Dual Alignment Between Language Model Layers and Human Sentence Processing

Tatsuki Kuribayashi[1,2] **Alex Warstadt**[3] **Yohei Oseki**[4] **Ethan Gotlieb Wilcox**[5]
[1]MBZUAI [2]Tohoku University [3]UC San Diego
[4]The University of Tokyo [5]Georgetown University
tatsuki.kuribayashi@mbzuai.ac.ae awarstadt@ucsd.edu
oseki@g.ecc.u-tokyo.ac.jp ethan.wilcox@georgetown.edu

Abstract

A recent study (Kuribayashi et al., 2025) has shown that human sentence processing behavior, typically measured on syntactically unchallenging constructions, can be effectively modeled using surprisal from early layers of large language models (LLMs). This raises the question of whether such advantages of internal layers extend to more syntactically challenging constructions, where surprisal has been reported to underestimate human cognitive effort. In this paper, we begin by exploring internal layers that better estimate human cognitive effort observed in syntactic ambiguity processing in English. Our experiments show that, in contrast to naturalistic reading, later layers better estimate such a cognitive effort, but still underestimate the human data. This *dual alignment* sheds light on different modes of sentence processing in humans and LMs: naturalistic reading employs a somewhat weak prediction akin to earlier layers of LMs, while syntactically challenging processing requires more fully-contextualized representations, better modeled by later layers of LMs. Motivated by these findings, we also explore several probability-update measures using shallow and deep layers of LMs, showing a complementary advantage to single-layer's surprisal in reading time modeling.

https://github.com/kuribayashi4/internal_surprisal_targeted_assessment

1 Introduction

A central goal in computational psycholinguistics is to understand human sentence processing through constructing a computational model that can simulate it (Crocker, 2007). Language models (LMs) have offered a framework to explore the candidates for cognitively plausible models, motivated by the widely held view that *prediction* is a core principle of human sentence processing (Clark, 2013; Levy, 2008a; Smith and Levy, 2013). As LMs are fundamentally designed as next-word prediction machines, they have served as a tool to estimate the predictability of words and contributed to exploring the role of prediction in human language, particularly in online sentence processing (Frank and Bod 2011; Goodkind and Bicknell 2018; Hale et al. 2018; Wilcox et al. 2020; Oh and Schuler 2023; Kuribayashi et al. 2025; i.a.).

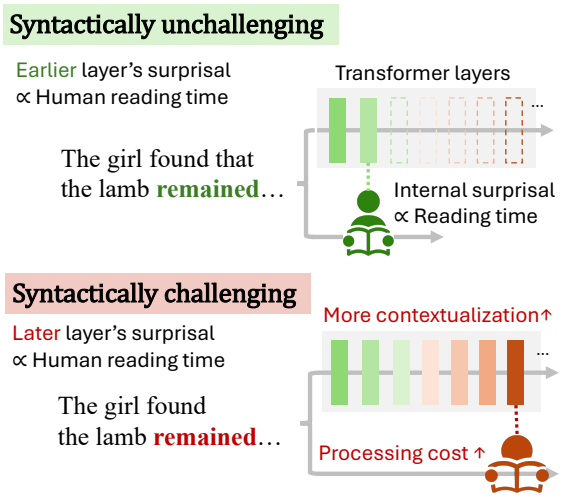


Figure 1: We examine surprisal from internal layers of Transformer LMs to better capture human sentence processing behavior and find that deeper layers align better in syntactically challenging constructions.

Existing studies have demonstrated both the successes and limitations of accurate predictability estimation by modern LMs in cognitive modeling. In particular, surprisal, defined as $-\log p(w_t \mid \mathbf{w}_{<t})$ for a word w_t and a context of previous words $\mathbf{w}_{<t}$ and estimated from some LMs, has proven to be a strong predictor of human reading behavior (Frank and Bod, 2011; Goodkind and Bicknell, 2018; Wilcox et al., 2023). However, surprisal from very large LMs, despite arguably better alignment with ground truth text distribution, has been shown empirically to misalign with human naturalistic reading behaviors throughout an entire corpus (Kuribayashi et al., 2021; Oh and Schuler,

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 46207–46223

2023; Shain et al., 2024; de Varda and Marelli, 2023; Boeve and Bogaerts, 2025). We refer to this type of shortcoming as **holistic misalignment**. At the same time, surprisal from all LMs so far tested has been found to underestimate the cognitive load associated with syntactically challenging constructions, such as garden-path sentences or ungrammatical sentence regions (van Schijndel and Linzen, 2021; Wilcox et al., 2021; Arehalli et al., 2022; Huang et al., 2024; Timkey et al., 2025). We refer to this as **targeted misalignment**. Targeted misalignment has been found in contexts where humans experience high cognitive load and exhibit a substantial slowdown in reading, the magnitude of which is not reflected in models' surprisal values. A recent study (Kuribayashi et al., 2025) addressed the holistic misalignment issue by showing that surprisal decoded from earlier layers of LMs, rather than final layers, better matches human-like reading behavior on syntactically unchallenging, naturalistic corpora.

We ask whether targeted misalignment can be reconciled by using surprisal from internal model layers (§ 3 and § 4). Our experimental results demonstrate that, in contrast to the naturalistic reading results, earlier layers do not better simulate the contrastive human reading slowdown in syntactically challenging contexts. They compute almost the same surprisal in both syntactically ambiguous and unambiguous conditions, reflecting an overly severe recency bias and syntactic insensitivity. Our results, therefore, contrast with those presented in Kuribayashi et al. (2025); earlier layers alone are not a cognitively plausible model of human sentence processing when extended to syntactically challenging contexts.

Widening our investigation to the whole model, we find that later layers align better with syntactic ambiguity processing behavior. However, they still produce an *underestimate* of the human reading data. This *dual alignment* (Figure 1) between LM layers and human sentence processing stages suggests that different stages of human sentence processing may correspond to different layers of LMs; in particular, normal naturalistic processing can be aligned with earlier layers' prediction, while slower, late-stage processing (e.g., reanalysis) demands later layers' more contextualized representations. This partially supports the recent proposed correspondence between LMs' forward computation to human language processing stages (Tenney et al., 2019; Hu et al., 2026; Kuribayashi et al., 2025).

Combining our two empirical findings, we propose using contrastive surprisal from early vs. late layers as a measure of the *degree of belief update* between shallow and fully-contextualized processing. We hypothesize that this measure can be used to identify data points that will be contextually demanding for humans to process (§ 5). This is based on the observation that, generally, syntactically challenging constructions incurred greater qualitative change in surprisal across layers. We exemplify this proposal by showing the effectiveness of surprisal update as a predictor in reading time modeling, but leave a full analysis as a direction for future research.

2 Background

2.1 Surprisal theory

Humans exhibit different cognitive load (e.g., measured by reading time) for different interest areas (e.g., words or tokens) in a text during reading. Surprisal has proven to be a robust predictor of reading time across languages and experimental paradigms (Levy, 2008a; Demberg and Keller, 2008; Wilcox et al., 2023). The surprisal (Cover, 1999) of a word $w_t \in W$ in context $\mathbf{w}_{<t} := [w_0, \cdots, w_{t-1}]^\top$ is defined as $-\log P_t(W = w_t | \mathbf{w}_{<t})$, where $P_t : W \to [0, 1]$ is a family of conditional distributions assigning a probability to a word w at time step t given its prefix. Thus, the more unexpected w_t is, the more costly it is for humans to process. Empirically, this cost has been found to scale linearly with its negative log probability (Smith and Levy, 2013; Shain et al., 2024).

2.2 Internal surprisal from Transformers

To review the decoder-based Transformer architecture, the model consists of a stack of layers that parameterize the anticipation of the next word $P(\cdot \mid \mathbf{w}_{<t})$ by iteratively processing the context through self-attention. Specifically, in each layer l for each token i, the model integrates the previous layer's representations up to and including i to output a representation $\boldsymbol{h}_i^{(l)} \in \mathbb{R}^d$:

$$\boldsymbol{h}_i^{(l)} = \mathcal{F}^{(l)}\left(\boldsymbol{h}_0^{(l-1)}, \ldots, \boldsymbol{h}_i^{(l-1)}\right) \ , \quad (1)$$

$$\boldsymbol{h}_i^{(0)} = \mathrm{emb}(w_i) \ , \quad (2)$$

where $\mathcal{F}^{(l)}$ is the forward computation of layer l. $\boldsymbol{h}_i^{(0)} = \mathrm{emb}(w_i) \in \mathbb{R}^d$ is an input embedding of

Phenomena	Example
MVRR	D^+: The girl fed the lamb ***remained*** relatively calm before the sunset in silence. D^-: The girl who was fed the lamb ***remained*** relatively calm before the sunset in silence.
NPS	D^+: The girl found the lamb ***remained*** relatively calm near the wooden fence. D^-: The girl found that the lamb ***remained*** relatively calm near the wooden fence.
NPZ	D^+: When the girl attacked the lamb ***remained*** relatively calm despite the sudden noise. D^-: When the girl attacked, the lamb ***remained*** relatively calm despite the sudden noise.
RC	D^+: The bus driver that ***the*** kids followed waited patiently at dawn. D^-: The bus driver that ***followed*** the kids waited patiently at dawn.
Attachment	D^+: Janet charmed the executive of the assistants who ***decides*** almost everything during long weekly meetings. D^-: Janet charmed the executives of the assistant who ***decides*** almost everything during long weekly meetings.

Table 1: Examples of pairs of syntactically ambiguous ($\in D^+$) and unambiguous ($\in D^-$) sentences. The underlined parts are minimal differences between the two conditions. The bold part is the disambiguating point t^*, as well as the first word of the region of interest (RoI). Examples are borrowed from Huang et al. (2024).

the word w_i. Layer-specific next word probability is obtained from that layer's representation of the preceding context using logit-lens (LLens; nostalgebraist, 2020):

$$P^{(l)}(W = w_t \mid \mathbf{w}_{<t}) = \text{LLens}(\boldsymbol{h}_{t-1}^{(l)})^{\texttt{id}(w_t)} = \text{softmax}(\boldsymbol{W}_U \text{LayerNorm}(\boldsymbol{h}_{t-1}^{(l)}))^{\texttt{id}(w_t)}, \quad (3)$$

where $\boldsymbol{W}_U \in \mathbb{R}^{|\mathcal{V}| \times d}$ is an unembedding matrix obtained from the LM's output layer, and $|\mathcal{V}| \in \mathbb{R}$ is the model's vocabulary size. The superscript $\texttt{id}(w_t)$ denotes the element corresponding to word w_t in the resulting probability vector.[1] Layer-specific surprisal $S_t^{(l)} = -\log P^{(l)}(W = w_t | \mathbf{w}_{<t}) \in \mathbb{R}_{\geq 0}$ can also be computed. One limitation with Logit Lense is that it has been empirically found to be less reliable for decoding earlier layers, likely because embeddings do not exist in the same representation space (Belrose et al., 2023; Langedijk et al., 2024). We address this limitation in appendix B.1. Kuribayashi et al. (2025) explored which layer l exhibits a better fit to reading time data. They find that earlier layers typically result in the best prediction for the case of naturalistic reading.

2.3 Targeted misalignment of surprisal

There are notable cases in which human reading time patterns cannot be explained by (final layer) surprisal, leading to the criticism that surprisal-based predictability alone is insufficient to characterize total processing cost (van Schijndel and Linzen, 2021; Arehalli et al., 2022; Huang et al., 2024; Timkey et al., 2025; Wilcox et al., 2021). Specifically, these studies recorded contrastive reading behavior between two different conditions, where sentences either conform to or violate structural expectations determined by the grammar. While structural expectations can be based on pure grammaticality (as in Wilcox et al., 2021), they can also be determined by structural ambiguity. In our study, the test materials consist of pairs of syntactically ambiguous ($\in D^+$) and unambiguous sentences ($\in D^-$). Examples are given in Table 1. We choose these phenomena, as managing ambiguity can help us understand the role that context, representations, and prediction play during human language processing. Previous studies find that the reading time difference between D^+ and D^- is underestimated by the magnitude of the surprisal difference between the two conditions.

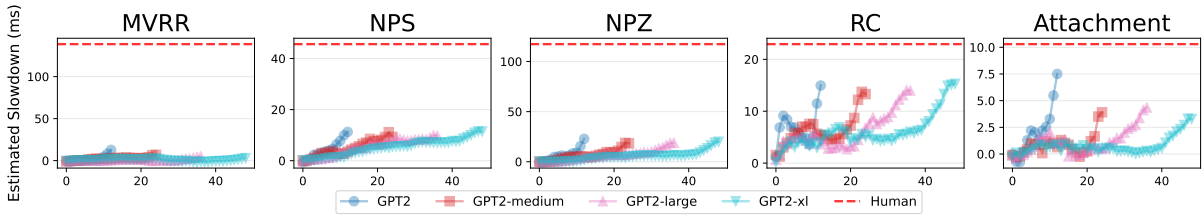
3 Experiment 1

We first extend existing experiments to compare the reading time slowdown in syntactic ambiguity processing with LM-computed surprisal, focusing on internal layers of LMs, not just the last layer.

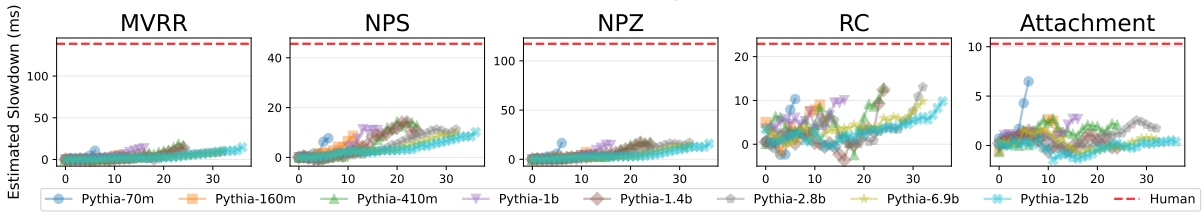
3.1 Settings

Data We use the syntactic ambiguity processing data from Huang et al. (2024), which covers five types of syntactically challenging constructions: (i) Main Verb/Reduced Relative (MVRR); (ii) Noun Phrase or Sentential Complement (NPS); (iii) Noun Phrase Complement or Zero Complement (NPZ); (iv) Object/Subject Relative Clause (RC); and (v)

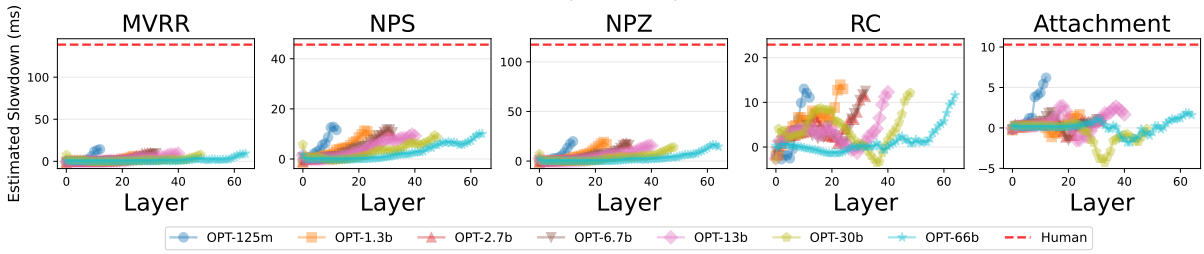
[1] When a word is tokenized into multiple subwords, we follow the existing studies (Oh and Schuler, 2023; Kuribayashi et al., 2025) to compute the joint probability of the subwords.



(a) GPT-2 family



(b) Pythia family



(c) OPT family

Figure 2: Estimated reading time slowdown by layers for each syntactic construction. The red dashed line shows the average observed human slowdown; y-axis varies between plots. Later layers show better alignment, but all model families and all layers underestimate the effect.

High/Low Attachment (Attachment).[2] For each construction, the dataset D contains matched pairs of sentences $(s^+, s^-)_1^{|D|} \in D^+ \times D^-$, where s^+ is from the syntactically challenging condition and s^- is from an unchallenging condition with explicit cues that resolve syntactic ambiguity (see Table 1). For example, in the MVRR construction shown in Table 1, the challenging version ($\in D^+$) is ambiguous: "fed" could be either a main transitive verb or a past participle in a relative clause without the relative pronoun and copula. This ambiguity is resolved only when readers reach "remained." In contrast, the unambiguous version ($\in D^-$) includes the relative pronoun and copula, making the structure immediately clear. Each sentence pair has an annotated disambiguating point t^* where ambiguity is resolved on the D^+ side (and its corresponding position on the D^- side), as shown in bold in Table 1. Slowdowns are observed around t^* in the D^+ condition compared to D^-, and quantifying this magnitude of slowdown is our focus.

[2]We excluded the Agreement part, which targets the effort for processing grammatical violations, as our initial focus is on syntactic ambiguity processing.

The data contains 24 unique sentence pairs for each syntactic phenomenon, resulting in a total of 120 unique pairs with a total of 3,371 tokens. Sentences are annotated with token-level human reading times. Our human reading data comes from Huang et al. (2024), who used a web-based self-paced reading paradigm on over 2K participants, resulting in around 1.2M data points across all the tokens in the dataset, and around 87K data points at the disambiguating points (including the corresponding point in D^- side). In our study, as a preprocessing step, reading times are averaged across participants prior to analysis.

Procedure Our analysis closely resembles that of Wilcox et al. (2021), who estimated predicted reading time slowdowns from LM surprisal. Let $\boldsymbol{w} = [w_1, \cdots, w_n]^\top$ be tokens in the held-out corpus, and let $\boldsymbol{y} = [y_1, \cdots, y_n]^\top$ be their respective reading times. For each token, we select a set of word-level linguistic features $\boldsymbol{f}(w_k) \in \mathcal{R}^m$, including its length in characters, unigram frequency, and surprisal. We then fit a regression model to predict by-token reading time $g : \boldsymbol{f}(w_k) \mapsto \hat{y}_k$

from features.[3] The regression model is trained on the filler-sentence part of the dataset (Huang et al., 2024). Then, this regression model is run on the target data: $s^+ = [w_1^+, \cdots, w_n^+]^\top \in D^+$ and $s^- = [w_1^-, \cdots, w_n^-]^\top \in D^-$. Estimated reading times for each token are obtained ($\hat{\boldsymbol{y}}^+ = [\hat{y}_1^+, \cdots, \hat{y}_n^+]^\top$ and $\hat{\boldsymbol{y}}^- = [\hat{y}_1^-, \cdots, \hat{y}_n^-]^\top$), and we compute the reading time difference at the disambiguating point t^* between the two conditions, yielding surprisal-estimated reading time slowdowns. Note that we compared the reading time difference summed over t^* and $t^* + 1$, given the presence of spillover in the data (Huang et al., 2024). We refer to these as regions of interest (RoIs) in this section. We average this estimated slowdown across all sentence pairs in the dataset. We repeatedly conducted this procedure using surprisal $S_t^{(l)}$ from each layer l obtained with logit lens (§ 2.2) to examine which layer's surprisal better estimates the human reading slowdown between D^+ and D^- conditions.

LMs We examine 19 open-source Transformer LMs including GPT-2 (124M, 355M, 774M, and 1.5B parameters; Radford et al., 2019), OPT (125M, 1.3B, 2.7B, 6.7B, 13B, 30B, and 66B parameters; Zhang et al., 2022), Pythia (70M, 160M, 410M, 1B, 1.4B, 2.8B, 6.9B, and 12B parameters; Biderman et al., 2023). We excluded instruction-tuned models (Kuribayashi et al., 2024).

Surprisal We compute layer-specific surprisal $S_t^{(l)}$ for each token at each layer using the Logit-Lens method (nostalgebraist, 2020) (§ 2.2). Due to the known limitations for using Logit-Lens to examine early layers, we conducted an additional analysis using Tuned-Lens (Belrose et al., 2023), and found no substantial difference in the results (appendix B.1); thus, results in the main text are all from the simpler, Logit-Lens version. We applied Whitespace-Trailing Decoding (Oh and Schuler, 2024) to compute the accurate next-word probabilities for exact disambiguating tokens.

3.2 Results

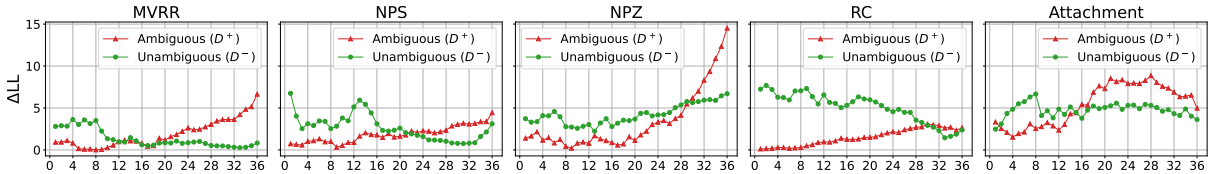
Figure 2 presents the averaged reading time differences between syntactically challenging (D^+) and unchallenging (D^-) conditions by LM surprisal across different layers, alongside the actual human reading time difference (red line; note that the y-axis differs across phenomena). First, it is evident that surprisal from all layers underestimates the human reading time difference, consistent with prior findings on targeted misalignment (van Schijndel and Linzen, 2021; Huang et al., 2024). Second, later layers consistently provide relatively better (even if underestimated) predictions of the reading time difference between D^+ and D^- conditions. This contrasts with previous findings on naturalistic reading, where earlier layers typically yielded superior predictions of human reading times (Kuribayashi et al., 2025). That is, the best layer is notably later in syntactically challenging contexts compared to naturalistic reading scenarios.

3.3 Interim Discussion

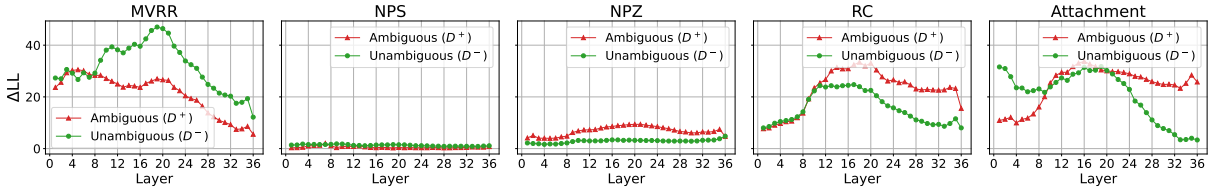
Syntactic insensitivity of earlier layers. Earlier layers' surprisal in RoIs was almost the same between D^+ and D^- conditions (Figure 2), leading to a failure in simulating the contrastive reading time slowdown. One possible explaination is that earlier layers are not sensitive enough to long dependencies and are distracted by local co-occurrences. For example, given the MVRR construction "The girl fed the lamb **remained**...,", an earlier layer might only consider the local co-occurrence of "the lamb **remained**," and therefore assign lower surprisal to "remained" even though it is implausible given the larger context.

Dual alignment. We also observe a systematic shift in which layer's surprisal best approximates human behavior: later layers are more effective for syntactically challenging constructions, whereas prior work has shown that earlier layers better capture naturalistic reading (Kuribayashi et al., 2025). This *dual alignment* suggests that if we attempt to model the reading behavior of syntactically challenging constructions through the lens of prediction, more extensively contextualized representations would be selectively required. This corresponds to a dual-mechanism perspective on human sentence processing (Narayanan and Jurafsky, 1998; van Schijndel and Linzen, 2021), wherein humans may usually read sentences using a relatively shallow processing strategy (aligned with earlier layers) and switch to a deeper, more contextually integrated processing mode (somewhat better aligned with later layers) when confronted

[3] Following work that establishes a linear link between surprisal and reading times (Shain et al., 2024), we use a linear regression model: $\texttt{RT}(w_t) = \beta_0 + \beta_1 \cdot \texttt{Surprisal}(w_t) + \beta_2 \cdot \texttt{Length}(w_t) + \beta_3 \cdot \texttt{LogFreq}(w_t) + \beta_4 \cdot \texttt{Surprisal}(w_{t-1}) + \beta_5 \cdot \texttt{Length}(w_{t-1}) + \beta_6 \cdot \texttt{LogFreq}(w_{t-1}) + \beta_7 \cdot \texttt{Surprisal}(w_{t-2}) + \beta_8 \cdot \texttt{Length}(w_{t-2}) + \beta_9 \cdot \texttt{LogFreq}(w_{t-2}) + \epsilon$. (Appendix A.4)



(a) RoI data points



(b) $\overline{\text{RoI}}$ data points

Figure 3: By-layer PPP of Pythia 12B in the four conditions: $D^+ \cap$ RoI (upper red), $D^- \cap$ RoI (upper green), $D^+ \cap \overline{\text{RoI}}$ (bottom red), and $D^- \cap \overline{\text{RoI}}$ (bottom green). Only in $D^+ \cap$ RoI, better PPPs are from deeper layers.

with syntactically challenging constructions that require reanalysis or complex integration of contextual information. The following sections (§ 4 and § 5) conduct a follow-up analysis to further explore this hypothesis.

4 Experiment 2: Psychometric predictive power analysis with layer shift

The results of our previous study challenge the existing paradigm in LM-based cognitive modeling, which uses a single layer to model all data points. Rather, they suggest that the part of the LM used for modeling human cognition may need to change dynamically depending on the phenomena. Given this view, we conduct a follow-up experiment to extend and clarify these results. Rather than looking at predicted human reading slowdowns, we instead measure models' psychometric predictive power (PPP; Goodkind and Bicknell, 2018), and break results down into our relevant experimental conditions. Given our earlier results, we expect to find that earlier layers better simulate human reading behavior as observed in naturalistic reading (Kuribayashi et al., 2025), while deeper processing is recruited for more syntactically challenging sentences; once the difficulty is resolved, earlier layers again become effective. In this study, we simply ask whether such a shift can be modeled using LMs' layer-wise surprisal, and leave questions about control between shallow vs. deep processing for future research.

4.1 Problem setting

For a simple model of context-dependent layer shift, we split the data points into four conditions based on whether they come from ambiguous sentences, and from regions of interest: $\{D^+, D^-\} \times \{\text{RoI}, \overline{\text{RoI}}\}$. Then, we examine which layer's surprisal best fits data points in each condition based on PPP. If our hypothesis — the advantage of a later layer is a signature of processing difficulty — holds, we expect that the deeper layer yields a better PPP only for $D^+ \times$RoI combination. Hereafter, we include tokens at t^*-2, t^*-1, t^*, t^*+1, and t^*+2 as RoI to consider relative reading time magnitude change around the disambiguating point; all other words constitute the $\overline{\text{RoI}}$ group.

Psychometric predictive power (PPP) We quantify the goodness-of-fit of layer-specific surprisal to data points in each condition: $\{D^+, D^-\} \times \{\text{RoI}, \overline{\text{RoI}}\}$. This is measured by a log-likelihood-based score, ΔLL (i.e., psychometric predictive power; PPP), following existing studies (Goodkind and Bicknell, 2018; Wilcox et al., 2020; Kuribayashi et al., 2021, 2022, 2025; Oh and Schuler, 2023).[4] Specifically, we fit two linear regression models to predict word-by-word reading times: a full model that includes both surprisal and baseline linguistic features, and a reduced model that includes only baseline features.[5] The PPP score is defined as the difference in log-likelihood between these two models: $\Delta\text{LL} = \text{LL}_{\text{full}} - \text{LL}_{\text{baseline}}$, which quantifies how much the addition of surprisal improves the model fit (see Appendix A in Kuribayashi et al., 2025). Higher ΔLL values indicate that surprisal better captures human reading behavior. We repeat-

[4] The total ΔLL over the dataset, not token-level average.

[5] We use the same regression model as in § 3, and for baseline features, we excluded all the surprisal factors.

	MVRR				NPS				NPZ				RC				Attachment			
	D^+		D^-		D^+		D^-		D^+		D^-		D^+		D^-		D^+		D^-	
Model	RoI	$\overline{\text{RoI}}$	RoI	$\overline{\text{RoI}}$	RoI	$\overline{\text{RoI}}$	RoI	$\overline{\text{RoI}}$	RoI	$\overline{\text{RoI}}$	RoI	$\overline{\text{RoI}}$	RoI	$\overline{\text{RoI}}$	RoI	$\overline{\text{RoI}}$	RoI	$\overline{\text{RoI}}$	RoI	$\overline{\text{RoI}}$
GPT2-sm	0.46	−0.99	**0.61**	−0.7	**0.87**	−0.72	−0.15	0.01	−0.69	0.05	**0.89**	0.38	**0.82**	−0.37	−0.87	−0.78	**0.72**	0.39	−0.39	−0.83
GPT2-md	**0.59**	−0.91	0.34	−0.87	**0.90**	−0.68	−0.7	−0.42	**0.74**	−0.24	0.52	0.39	**0.81**	−0.43	−0.84	−0.82	**0.42**	−0.08	−0.54	−0.95
GPT2-lg	**0.53**	−0.98	−0.33	−0.91	**0.88**	−0.58	−0.62	0.71	**0.88**	−0.65	0.13	0.33	**0.91**	−0.82	−0.85	−0.83	**0.08**	−0.23	**0.08**	−0.96
GPT2-xl	**0.88**	−0.96	−0.72	−0.86	**−0.07**	−0.3	−0.85	−0.1	**0.88**	−0.73	0.57	0.33	**0.96**	−0.76	−0.44	−0.85	**−0.32**	−0.45	−0.51	−0.97
OPT-125m	**0.73**	−0.57	0.31	−0.26	**0.97**	−0.14	−0.48	0.13	**0.89**	−0.2	0.75	0.55	**0.81**	−0.09	−0.56	−0.12	0.51	**0.92**	−0.91	−0.84
OPT-1.3b	**0.67**	−0.97	−0.77	−0.98	**0.81**	0.52	0.23	−0.44	**0.83**	−0.89	−0.46	0.32	**0.83**	−0.65	0.71	−0.75	**0.40**	−0.0	−0.65	−0.97
OPT-2.7b	**0.70**	−0.95	−0.78	−0.98	**0.86**	0.35	0.14	−0.27	**0.86**	−0.89	0.21	−0.3	**0.81**	−0.15	0.57	−0.48	**0.18**	0.05	−0.63	−0.96
OPT-6.7b	**0.23**	−0.96	0.06	−0.98	**0.74**	0.43	−0.4	−0.42	**0.49**	−0.91	−0.57	−0.35	**0.87**	−0.59	0.25	−0.73	**−0.20**	−0.26	−0.26	−0.93
OPT-13b	**0.09**	−0.96	−0.71	−0.97	**0.71**	0.14	0.15	−0.76	**0.81**	−0.87	−0.19	0.03	**0.88**	−0.65	0.74	−0.83	**0.26**	−0.43	−0.22	−0.97
OPT-30b	−0.13	−0.86	**0.29**	−0.9	**0.61**	0.39	−0.52	0.16	0.13	−0.76	−0.08	**0.32**	**0.82**	−0.61	−0.38	−0.8	−0.42	**−0.22**	−0.35	−0.91
OPT-66b	**0.07**	−0.88	−0.07	−0.83	0.59	**0.65**	−0.49	−0.49	**0.48**	−0.96	−0.2	−0.61	**0.77**	−0.36	0.53	−0.59	**0.66**	0.32	0.19	−0.93
PYT-70m	**0.04**	−0.86	−0.02	−0.86	**0.74**	−0.69	0.41	−0.42	0.3	0.39	**0.76**	0.41	**−0.13**	−0.62	−0.85	−0.87	**0.88**	0.59	−0.51	−0.14
PYT-160m	**0.31**	−0.93	0.16	−0.67	0.24	−0.11	**0.50**	−0.49	0.29	0.19	0.09	**0.58**	**0.52**	−0.5	−0.76	−0.06	**0.91**	0.6	−0.55	−0.42
PYT-410m	**0.05**	−0.89	−0.34	−0.86	**0.82**	−0.61	−0.2	−0.55	0.65	0.58	0.65	**0.66**	**0.92**	0.28	−0.67	−0.33	0.06	**0.49**	−0.69	−0.94
PYT-1b	**0.70**	−0.95	−0.46	−0.94	**0.69**	−0.01	−0.58	−0.62	**0.57**	−0.64	−0.56	0.17	**0.90**	−0.87	−0.81	−0.89	**0.70**	0.6	−0.76	−0.85
PYT-1.4b	**0.55**	−0.96	−0.15	−0.91	**0.80**	−0.27	0.51	−0.57	**0.50**	−0.15	−0.71	0.01	**0.91**	0.8	−0.86	−0.39	**0.46**	0.28	−0.83	−0.9
PYT-2.8b	**0.80**	−0.95	−0.46	−0.88	0.44	−0.58	**0.51**	−0.4	**0.64**	−0.65	−0.54	0.25	**0.90**	−0.68	−0.94	−0.85	**0.72**	0.3	−0.46	−0.86
PYT-6.9b	**0.54**	−0.75	−0.21	−0.82	**0.96**	−0.56	−0.7	−0.82	0.75	0.55	0.7	**0.96**	**0.94**	0.28	−0.94	−0.54	**0.82**	0.53	−0.3	−0.77
PYT-12b	**0.88**	−0.89	−0.83	−0.41	**0.93**	−0.43	−0.69	−0.82	**0.79**	0.38	0.78	0.74	**0.97**	0.46	−0.92	−0.22	**0.80**	0.56	0.04	−0.76

Table 2: Correlation between layer depth and PPP by model and condition. $D^+ \cap$ RoI exhibits positive correlations.

edly compute PPP for each layer in each of the four conditions: $\{D^+, D^-\} \times \{\text{RoI}, \overline{\text{RoI}}\}$.

Measure In each condition, we report Pearson's correlation coefficient between layer depth and PPP. A higher correlation indicates the tendency of later layers to better simulate the respective reading time, which is expected only in the $D^+ \cap$ RoI condition.

4.2 Results

Let us begin with observing a representative pattern from Pythia 12B model in the four different conditions (Figure 3). The figure reveals that the increasing PPP toward deeper layers is distinctive in the $D^+ \cap$ RoI condition (top red lines). Table 2 summarizes the correlation between layer depth and ΔLL, providing a complementary view of the layer-wise trend. A positive correlation indicates that deeper layers progressively improve the prediction of human reading behavior. Positive correlations are consistently observed for $D^+ \cap$ RoI condition across all models and constructions. This corroborates that human reading behavior under syntactic ambiguity particularly aligns with deeper layers, consistent with our expectation.

Notably, the contrast in correlation across conditions $\{D^+, D^-\} \times \{\text{RoI}, \overline{\text{RoI}}\}$ becomes more pronounced in larger models. For example, Pythia-12B shows correlations of 0.88 (RoI in D^+) vs. -0.83 (RoI in D^-) for MVR, and 0.93 (RoI in D^+) vs. -0.69 (RoI in D^-) for NPS, while these are somewhat attenuated in smaller ones, indicating that as models scale up, they develop a clearer differentiation in how layer depth relates to syntactically challenging versus unchallenging reading.

5 Experiment 3: Probability-update as processing effort

Our experiments so far have converged on the finding that surprisal from deeper layers with extensive contextualization better captures syntactic ambiguity processing, a behavior that requires substantial cognitive effort. One interpretation of this result is that it arises because human processing in such regions involves an initial shallow prediction with surface-level features, such as unigram frequency or local co-occurrence information, which must be subsequently revised by considering a broader linguistic context, incurring a higher processing cost. In this section, we propose a method for identifying such data points by inspecting the difference in predictions between LM layers, i.e., the advantage that deeper contextualization buys you for prediction.

5.1 Measurements

We explore several information-theoretic measures that quantify the change in predictive distributions between shallow and deep processing. Our first formulation computes the *change* in surprisal of a word w_t between a shallow and a deep layer, rather than just computing surprisal at a certain layer. We term this the **surprisal update** (SU) defined as:

$$\begin{aligned} \text{SU}(w_t|\mathbf{w}_{<t}) &= S_t^{\text{shallow}} - S_t^{\text{deep}} \\ &= -\log P_t(w_t|\mathbf{w}_{<t}) - (-\log Q_t(w_t|\mathbf{w}_{<t})) \\ &= \log \frac{Q_t(w_t|\mathbf{w}_{<t})}{P_t(w_t|\mathbf{w}_{<t})} \quad . \end{aligned} \tag{4}$$

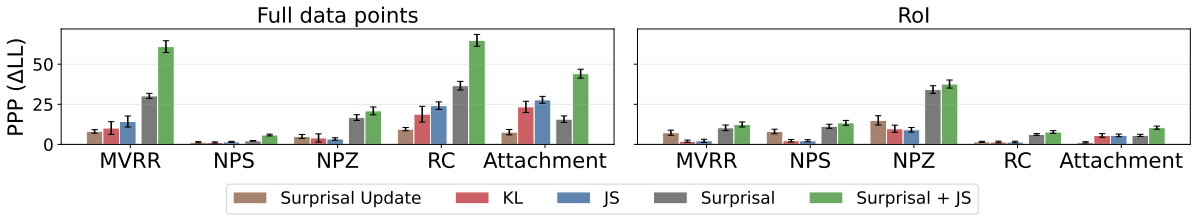


Figure 4: PPP obtained by probability-update measurements introduced in § 5.1 on ten different data conditions {5 phenomena}×{Full, RoI}. PPP scores from 19 LMs are averaged, and error bars indicate 95% confidence interval. PPP with surprisal (gray) and the one with both surprisal and JS divergence (green) are also reported.

where $P_t(w_t|\mathbf{w}_{<t})$ and $Q_t(w_t|\mathbf{w}_{<t})$ denote the probability of w_t in context estimated by the first and last layer, respectively.[6] We henceforth denote the corresponding probability distributions over the model's subword vocabulary as P_t and Q_t.

As a second measure, we extend surprisal update to the full next-(sub)word distribution using the **Kullback-Leibler divergence** (KL), i.e., the expected value of SU under Q_t:

$$\begin{aligned}\mathrm{KL}(Q_t||P_t) &= \sum_{w\in W} Q_t(w|\mathbf{w}_{<t}) \log \frac{Q_t(w|\mathbf{w}_{<t})}{P_t(w|\mathbf{w}_{<t})} \\ &= \mathbb{E}_{w\sim Q_t}\left[\mathrm{SU}(w|\mathbf{w}_{<t})\right] \quad . \end{aligned} \tag{5}$$

This quantifies the discrepancy between the final prediction Q_t and the initial one P_t. As the KL is asymmetric, we also examine its symmetric version, namely, **Jensen–Shannon divergence** (JS):

$$\begin{aligned}\mathrm{JS}(Q_t||P_t) &= \frac{1}{2}\left(\mathrm{KL}(Q_t||M_t) + \mathrm{KL}(P_t||M_t)\right) \\ M_t &:= \frac{1}{2}\left(P_t + Q_t\right) \quad . \end{aligned} \tag{6}$$

We compute these measures for each token[7] and assess their effectiveness in modeling human reading times. We hypothesize that larger discrepancies between shallow and deep predictions, as quantified by SU, KL, and JS, may be associated with higher processing effort, as they indicate a greater room to revise initial predictions based on broader context. Conversely, if surprisal or probability distribution does not change much after extensive contextualization, it suggests that the word is easier to integrate, requiring less cognitive effort. Our proposal is similar in spirit to the information-theoretic model of shallow vs. deep processing presented in Li and Futrell (2024).

[6] In the regression analysis, we applied Z-score normalization to each layer's surprisals before computing surprisal-update to mitigate layer-dependent scale differences.

[7] KL and JS are computed for each subword position, and if a token consists of multiple subwords, we simply sum up the subword-level scores, similarly to cumulative surprisal.

5.2 Problem setting

As in Experiment 2, we model reading time in data from Huang et al. (2024) using a linear regression model. We replace surprisal with one of our probability-update predictors (i.e., SU, KL, or JS). To evaluate whether the probability-update captures processing effort difference across challenging and unchallenging conditions, we report results for RoI regions (as defined in § 4) as well as for full data points, yielding ten conditions: {5 phenomena}×{Full, RoI}. Note that each condition includes both syntactically challenging and unchallenging items.

5.3 Results

Figure 4 shows the average PPP across 19 models for each data condition. Detailed results with likelihood ratio test (Wilks, 1938) are also shown in Tables 6, 7, and 8 in Appendix B.2. Consistent with our hypothesis, probability-update measures significantly improve the model fit in most cases, in both full and RoI conditions. This indicates that words associated with larger probability updates correspond to increased reading times in humans. Among the three measures, JS typically exhibited the best PPP empirically.

Nevertheless, following the setting of § 3, the estimated slowdown by probability-update measures was under 10 ms (Appendix B.3). In addition, as shown in Figure 4, PPP gains are more nuanced for the RoI data points compared to the full data conditions. Thus, we tentatively conclude that, while probability-update measures explored in this section are potentially effective features for reading time modeling, their advantage is not specifically associated with the processing cost of syntactically ambiguous sentences.

5.4 Analysis

Is the advantage of the probability-update measure in terms of PPP orthogonal to surprisal? Taking the JS value as an example, we evaluate the additive effect of the JS to the last layer's surprisal. Figure 4 also includes the results for surprisal (gray) and both surprisal and JS (green), where these features are added to the same baseline regression model (§ 3.1). As shown in Figure 4, the Surprisal + JS (green) setting typically exhibits an advantage over the Surprisal ones, supporting the complementary effect of JS-based probability-update to the commonly-used surprisal feature. In Appendix B.4, likelihood ratio tests are conducted between the two nested regression models: one with surprisal plus baseline features and the other with surprisal, JS, and baseline features, showing that MVR (Full), RC (Full), and Attachment (Full/RoI) settings tend to yield statistical significance.

6 Discussion

We have investigated the alignment between internal-layer surprisal from LMs and human reading behavior in syntactically challenging constructions. It is a natural extension of previous work connecting layer-wise dynamics to *offline* language processing (Tenney et al., 2019; He et al., 2024; Hu et al., 2026). On the one hand, the dual alignment we observe is consistent with the shallow vs. deep processing that characterizes human language comprehension (Barton and Sanford, 1993; Christianson et al., 2001; Ferreira et al., 2002). This suggests that LMs have stages of processing, perhaps analogous to the two-stage model of human sentence processing (see discussion in van Schijndel and Linzen, 2021). On the other hand, our results pose challenges to cognitive modeling with LMs. First, while human processing dynamics unfold over time, model dynamics unfold over internal layers, thus requiring another (perhaps thorny) theoretical link between the two. Second, we are unable to identify a single component of an LM that is the optimal basis for modeling (all) human language processing behavior. While early layers function best for naturalistic reading are relatively worse predictors for reading times of ambiguous sentences. Finally, in human reading, reanalysis is thought to be a specialized operation, triggered by abnormally high surprisal (Warner and Glass, 1987; Levy et al., 2008) or entropy (Botvinick et al., 2001; Ness et al., 2025). In contrast, LM architectures apply essentially the same computational operations, i.e., highly contextualized processing, to all inputs.

Our contribution sharpens what commitments must be made about LMs when using them as cognitive models of human processing. Rather than using LM surprisal as one single model to capture human sentence processing, we should look to *parts* of LMs as models of sub-processes or components of human language processing. In this spirit, we suggest several probability-update measures over internal layers as a way to model the cognitive *distance* between early- and late-stage processing when processing a word.

Our exploration of probability updates across layers in relation to cognitive cost (§ 5) connects to the Bayesian approaches to sentence processing (Ratcliff, 1978; Narayanan and Jurafsky, 1998; Levy, 2008b; Ratcliff and McKoon, 2008; Norris, 2006, 2009; Itti and Baldi, 2009). While these theories typically focus on incremental, input-driven updates to the probability distribution, our analysis may align with the memory-based contextual integration, whereby internal layers iteratively refine representations and update predictive distribution. Establishing theoretical links between our approach and psycholinguistic theories is an important future work, both for better situating LMs as a tool for psycholinguistics and for addressing an NLP interpretability question — if the goal is to identify syntactic ambiguity processing within an LM, which approach is more appropriate?

7 Conclusions

This study provides evidence that surprisal from later layers with richer contextualization better captures human ambiguity processing. This contrasts with prior work on holistic reading-time modeling, which reported stronger alignment with earlier-layer surprisal (Kuribayashi et al., 2025), and we introduce a *dual alignment* between LM layers and human sentence processing. These findings suggest that, rather than treating LM surprisal as a single unified predictor of human sentence processing, it may be more fruitful to look to *parts* of LMs as models of sub-processes or components of human language processing. In this spirit, we propose probability-update measures across layers as a way to quantify the cognitive distance between early- and late-stage processing, demonstrating the potential of layer-contrastive information-theoretic measures for modeling sentence processing effort.

Limitations

We only investigated English LMs and human reading data in the English syntactic ambiguity processing. Extending the analysis to other languages with different syntactic structures and ambiguity types will enhance the universality of our findings. We used self-paced reading time data from Huang et al. (2024), and a concurrent study has released eye-tracking corpora (Timkey et al., 2025). The use of eyetracking data will provide fine-grained information about layer-wise alignment; for example, it is likely that first-pass forward reading time aligns with earlier layers, while later regressive behavior relatively aligns with deeper layers. Integrating reading behavior data over grammatical violations (Wilcox et al., 2021), which has also been underestimated by surprisal, will be worth exploring, and similar results may be expected, given that earlier layers' surprisal was not sensitive to grammatical structure. More generally, our paper alone does not answer whether the dual alignment is specifically related to syntactic ambiguity or generalizes to other well-studied sources of processing difficulty, such as agreement attraction, long-distance dependencies, or similarity-based interference. More controlled experiments will clarify our findings.

On the LM side, there are several technical issues. First, internal probabilities are obtained from logit-lens (nostalgebraist, 2020). Although, at least compared to tuned-lens (Belrose et al., 2023), the selection of the probability extraction method did not substantially affect the results, our analysis relies heavily on the method for obtaining internal probabilities, potentially leading to methodological biases. Second, for KL- and JS-based probability update measures, we treat the LMs' subword vocabulary as the vocabulary space, which would not be cognitively plausible. More generally, the token granularity of LMs has been reported to affect the quality of information-theoretic values, e.g., surprisal, and cognitive modeling results (Nair and Resnik, 2023; Oh and Schuler, 2025), and this issue also applies to our experiments. Third, exploring other variants of layer-contrastive information-theoretic measures beyond the metrics studied in this paper and establishing their theoretical connections to psycholinguistic theories will also be a promising future direction.

Ethical Statement

This study conducts the analysis of publicly available datasets of human reading behavior and LMs. We expect that these artifacts were collected and released following appropriate ethical guidelines in existing studies.

AI Writing/Coding Assistance Policy

We used generative AI tools solely for the purpose of adjusting the grammar and phrasing of the manuscript, and a coding assistant for formatting tables and figures.

Acknowledgements

This work was supported by JSPS Grant-in-Aid for Early Career Scientists Grant Number JP23K16938, JSPS KAKENHI Grant Number JP24H00087, JST CREST Grant Number JPMJCR2565, JST PRESTO Grant Number JPMJPR21C2, and JST BOOST Grant Number JPMJBY24B2.

References

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of CoNLL 2022*, pages 301–313.

S B Barton and A J Sanford. 1993. A case study of anomaly detection: shallow semantic processing and cohesion establishment. *Mem. Cognit.*, 21(4):477–487.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Lev McKinney, Igor Ostrovsky, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint.*

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of ICML 2023*, pages 2397–2430. PMLR.

Sam Boeve and Louisa Bogaerts. 2025. A systematic evaluation of dutch large language models' surprisal estimates in sentence, paragraph and book reading. *Behav. Res. Methods*, 57(9):266.

Matthew M Botvinick, Todd S Braver, Deanna M Barch, Cameron S Carter, and Jonathan D Cohen. 2001. Conflict monitoring and cognitive control. *Psychol. Rev.*, 108(3):624–652.

K Christianson, A Hollingworth, J F Halliwell, and F Ferreira. 2001. Thematic roles assigned along the garden path linger. *Cogn. Psychol.*, 42(4):368–407.

Andy Clark. 2013. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.*, 36(3):181–204.

T M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

Matthew W Crocker. 2007. Computational psycholinguistics. *The Handbook of Computational Linguistics and Natural Language Processing*.

Andrea de Varda and Marco Marelli. 2023. Scaling in cognitive modelling: a multilingual approach to human reading times. In *Proceedings of ACL 2023*, pages 139–149.

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

Fernanda Ferreira, Karl G.D. Bailey, and Vittoria Ferraro. 2002. Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11(1):11–15.

Stefan L Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–834.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2018. The Natural Stories Corpus. In *Proceedings of LREC 2018*, pages 76–82.

Richard Futrell, Roger P. Levy, and Edward Gibson. 2020. Dependency locality as an explanatory principle for word order. *Journal of Language*.

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*, pages 10–18.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan R. Brennan. 2018. Finding Syntax in Human Encephalography with Beam Search. In *Proceedings of ACL 2018*, pages 2727–2736.

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R. Brennan. 2024. Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. In *Proceedings of LREC-COLING 2024*, pages 4488–4497.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Jennifer Hu, Michael A. Lepori, and Michael Franke. 2026. Signatures of human-like processing in transformer forward passes. In *First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models*.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Laurent Itti and Pierre Baldi. 2009. Bayesian surprise attracts human attention. *Vision Res.*, 49(10):1295–1306.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. 2024. Psychometric predictive power of large language models. In *Findings of NAACL 2024*, pages 1983–2005.

Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. Context limitations make neural language models more human-like. In *Proceedings of EMNLP 2022*, pages 10421–10436.

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.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *TACL*, 13:1743–1766.

Anna Langedijk, Hosein Mohebbi, Gabriele Sarti, Willem Zuidema, and Jaap Jumelet. 2024. DecoderLens: Layerwise interpretation of encoder-decoder transformers. In *Findings of NAACL 2024*, pages 4764–4780.

Roger Levy. 2008a. Expectation-based syntactic comprehension. *Journal of Cognition*, 106(3):1126–1177.

Roger Levy. 2008b. A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of EMNLP 2008*, pages 234–243.

Roger Levy, Florencia Reali, and Thomas Griffiths. 2008. Modeling the effects of memory on human online sentence processing with particle filters. *Proceedings of NIPS 2008*, 21.

Jiaxuan Li and Richard Futrell. 2024. An information-theoretic model of shallow and deep language comprehension. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Sathvik Nair and Philip Resnik. 2023. Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship? In *Findings of EMNLP2023*.

Srini Narayanan and Daniel Jurafsky. 1998. Bayesian models of human sentence processing. In *Proceedings of CogSci 1998*, pages 752–757. Routledge.

Tal Ness, Valerie J Langlois, Albert E Kim, and Jared M Novick. 2025. The state of cognitive control in language processing. *Perspect. Psychol. Sci.*, 20(2):219–240.

Dennis Norris. 2006. The bayesian reader: explaining word recognition as an optimal bayesian decision process. *Psychological review*, 113(2):327.

Dennis Norris. 2009. Putting it all together: a unified account of word recognition and reaction-time distributions. *Psychol. Rev.*, 116(1):207–219.

nostalgebraist. 2020. interpreting GPT: the logit lens. Blog post, retrieved 20 June, 2025.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *TACL*, 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 the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4150–4162, Vienna, Austria. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. OpenAI blog.

Roger Ratcliff. 1978. A theory of memory retrieval. *Psychol. Rev.*, 85(2):59–108.

Roger Ratcliff and Gail McKoon. 2008. The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput.*, 20(4):873–922.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*.

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

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

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of ACL 2019*, pages 4593–4601.

William Timkey, Kuan-Jung Huang, Byung-Doh Oh, Grusha Prasad, Suhas Arehalli, Tal Linzen, and Brian Dillon. 2025. Eye movements reveal a dissociation between prediction and structural processing in language comprehension. *PsyArXiv*.

Marten van Schijndel and Tal Linzen. 2021. Single-Stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6):e12988.

John Warner and Arnold L Glass. 1987. Context and distance-to-disambiguation effects in ambiguity resolution: Evidence from grammaticality judgments of garden path sentences. *J. Mem. Lang.*, 26(6):714–738.

Ethan Wilcox, Pranali Vani, and Roger Levy. 2021. A targeted assessment of incremental processing in neural language models and humans. In *Proceedings of ACL 2021*, pages 939–952.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the Predictions of Surprisal Theory in 11 Languages. *TACL*, 11:1451–1470.

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 CogSci 2020*, pages 1707–1713.

S S Wilks. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.*, 9(1):60–62.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP 2020: System Demonstrations*, pages 38–45.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. *arXiv preprint*, cs.CL/2205.01068v4.

Appendix

A Artifacts

A.1 Languauge models

Table 3 lists the LMs we used.

A.2 Data and tools

Table 4 lists data and tools we used. The experiments used a single NVIDIA RTX 6000 Ada GPU for several hours for surprisal computation.

A.3 Reading material statistics

We posit the assumption that the SAP dataset (Huang et al., 2024) involves more syntactically challenging constructions than other naturalistic reading corpora. As a quantitative support for this assumption, we computed two general syntactic complexity measures for different corpora: (i) average syntactic tree depth and (ii) average dependency length of a sentence. Note that these two scores are normalized by average sentence length, and length control is usually done for a fair inter-corpus comparison (Futrell et al., 2020). These measures are computed with the Spacy `en_core_web_sm` parser (Honnibal et al., 2020). In addition to SAP data, we analyzed two other naturalistic reading time corpora of the Natural Stories Corpus (NSC) (Futrell et al., 2018) and the Dundee Corpus (Kennedy et al., 2003).

Table 5 shows the statistics. These statistics show that SAP is, on average, syntactically more complex. Note that even in the NSC, one of the relatively syntactically challenging corpus used in naturalistic reading time modeling experiments, including Kuribayashi et al. (2025), only 0.02% sentences, for example, have MVRR ambiguity (see Appendix in Futrell et al. (2018)); we believe that large-scale targeted datasets such as SAP would still be a reasonable option for our research purpose to obtain statistically reliable results.

It will also be interesting to analyze the effective layer depth even within the naturalistic reading time corpus by dividing it into relatively syntactically challenging data points and others, which will be future work, and our study opens such new analysis directions.

A.4 Regression models

In § 3 and § 4, we use a linear regression model:

$$\begin{aligned}\texttt{RT}(w_t) &= \beta_0 + \beta_1 \cdot \texttt{Surprisal}(w_t) + \beta_2 \cdot \texttt{Length}(w_t) \\ &+ \beta_3 \cdot \texttt{LogFreq}(w_t) + \beta_4 \cdot \texttt{Surprisal}(w_{t-1}) \\ &+ \beta_5 \cdot \texttt{Length}(w_{t-1}) + \beta_6 \cdot \texttt{LogFreq}(w_{t-1}) \\ &+ \beta_7 \cdot \texttt{Surprisal}(w_{t-2}) + \beta_8 \cdot \texttt{Length}(w_{t-2}) \\ &+ \beta_9 \cdot \texttt{LogFreq}(w_{t-2}) + \epsilon. \end{aligned} \tag{7}$$

For the baseline, we used the regression model without $\texttt{Surprisal}(w_t)$ $\texttt{Surprisal}(w_{t-1})$, and $\texttt{Surprisal}(w_{t-2})$. Word frequency is computed with `wordfreq` package (Speer, 2022), and word length is character-based. Surprisal is computed with an intra-sentential context.

In § 5.4, we analyzed the additive effect of JS to surprisal, where the full model is:

$$\begin{aligned}\texttt{RT}(w_t) &= \beta_0 + \beta_1 \cdot \texttt{Surprisal}(w_t) + \beta_2 \cdot \texttt{JS}(w_t) \\ &+ \beta_3 \cdot \texttt{Length}(w_t) + \beta_4 \cdot \texttt{LogFreq}(w_t) \\ &+ \beta_5 \cdot \texttt{Surprisal}(w_{t-1}) + \beta_6 \cdot \texttt{JS}(w_{t-1}) \\ &+ \beta_7 \cdot \texttt{Length}(w_{t-1}) + \beta_8 \cdot \texttt{LogFreq}(w_{t-1}) \\ &+ \beta_9 \cdot \texttt{Surprisal}(w_{t-2}) + \beta_{10} \cdot \texttt{JS}(w_{t-2}) \\ &+ \beta_{11} \cdot \texttt{Length}(w_{t-2}) + \beta_{12} \cdot \texttt{LogFreq}(w_{t-2}) + \epsilon. \end{aligned} \tag{8}$$

B Supplimentary results

B.1 Tuned-lens

We also preliminarily performed the main experiment § 3 with TunedLens (Belrose et al., 2023). Figure 5 shows the results for models whose tuned-lens parameters are publicly available. The patterns generally hold the same as those with LogitLens (Figure 2); no layer can simulate the degree of human reading slowdown. These results motivate us to focus on experiments with the simpler LogitLens in the subsequent analyses.

B.2 PPP by probability-update measures

Tables 6, 7, and 8 show the model-wise breakdown of PPPs, which are aggregated in Figure 4.

B.3 Slowdown estimates by probability-update measures

Tables 9, 10, and 11 show the estimated reading time slowdowns for t* and t*+1 data points, following the method of § 3.

B.4 Additive effect of JS to surprisal

Table 12 shows the model-wise breakdown of the additive effect of the JS value on top of surprisal. For the log likelihood ratio test, we compare the likelihood of two nested models: (i) the above full model vs. (ii) the model without JS factors (but still with surprisals). That is, the difference in degrees of freedom between the two models is three.

Model	URL	#params
GPT2-small	https://huggingface.co/gpt2	117M
GPT2-medium	https://huggingface.co/gpt2-medium	345M
GPT2-large	https://huggingface.co/gpt2-large	774M
GPT2-xl	https://huggingface.co/gpt2-xl	1B
OPT-125m	https://huggingface.co/facebook/opt-125m	125M
OPT-1.3b	https://huggingface.co/facebook/opt-1.3b	1.3B
OPT-2.7b	https://huggingface.co/facebook/opt-2.7b	2.7B
OPT-6.7b	https://huggingface.co/facebook/opt-6.7b	6.7B
OPT-13b	https://huggingface.co/facebook/opt-13b	13B
OPT-30b	https://huggingface.co/facebook/opt-30b	30B
OPT-66b	https://huggingface.co/facebook/opt-66b	66B
Pythia-70m-deduped	https://huggingface.co/EleutherAI/pythia-70m-deduped	70M
Pythia-160m-deduped	https://huggingface.co/EleutherAI/pythia-160m-deduped	160M
Pythia-410m-deduped	https://huggingface.co/EleutherAI/pythia-410m-deduped	410M
Pythia-1b-deduped	https://huggingface.co/EleutherAI/pythia-1b-deduped	1B
Pythia-1.4b-deduped	https://huggingface.co/EleutherAI/pythia-1.4b-deduped	1.4B
Pythia-2.8b-deduped	https://huggingface.co/EleutherAI/pythia-2.8b-deduped	2.8B
Pythia-6.9b-deduped	https://huggingface.co/EleutherAI/pythia-6.9b-deduped	6.9B
Pythia-12b-deduped	https://huggingface.co/EleutherAI/pythia-12b-deduped	12B

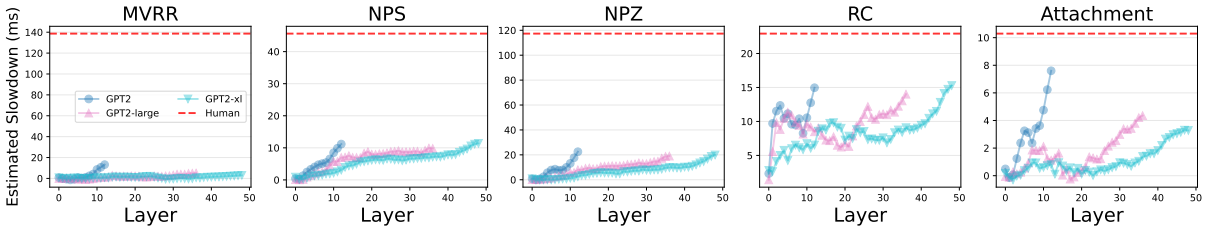
Table 3: LM details

Artifact	License	Usage
Statsmodels (Seabold and Perktold, 2010)	BSD 3-Clause "New" or "Revised" License	To train and run regression models
Syntactic Ambiguity Processing Benchmark (Huang et al., 2024) (https://github.com/caplabnyu/sapbenchmark)	MIT License	To use human reading time data
TunedLens package (Belrose et al., 2023) (https://github.com/AlignmentResearch/tuned-lens)	MIT License	To compute next-word probabilities
WordFreq (Speer, 2022) (https://github.com/rspeer/wordfreq)	Apache 2.0	To compute unigram frequency of words
Transformers (Wolf et al., 2020) (https://github.com/huggingface/transformers)	Apache 2.0	To download and run models

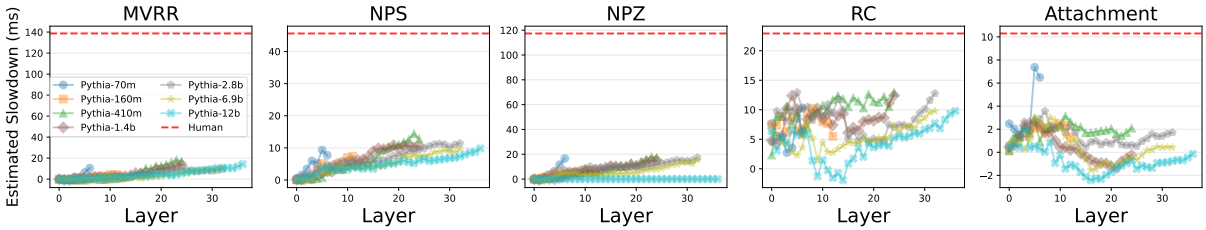
Table 4: Artifacts used in this paper.

Corpus	Tree depth	Dep. length
SAP	**0.396**	**0.152**
NSC	0.302	0.139
Dundee corpus	0.313	0.139

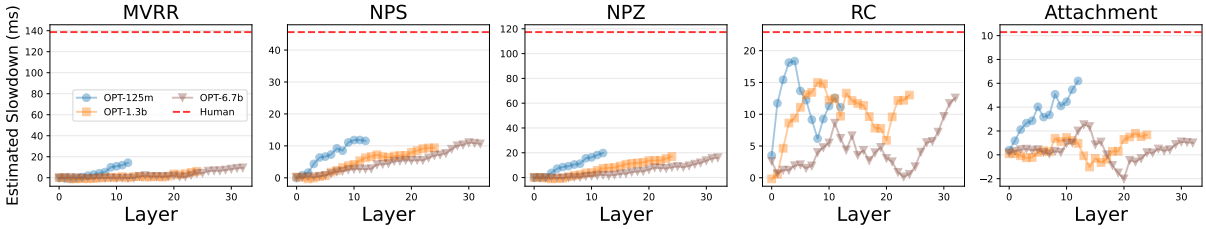
Table 5: Corpus statistics on syntactic complexity



(a) GPT-2 family



(b) Pythia family



(c) OPT family

Figure 5: Estimated reading time slowdown by layers for each syntactic construction with TunedLens (Belrose et al., 2023)

	MVRR		NPS		NPZ		RC		Att.	
Model	All	RoI	All	RoI	All	RoI	All	RoI	All	RoI
G2-sm	11.7†	14.1†	1.1	13.8†	7.8†	24.5†	12.0†	1.7	4.5†	0.8
G2-md	10.8†	8.7†	0.8	6.9†	12.5†	26.3†	13.6†	1.5	4.9†	1.9
G2-lg	11.3†	7.0†	0.7	10.3†	10.5†	27.0†	14.1†	1.8	3.2	2.2
G2-xl	11.4†	5.5†	0.8	7.3†	10.5†	28.1†	12.3†	1.6	4.2†	2.1
OT-125m	11.5†	13.5†	0.4	11.9†	3.8	20.1†	11.6†	1.9	4.5†	1.1
OT-1.3b	7.8†	7.8†	0.6	8.2†	6.8†	20.9†	12.7†	2.1	4.8†	1.9
OT-2.7b	8.3†	11.0†	0.5	7.0†	6.3†	20.0†	11.1†	1.7	4.9†	1.7
OT-6.7b	8.2†	7.1†	0.8	8.3†	5.9†	19.0†	10.1†	1.3	5.6†	1.8
OT-13b	8.4†	11.1†	0.4	6.6†	5.5†	17.8†	8.4†	2.6	7.3†	1.8
OT-30b	6.7†	3.1	0.7	3.7	5.9†	10.5†	7.3†	2.0	7.8†	1.4
OT-66b	7.9†	7.7†	0.6	5.1†	4.3†	15.3†	10.7†	1.8	6.9†	1.1
PT-70m	6.8†	2.6	1.6	1.5	3.5	8.7†	9.6†	0.3	2.4	0.6
PT-160m	7.8†	1.6	1.5	1.6	11.3†	0.9	6.7†	0.7	5.8†	4.2†
PT-410m	8.9†	8.8†	0.1	11.9†	2.4	14.8†	14.0†	0.8	9.3†	0.9
PT-1b	9.5†	8.7†	0.6	8.5†	1.8	6.2†	9.1†	0.7	2.9	1.7
PT-1.4b	6.0†	9.1†	1.5	6.6†	1.1	11.1†	7.0†	0.3	6.2†	2.2
PT-2.8b	5.8†	3.7	0.1	7.3†	4.3†	13.8†	6.9†	0.3	4.8†	1.2
PT-6.9b	8.5†	8.6†	2.2	10.8†	1.9	15.0†	11.6†	0.6	8.9†	1.5
PT-12b	6.8†	10.4†	1.0	9.8†	2.7	10.8†	8.2†	0.3	10.8†	1.8

Table 6: PPP (ΔLL) of surprisal update by model, phenomenon, and data group. The value with † indicates statistical significance ($p < 0.05$) with a likelihood ratio test.

	MVRR		NPS		NPZ		RC		Att.	
Model	All	RoI	All	RoI	All	RoI	All	RoI	All	RoI
G2-sm	10.0†	2.5	0.4	0.7	1.7	9.7†	11.2†	0.6	5.3†	2.6
G2-md	9.7†	2.0	0.8	5.1†	2.6	15.0†	18.9†	0.2	20.9†	4.9†
G2-lg	7.4†	2.2	0.5	0.8	1.8	16.0†	16.5†	0.1	16.3†	7.4†
G2-xl	3.1	2.2	1.0	0.3	1.4	15.5†	15.5†	0.4	16.2†	4.3†
OT-125m	7.7†	0.2	0.4	2.1	2.7	8.5†	8.5†	2.4	19.1†	6.0†
OT-1.3b	9.6†	1.6	2.6	1.4	4.3†	13.3†	15.4†	0.5	34.8†	4.0†
OT-2.7b	9.5†	1.9	2.1	1.2	4.4†	10.8†	11.5†	0.6	34.3†	3.7
OT-6.7b	2.7	2.9	0.8	4.0†	1.1	8.9†	10.2†	0.7	24.6†	2.4
OT-13b	6.5†	4.6†	1.7	2.8	3.8	8.2†	17.5†	0.7	26.4†	2.9
OT-30b	4.8†	5.9†	1.3	7.4†	3.1	7.1†	17.0†	0.6	31.4†	3.3
OT-66b	2.2	1.7	0.7	1.2	2.9	12.1†	9.3†	1.2	17.5†	3.5
PT-70m	10.6†	2.0	1.6	1.4	0.7	0.8	22.5†	3.3	22.6†	7.0†
PT-160m	24.5†	0.5	3.1	2.6	26.1†	1.8	50.3†	3.8	35.1†	9.7†
PT-410m	9.9†	2.9	1.2	1.7	0.6	2.4	9.2†	1.2	26.7†	9.2†
PT-1b	38.5†	3.2	1.9	1.9	8.6†	19.6†	34.8†	2.1	22.3†	11.7†
PT-1.4b	7.6†	0.3	0.1	1.9	1.1	5.9†	14.9†	1.8	16.0†	6.2†
PT-2.8b	20.3†	0.7	1.4	3.3	6.1†	12.7†	38.0†	3.7	31.7†	5.0†
PT-6.9b	4.3†	0.2	0.9	2.2	1.8	8.8†	18.7†	2.2	21.6†	6.6†
PT-12b	4.5†	0.3	0.2	1.3	1.5	8.0†	18.2†	2.3	21.9†	5.5†

Table 7: PPP (ΔLL) of KL-divergence $\mathrm{KL}(Q||P)$ by model, phenomenon, and data group. The value with † indicates statistical significance ($p < 0.05$) with a likelihood ratio test.

	MVRR		NPS		NPZ		RC		Att.	
Model	All	RoI	All	RoI	All	RoI	All	RoI	All	RoI
G2-sm	10.4†	2.4	0.42.5	2.2	12.3†	15.4†	0.5	17.2†	3.5	
G2-md	17.2†	3.0	2.52.8	6.6†	13.9†	30.8†	0.4	32.2†	6.5†	
G2-lg	9.5†	5.2†	1.61.3	1.9	12.0†	19.9†	0.1	27.4†	5.1†	
G2-xl	7.7†	2.6	1.50.9	2.2	11.9†	22.4†	0.3	29.5†	5.0†	
OT-125m	33.7†	0.8	2.12.3	2.2	8.1†	23.6†	0.5	29.5†	5.3†	
OT-1.3b	12.7†	2.1	2.60.4	6.2†	10.9†	30.1†	0.7	30.7†	4.6†	
OT-2.7b	12.1†	1.6	2.00.9	6.7†	9.8†	30.1†	1.0	30.3†	4.3†	
OT-6.7b	6.1†	3.4	1.13.4	3.1	13.3†	24.1†	0.9	25.8†	3.1	
OT-13b	9.1†	4.7†	1.62.1	5.2†	9.2†	28.5†	0.8	28.7†	3.4	
OT-30b	7.1†	8.8†	1.06.5†	4.3†	14.2†	25.0†	0.7	28.4†	3.3	
OT-66b	2.4	2.6	0.21.7	2.5	9.7†	11.6†	1.3	16.0†	3.8	
PT-70m	24.4†	0.6	1.81.5	3.7	6.0†	28.5†	3.5	33.5†	8.8†	
PT-160m	20.9†	0.2	2.34.0†	3.7	6.9†	26.0†	3.1	30.9†	7.3†	
PT-410m	20.8†	1.5	1.72.2	1.6	5.3†	19.9†	2.3	28.6†	5.3†	
PT-1b	18.9†	0.6	1.82.1	3.0	7.4†	26.5†	2.4	32.4†	7.9†	
PT-1.4b	14.4†	0.7	0.94.0†	1.5	5.9†	19.8†	1.8	24.6†	7.2†	
PT-2.8b	22.2†	0.4	2.32.7	3.9	5.8†	30.4†	3.4	32.5†	7.7†	
PT-6.9b	8.7†	0.9	1.10.4	1.8	5.4†	22.4†	1.9	23.8†	7.1†	
PT-12b	12.9†	0.4	1.01.1	1.7	5.1†	23.7†	2.3	25.8†	7.5†	

Table 8: PPP (ΔLL) of JS-divergence by model, phenomenon, and data group. The value with † indicates statistical significance ($p < 0.05$) with a likelihood ratio test.

Models	MVRR	NPS	NPZ	RC	Attach.
GPT2	0.07	−0.51	−0.86	3.36	0.08
GPT2-medium	−0.34	−1.42	−1.85	0.52	0.06
GPT2-large	0.20	−0.63	0.46	1.85	0.29
GPT2-xl	−0.31	−0.42	0.47	0.81	0.55
OPT-125m	0.08	−0.33	0.14	−0.86	−0.05
OPT-1.3b	−0.39	−1.12	−1.16	0.08	−0.23
OPT-2.7b	0.26	−1.23	−0.83	0.27	−0.01
OPT-6.7b	0.44	−0.85	−1.06	−0.98	0.02
OPT-13b	−0.60	−1.77	−0.68	−0.86	−0.15
OPT-30b	−1.08	−1.76	−1.58	−0.49	−0.15
OPT-66b	0.85	−0.58	−0.54	−1.27	−0.13
Pythia-70m	2.36	2.20	1.52	−0.10	0.38
Pythia-160m	0.50	0.91	0.32	−3.01	0.22
Pythia-410m	0.98	−0.65	−1.09	−0.77	−0.26
Pythia-1b	3.83	1.03	4.15	−0.19	0.30
Pythia-1.4b	2.07	0.77	0.50	−0.34	−0.31
Pythia-2.8b	1.98	0.69	3.84	−2.67	−0.29
Pythia-6.9b	1.21	−0.19	−0.67	−0.99	−0.66
Pythia-12b	2.24	−0.39	−0.19	−1.18	−0.83

Table 10: Estimated slowdown by KL-divergence $\mathrm{KL}(Q||P)$

Models	MVRR	NPS	NPZ	RC	Attach.
GPT2-small	5.00	4.30	7.97	5.91	2.61
GPT2-medium	2.13	2.25	5.40	5.29	1.55
GPT2-large	1.18	1.30	3.66	5.02	1.01
GPT2-xl	−0.27	1.14	2.52	5.16	0.34
OPT-125m	3.41	2.53	2.41	4.84	1.06
OPT-1.3b	0.01	0.80	0.77	3.99	−0.43
OPT-2.7b	0.23	0.79	0.78	2.70	−0.62
OPT-6.7b	1.39	2.29	2.72	3.86	−0.38
OPT-13b	−0.20	1.09	0.64	2.33	−0.80
OPT-30b	−0.41	1.98	1.87	2.58	−1.11
OPT-66b	−0.83	−0.38	−1.56	0.38	−0.70
Pythia-70m	1.36	2.14	3.70	0.04	1.26
Pythia-160m	0.59	0.65	3.06	−3.73	−0.10
Pythia-410m	−2.18	0.28	−0.33	−0.30	−1.74
Pythia-1b	−2.80	−1.90	−2.59	−0.79	−0.55
Pythia-1.4b	−2.67	−0.92	−2.83	−1.01	−0.59
Pythia-2.8b	−1.91	0.16	−1.11	−0.58	−1.61
Pythia-6.9b	−1.51	−0.03	−1.38	−1.21	−1.07
Pythia-12b	−0.80	1.02	−0.25	−1.22	−1.03

Table 9: Estimated slowdown by surprisal-update

Models	MVRR	NPS	NPZ	RC	Attach.
GPT2	−0.06	−0.56	−0.59	2.97	0.19
GPT2-medium	−0.59	0.28	0.38	0.69	0.15
GPT2-large	−0.58	−0.33	0.08	1.27	0.14
GPT2-xl	−0.40	0.23	0.31	0.58	0.49
OPT-125m	0.04	−0.01	−0.03	0.65	0.09
OPT-1.3b	−0.39	−0.28	0.27	−0.08	−0.09
OPT-2.7b	0.09	−0.98	0.06	0.23	−0.04
OPT-6.7b	0.22	−1.24	−1.26	−0.28	−0.13
OPT-13b	−0.53	−1.21	0.03	−0.81	−0.09
OPT-30b	−1.56	−2.70	−1.81	−0.45	−0.15
OPT-66b	0.53	−0.80	−0.65	−1.36	−0.10
Pythia-70m	0.20	0.08	0.21	0.47	−0.28
Pythia-160m	−0.01	0.07	0.15	0.09	−0.04
Pythia-410m	0.39	0.05	0.23	0.49	−0.22
Pythia-1b	−0.17	0.22	0.90	0.41	−0.17
Pythia-1.4b	0.35	0.00	0.01	0.03	−0.38
Pythia-2.8b	0.00	−0.03	0.26	0.44	−0.21
Pythia-6.9b	0.24	−0.36	−0.48	−0.39	−0.50
Pythia-12b	0.26	−0.36	−0.14	−0.26	−0.44

Table 11: Estimated slowdown by JS-divergence

	MVRR		NPS		NPZ		RC		Att.	
Model	All	RoI	All	RoI	All	RoI	All	RoI	All	RoI
G2-sm	31.9†	0.5	2.2	3.7	1.7	4.2†	24.7†	0.7	23.5†	3.6
G2-md	33.2†	2.1	4.4†	4.6†	6.3†	6.1†	30.5†	0.4	33.0†	5.8†
G2-lg	34.8†	3.9	3.3	2.2	4.5†	3.0	35.0†	0.6	33.9†	4.5†
G2-xl	31.4†	2.2	4.2†	1.0	5.5†	2.5	36.0†	0.4	33.7†	4.3†
OT-125m	46.0†	0.2	3.2	1.4	1.9	5.4†	24.8†	0.2	28.2†	3.8
OT-1.3b	32.4†	2.7	5.8†	0.9	6.6†	2.3	29.2†	0.6	30.8†	3.9
OT-2.7b	33.0†	2.3	5.2†	0.9	7.8†	2.7	30.5†	0.9	29.6†	4.1†
OT-6.7b	28.6†	3.4	4.6†	2.8	7.0†	4.2†	29.9†	0.5	27.9†	3.1
OT-13b	31.4†	4.0†	5.0†	0.6	7.3†	3.2	29.1†	0.2	28.7†	3.3
OT-30b	31.0†	9.5†	3.8	3.3	6.0†	5.0†	25.4†	0.3	27.6†	3.3
OT-66b	28.5†	2.4	2.7	0.3	2.9	4.9†	21.5†	1.7	21.6†	4.1†
PT-70m	17.5†	0.4	1.6	1.2	1.7	1.2	21.0†	1.6	24.8†	7.0†
PT-160m	15.4†	0.2	2.0	5.3†	3.7	2.2	23.6†	2.6	30.8†	6.1†
PT-410m	30.3†	0.3	3.3	0.7	1.5	2.9	27.2†	3.3	28.2†	4.8†
PT-1b	29.5†	1.1	3.6	2.1	3.3	5.9†	31.6†	2.9	31.6†	6.5†
PT-1.4b	38.1†	1.0	3.4	4.5†	1.9	2.9	25.6†	2.6	24.5†	5.0†
PT-2.8b	26.5†	0.1	3.3	2.2	3.9†	2.7	32.3†	2.7	29.5†	6.4†
PT-6.9b	29.8†	2.2	4.4†	1.3	2.9	1.2	30.0†	3.4	25.6†	6.7†
PT-12b	34.2†	0.9	3.2	1.4	2.9	2.9	28.9†	3.2	25.2†	5.1†

Table 12: ΔLL between Surprisal vs. Surprisal+JS settings. † indicates statistical significance ($p < 0.05$) with a likelihood ratio test.