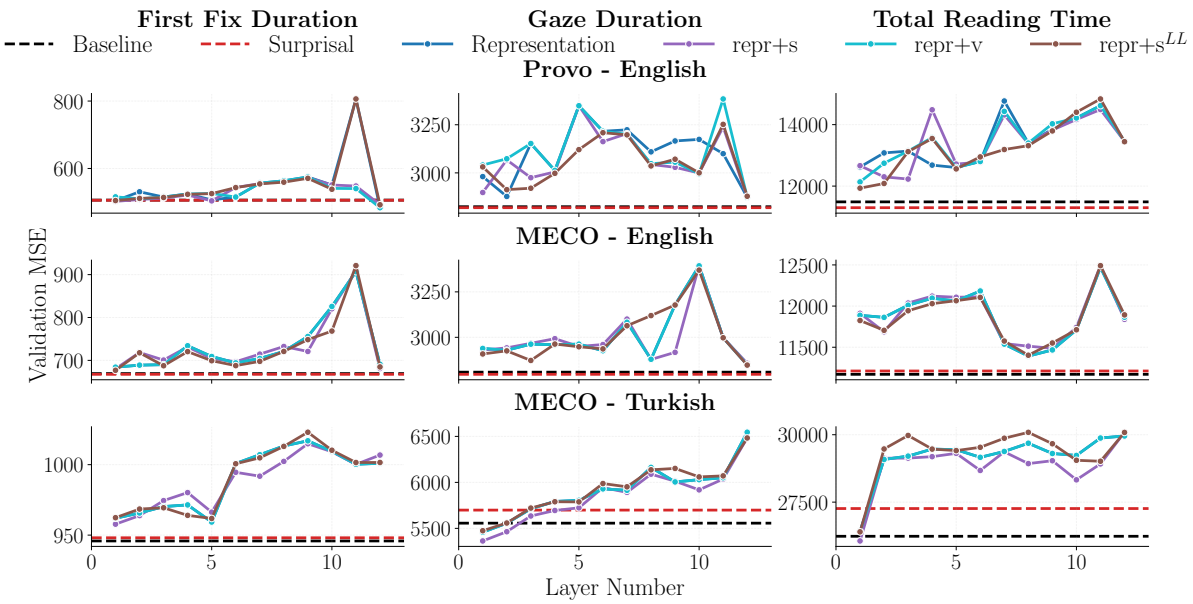

Figure 10: MSE for baseline, surprisal, and combined settings (representations with surprisal, information value, and logit-lens surprisal) on the Provo and English MECO data with GPT-2 and Turkish MECO data with cosmosGPT, across the 12 layers of each language model and eye-tracking measures **with reading times randomly permuted during training**.

Measure	Surprisal	Best $\mathbf{h}$ (ℓ)	Best repr+s (ℓ)	Best repr+v (ℓ)	Best repr+s^{LL} (ℓ)
			Provo—English		
FFD	$-0.54_{3.45}$	**$-22.56_{18.65}^{\bullet}$ (12)**	$-13.52_{16.36}^{\bullet}$ (12)	**$-22.56_{18.65}^{\bullet}$ (12)**	$-13.91_{16.25}^{\bullet}$ (12)
GD	**$-5.04_{11.92}$**	$49.20_{91.78}$ (12)	$48.83_{98.24}$ (12)	$49.20_{91.78}$ (12)	$54.42_{108.56}$ (12)
TRT	**$-189.62_{137.32}^{\bullet}$**	$1117.49_{686.19}$ (5)	$742.37_{531.21}^{\ddagger}$ (3)	$648.53_{602.57}^{\ddagger}$ (1)	$446.26_{503.46}^{\ddagger}$ (1)
			MECO—English		
FFD	**$-1.73_{8.67}$**	$14.93_{36.99}$ (1)	$11.08_{39.49}^{\ddagger}$ (1)	$14.93_{36.99}^{\ddagger}$ (1)	$7.67_{36.17}$ (1)
GD	**$-11.95_{30.49}^{\bullet}$**	$40.69_{81.35}$ (12)	$48.68_{83.21}$ (12)	$40.69_{81.35}$ (12)	$38.91_{76.28}$ (12)
TRT	$40.57_{151.86}$	$221.43_{666.71}$ (8)	$316.13_{817.56}$ (9)	$221.43_{666.71}$ (8)	$231.97_{700.13}$ (8)
			MECO—Turkish		
FFD	$2.27_{8.06}$	$13.45_{31.13}$ (5)	$11.92_{23.33}$ (1)	$13.45_{31.13}$ (5)	$15.96_{33.27}$ (5)
GD	$141.95_{158.25}$	$-98.55_{123.02}$ (1)	**$-193.53_{118.33}^{\bullet\ddagger}$ (1)**	$-98.70_{123.15}$ (1)	$-82.34_{150.56}$ (1)
TRT	$1028.82_{1222.60}$	$184.14_{976.70}$ (1)	**$-173.65_{901.31}^{\ddagger}$ (1)**	$183.57_{976.36}$ (1)	$164.47_{1191.25}$ (1)

Table 8: Δ_{MSE} (baseline $-$ target) of ten-fold cross-validation for models trained on baseline features and surprisal, as well as combined settings: representations + surprisal (repr+s), representations + information value (repr+v), and representations + logit-lens surprisal (repr+s^{LL}) derived from GPT-2 for the Provo and English MECO data, and from cosmosGPT for Turkish MECO data **with reading times randomly permuted during training**. For each measure, we report the lowest MSE over layers and the corresponding layer index ℓ. Bold indicates the best condition per row. Bullets ($\bullet$) denote models that significantly outperform the baseline, according to a one-sided paired t-test ($\alpha = 0.001$). Similarly, for combined settings, double daggers ($\ddagger$) indicate significance over representation-trained models.

F Linear Mixed-Effects Models

Measure	Surprisal	Best $\mathbf{h}$ (ℓ)	Best v (ℓ)	Best s^{LL} (ℓ)
		MECO—English		
FFD	$-19.67_{25.32}$*	**$-37.84_{38.14}$ (12)**	$-18.90_{20.13}$* (11)	$-19.78_{25.58}$* (24)
GD	$-127.14_{126.48}$*	**$-150.98_{147.34}$* (6)**	$-65.36_{48.64}$* (9)	$-128.45_{128.23}$* (24)
TRT	$-695.30_{613.28}$*	**$-1029.41_{1181.09}$* (11)**	$-361.08_{256.73}$*• (8)	$-700.19_{622.56}$* (24)
		MECO—Greek		
FFD	$-5.66_{13.18}$*	**$-13.01_{28.70}$* (7)**	$-8.59_{7.77}$* (1)	$-7.03_{14.01}$* (23)
GD	$-301.80_{271.98}$*	**$-306.80_{196.41}$*• (9)**	$-36.21_{39.22}$* (6)	$-304.54_{248.30}$* (23)
TRT	$-1321.44_{1322.95}$*	$-1295.00_{1436.95}$* (17)	$-249.45_{183.68}$* (6)	**$-1332.82_{1225.38}$* (23)**
		MECO—Hebrew		
FFD	$-5.04_{7.55}$*	**$-16.51_{24.18}$* (13)**	$-5.07_{5.21}$* (6)	$-5.64_{6.74}$* (19)
GD	$-31.10_{54.09}$*	**$-55.32_{79.07}$* (13)**	$-14.76_{14.48}$* (6)	$-30.46_{53.98}$* (24)
TRT	$-361.63_{708.45}$*	**$-1395.11_{1771.37}$* (13)**	$-154.15_{176.65}$* (2)	$-361.52_{714.79}$* (24)
		MECO—Russian		
FFD	$-2.51_{7.49}$	$27.48_{92.70}$ (4)	$-2.60_{6.62}$ (12)	**$-3.19_{7.67}$ (23)**
GD	$-3.82_{34.12}$*	**$-3772.16_{6778.05}$ (18)**	$3.25_{11.56}$* (12)	$-27.24_{59.93}$* (8)
TRT	$-126.33_{327.15}$*	**$-8452.38_{17523.53}$ (18)**	$-4.75_{57.05}$* (7)	$-190.15_{333.11}$* (8)
		MECO—Turkish		
FFD	**$-7.42_{6.62}$***	$-1.99_{22.27}$* (9)	$-5.54_{7.92}$* (12)	$-7.25_{6.48}$* (24)
GD	**$-238.29_{137.52}$*•**	$-149.02_{133.32}$* (12)	$-102.56_{64.73}$*• (9)	$-236.52_{139.32}$*• (24)
TRT	**$-1503.51_{1502.65}$***	$-642.16_{1890.21}$* (18)	$-583.09_{474.45}$* (6)	$-1494.48_{1500.08}$* (24)

Table 9: Δ_{MSE} of 10-fold cross-validation using linear mixed-effects models (LMMs) on per-participant MECO reading times with mGPT-derived surprisal, representations ($\mathbf{h}$; PCA with K=25 components), logit-lens surprisal (s^{LL}), and information value (v). Unlike the main analyses (Table 1), which use regularized regression on reading times averaged across participants, here we retain individual observations and fit LMMs with random intercepts for subjects and documents $r_{s,i} = \mathbf{x}_i^\top \boldsymbol{\beta} + b_s + u_i + \varepsilon_{s,i}$, where $b_s \sim \mathcal{N}(0, \sigma^2_{\text{subj}})$ and $u_i \sim \mathcal{N}(0, \sigma^2_{\text{item}})$. Models are fit by maximum likelihood with `lme4`; test-set predictions use fixed effects only. For each predictor type, we report the best-performing layer and its index ℓ. Bold indicates the best condition per row. Asterisks (*) denote models that significantly outperform the respective models trained on permuted reading times, according to a one-sided paired t-test ($\alpha = 0.001$). Bullets (•) indicate significance over the baseline. Note that the representation results for Russian exhibit high variance across folds, likely due to overfitting of the 25 PCA components on the smaller Russian dataset.